



Spa2vec: Unsupervised representation of localized spatial gene expression signatures

Gabriele Partel  and Carolina Wählby 

Centre for Image Analysis, Department of Information Technology and SciLifeLab, Uppsala University, Uppsala, Sweden

Keywords

gene expression; graph representation learning; RNA profiling; spatial transcriptomics; tissue analysis

Correspondence

G. Partel, Department of Information Technology, Lägerhyddsvägen 2, 751 05 Uppsala, Sweden

Tel: +4618-471 3371

E-mail: gabriele.partel@it.uu.se

C. Wählby, Department of Information Technology, Lägerhyddsvägen 2, 751 05 Uppsala, Sweden

Tel: +4618-471 3473

E-mail: carolina.wahlby@it.uu.se

(Received 25 February 2020, revised 13 August 2020, accepted 17 September 2020)

doi:10.1111/febs.15572

Investigations of spatial cellular composition of tissue architectures revealed by multiplexed in situ RNA detection often rely on inaccurate cell segmentation or prior biological knowledge from complementary single-cell sequencing experiments. Here, we present spa2vec, an unsupervised segmentation-free approach for decrypting the spatial transcriptomic heterogeneity of complex tissues at subcellular resolution. Spa2vec represents the spatial transcriptomic landscape of tissue samples as a graph and leverages a powerful machine learning graph representation technique to create a lower dimensional representation of local spatial gene expression. We apply spa2vec to mouse brain data from three different in situ transcriptomic assays and to a spatial gene expression dataset consisting of hundreds of individual cells. We show that learned representations encode meaningful biological spatial information of re-occurring localized gene expression signatures involved in cellular and subcellular processes.

Database

Spatial gene expression data are available in Zenodo database at <https://doi.org/10.5281/zenodo.3897401>. Source code for reproducing analysis results and figures is available in Zenodo database at <http://www.doi.org/10.5281/zenodo.4030404>.

Recent advances in single-cell RNA (scRNA) sequencing [1,2] allow to dissect the cell-type heterogeneity of complex tissues at incredible pace. An international effort has started building comprehensive reference maps of gene expression at the cellular resolution to uncover the cell-type composition of entire organs and organisms [3]. However, in order to understand the functional architecture of a tissue it is essential to reconstruct the spatial organization of its constituent cell types. To this end, single-cell sequencing analyses are often complemented with imaging-based methods for spatially resolved multiplexed in situ RNA detection [4–8] that allow to map mRNA molecules directly in tissue samples and identify specific cell-type location, enabling the discovery of their functional role inside the tissue architecture.

Abbreviations

GNN, graph neural networks; GO, gene ontology; ISS, inonnonbreakingspacesitu sequencing; scRNA, single-cell RNA.

Previous attempts to map the spatial heterogeneity of cell types mostly relied on cell body segmentation algorithms and gene assignments to cells based on segmented cell boundaries [4–7]. Extracted per-cell gene expression profiles are successively clustered and annotated based on complementary scRNA sequencing analysis experiments or published literature [4–7].

This means that analysis of the spatial heterogeneity in tissue samples is limited by the accuracy of image segmentation algorithms to outline exact cell borders in dense and overlapping cell environments, with uneven illumination conditions and low signal-to-noise ratios. Moreover, while some cell types are defined by clear differences in their gene expression profiles, others differ by only a few genes in their transcriptome (e.g., like finely related neuronal subtypes) making their identification challenging.

Preliminary work from Park *et al.* [9] tries to address these problems proposing a segmentation-free spatial cell-type analysis based on cellular mRNA density estimation via Gaussian KDE [10], defining cell location as local maxima of mRNA-dense regions and extracting gene expression profiles for each cell (i.e., local maxima) as the averaged gene expression in that unit area. Qian *et al.* [11], instead, proposed a probabilistic framework for jointly assigning mRNAs to segmented cells and cells to cell types based on scRNA-seq cell-type priors, achieving a fine classification of interneuron subtypes of CA1 hippocampal region.

Despite these efforts for improving cell-type identification *in situ*, spatial cell-type analyses alone do not use the full power of *in situ* spatial transcriptomics: The subcellular resolution can reveal spatial heterogeneity also at subcellular levels. There is compelling evidence that many genes are expressed in a spatially dependent fashion independent of cell types [12], and this information is lost when analyzing transcriptional profiles of single cells. Moreover, there is a considerable amount of heterogeneity within each cell type explained by the balance between intrinsic regulatory networks and extrinsic subcellular processes depending on the local cellular microenvironment [13–17]. mRNA localization plays an important role in these cell differentiation processes as localization can vary during specific stages of cell development, and distinguishes cell phenotypes, activities, and communication. Specifically, mRNA localization is involved in cellular compartmentalization of gene expression into spatial functional domains involved in spatially targeted segregation of protein synthesis [18]. For example, mRNA localization is particularly diffused in neurons, where protein synthesis can take place at distal sites far away from the nucleus: Dendritic and axonal structures express several forms of plasticity that requires local translation [19–22]. Disruption of these subcellular biological processes (BP) was shown to be implicated in neurodevelopmental, psychiatric, or degenerative diseases [23–26]. It is thus important to take advantage of *in situ* mRNA detection methods to dissect the spatial heterogeneity of gene expression at subcellular resolution with respect to development and disease, and unveil the subcellular spatial domains underlying cell differentiation.

Here, we propose a novel segmentation-free approach for analyzing the spatial heterogeneity in gene expression of tissue samples that does not rely on the definition of cell types and cell segmentation but leverages the spatial organization of single mRNAs to define subcellular spatial domains involved in cellular differentiation. Specifically, we consider the spatial

organization of mRNAs inside tissues as local neighborhoods where groups of different mRNA types interact based on their spatial proximity (Fig. 1). These subcellular domains are shared or cell-type specific, and can therefore be expected to occur in several places inside a cell or across a tissue sample. In order to investigate the spatial mRNA network for recurrent gene expression signatures, we adopted a powerful graph representation learning technique [27] based on graph neural networks (GNN) [28], which has recently emerged as state-of-the-art machine learning technique for leveraging information from graph local neighborhoods. Therefore, each mRNA location is encoded in a graph as a node with a single feature representing the gene it belongs to and it is connected to all the other nodes representing the other mRNAs located in its neighborhood (Fig. 1A). During training, the GNN learns the topological structure of each node's local neighborhood as well as the distribution of node features in the neighborhood (i.e., local gene expression), and projects each node in a lower dimensional embedding space that encapsulates high-dimensional information about the node's neighborhood (Fig. 1B). We call this vectorization approach spatial gene expression to vector, or *spage2vec*, where geometric relations in this lower dimensional space correspond to higher order relationships in the local gene environment. We apply *spage2vec* to three publicly available mouse brain datasets [6,7,11] and compare the resulting gene expression signatures to cell-type maps presented in the respective publications. We further validate the method on a previously published spatial gene expression dataset of over 400 human fibroblast cells [29], conforming previously observed spatial gene expression patterns and identifying 26 new spatial domains involved in different molecular and BP.

Results

Spage2vec for *in situ* sequencing analysis of mouse hippocampal area CA1

We first analyzed published *in situ sequencing* (ISS) data of mouse hippocampal area CA1 [11], where transcripts of 99 genes were localized. After representing the spatial gene expression as a graph, we applied *spage2vec* to generate a 50-dimensional embedding for each mRNA spot (Materials and methods), encoding information of its local neighborhood. We then projected the 50-dimensional embedding to three dimensions in order to visualize spatial relationships learnt from the data as similar colors in RGB color

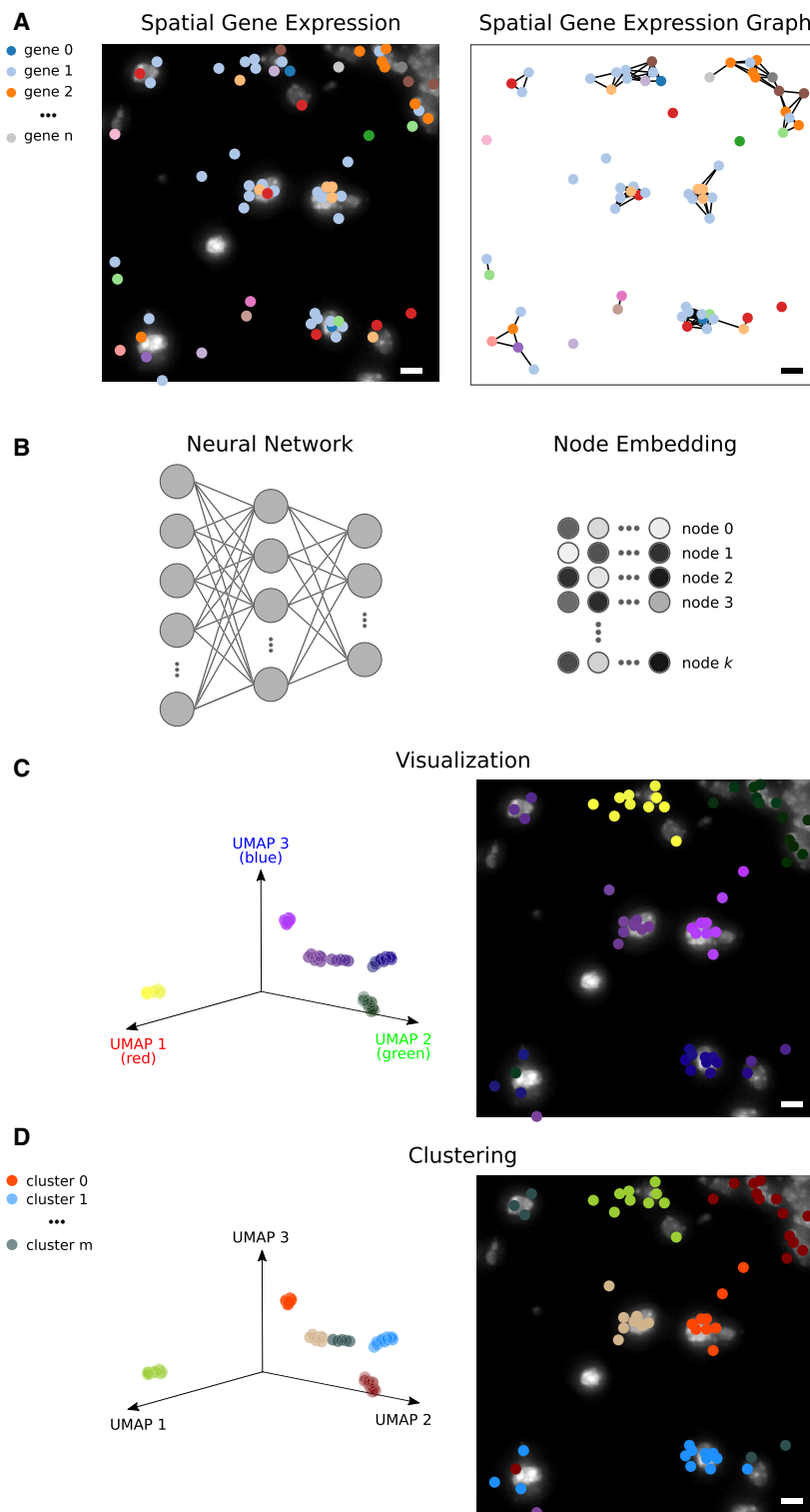


Fig. 1. Spage2vec workflow for detecting subcellular spatial domains from spatial gene expression data. (A) Left: Each colored dot represents a targeted gene, where color defines gene identity (targeted mRNAs representing n different genes, k dots). Cell nuclei are shown in grayscale in the background. Right: A graph connecting the neighboring dots from the left panel based on their spatial distances. (B) A lower dimensional representation of the graph is learnt for each of the k dots using a graph representation learning technique based on a GNN. The neural network predicts a node embedding vector for each dot of the graph representing high order spatial relationships with its local neighborhood ([Materials and method](#)). (C) Thereafter, the spatial gene expression variation is visualized at subcellular resolution by projecting the learnt node embedding vectors into a 3D RGB color space using UMAP. (D) Clusters representing localized gene expression signatures are obtained by unsupervised clustering analysis of the embedding. Scale bars 5 μm .

space (Fig. 2A,C). Next, in order to investigate whether the learnt lower dimensional embedding contains significant information of biological functional

domains, we clustered the spot embeddings directly in the 50-dimensional space ([Materials and methods](#)) and compared obtained spot cluster labels with cell-

type annotations of mRNA spots from Qian *et al.* We initially obtained 29 clusters, which reduced to 25 after merging highly correlated clusters (Fig. S1) (Materials and methods). Identified clusters can be interactively explored in TissUMaps [30] at https://tissumaps.research.it.uu.se/demo/ISS_Qian_et_al.html (Data S1). We then compared the 25 identified clusters with 20 cell-type and 69 subcell-type annotations defined in Qian *et al.*, excluding mRNA spots without cell-type labels (Fig. 2E,F). To demonstrate the ability of the model to generalize over unseen data, we used the spage2vec model trained on the right hemisphere mouse hippocampal area CA1 to predict the node embedding for the spatial gene expression graph of the left hemisphere CA1 area unseen during training (Fig. 2B,D). As can be seen in the figures (Fig. 2A–D), the node representation of the two spatial gene expression graphs projected and visualized in RGB color space shows that the model produces visually similar embeddings for data not available during training.

spage2vec for osmFISH analysis of mouse somatosensory cortex

In order to demonstrate the generalizability of spage2vec to other datasets, we also produced a lower dimensional representation of mRNAs from published osmFISH data of 33 cell-type marker genes targeted in mouse brain somatosensory cortex [7]. Again, we represented the gene expression as a graph and applied spage2vec, resulting in a 50-dimensional representation of each mRNA spot. We projected the 50 dimensions to three dimensions and visualized similar local gene expression signatures as similar colors in 3D RGB color space (Fig. 3A). Next, we clustered the learnt embedding space in 274 domains and reduced to 69 domains after merging highly correlated clusters (Fig. S2) (Materials and methods). Identified clusters can be interactively explored at https://tissumaps.research.it.uu.se/demo/osmFISH_Codeluppi_et_al.html (Data S1). We then compared the resulting 69 clusters with the 31 cell-type annotations defined in Codeluppi *et al.*, excluding spots without cell-type labels (Fig. 3B, C).

Spage2vec for MERFISH analysis of mouse hypothalamic preoptic region

We further applied spage2vec to a 3D mRNA localization dataset of hypothalamic preoptic region analyzed by MERFISH [6], where the transcripts of 135 targeted genes were localized in 3D. As for the

previous dataset, we applied spage2vec to the graph representation (in this case 3D) and projected the 50 dimensions into three for visualization (Fig. 4A). Leveraging the symmetry of the data, we trained a spage2vec model on approximately half the sample (0–956 μm) and tested on the other half. Clustering in 50-dimensional space resulted in 198 clusters, which were reduced to 121 after merging of clusters with a gene expression correlation greater than 95% (Fig. S3). Identified clusters can be interactively explored at https://tissumaps.research.it.uu.se/demo/MERFISH_Moffitt_et_al.html (Data S1). We compared the gene expression profiles of these 121 clusters with the 10 cell types and 76 subcell types presented in Ref. [6] (Fig. 4B–D).

Spage2vec for MERFISH analysis of human fibroblast cells (IMR90)

We performed a spatial gene expression analysis using spage2vec on a MERFISH dataset (Chen *et al.*) [29] consisting of over 400 human fibroblast cells (IMR90) targeted with a gene panel of 130 genes. In this case, a spatial gene expression graph is generated for each of the analyzed cells and then merged as disjoint sub-graphs in a bigger network. We then produced a 50-dimensional representation of the input spatial gene expression graph that clustered in 59 clusters. Next, we merged, by summing the expression, clusters with a pairwise correlation greater than 90%. And finally, we removed small spurious clusters containing < 1000 reads, resulting in a final set of 26 clusters.

Identified clusters can be interactively explored at https://tissumaps.research.it.uu.se/demo/MERFISH_Chen_et_al_2015.html.

We then validated the representation learned by spage2vec with respect to the two groups of genes (i.e., group I: THBS1, LRP1, GPR107, PAPP, FBN1, FBN2, group II: MYH10, DYNC1H1, CKAP5, FLNC, SPTBN1, FLNA, TLN1, SPTAN1) identified by Chen *et al.* as having similar subcellular spatial gene expression patterns. We show that these two groups of genes map on close and overlapping locations in the learnt spage2vec embedding space (Fig. 5A) and show distinct cluster expression profiles (Fig. 5B). Thus spage2vec successfully learnt a meaningful representation of the input spatial gene expression, where identified clusters group based on correlation of their gene expression profiles (Fig. S4) into two main groups dominated, respectively, by group I and group II genes, confirming spatial gene expression patterns identified by the Chen *et al.*'s analysis (Fig. 5C).

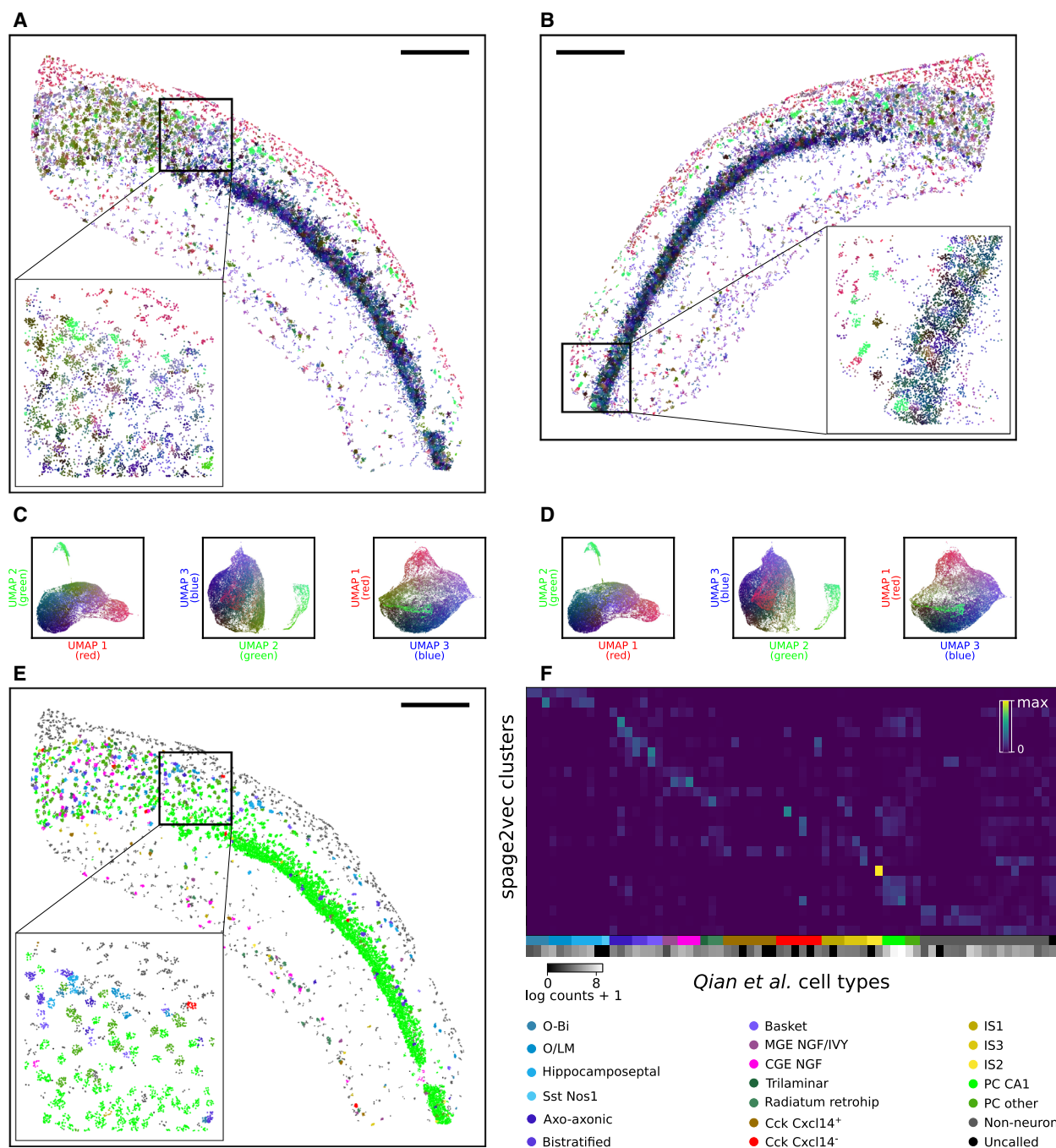


Fig. 2. Application of spage2vec to ISS data of mouse hippocampal area CA1. Visualization of functional variation of spatial gene expression at subcellular resolution in right (A) and left (B) hippocampal area CA1, where mRNAs are color-coded based on their node embedding projections in RGB color space for right (C) and left (D) hemisphere. (E) Spatial gene expression with mRNAs color-coded based on their cell-type annotation defined in Qian *et al.* (legend at bottom right). (F) Heat map showing distribution of gene counts in each spage2vec cluster (normalized by total gene count per cell type and cluster) and cell- and subcell-type annotation per gene transcript from Qian *et al.* (same legend as for (E)). The heat map is normalized to enhance how the spage2vec clusters correlate to cell types (or multiple cell types). Some cell types are not present in this part of the brain (representing only a subset of the data in Qian *et al.*), and therefore, some columns are empty. The grayscale at the bottom of the heat map shows total gene counts per cell subtype as found by Qian *et al.* Scale bars ~ 300 μm .

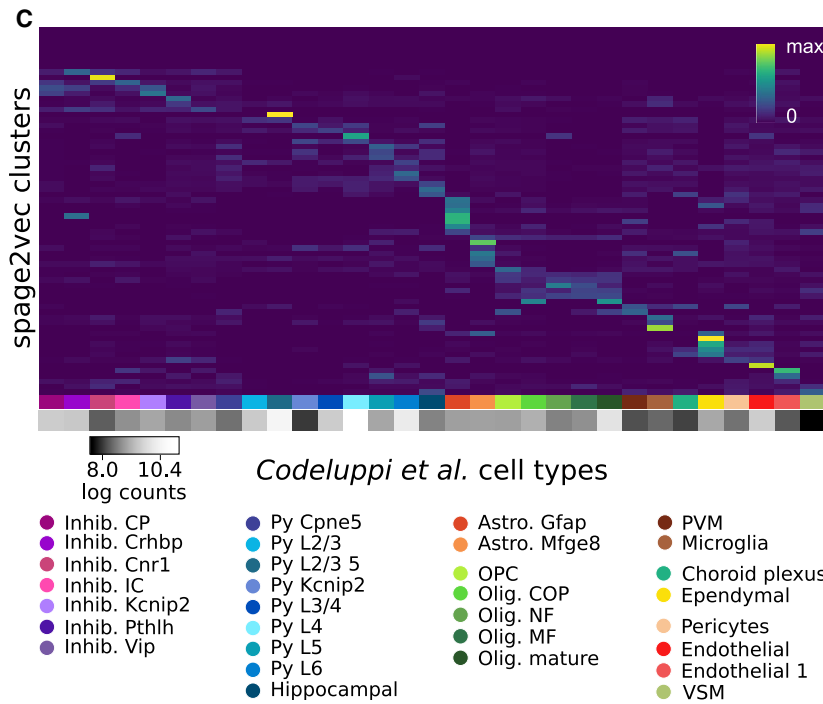
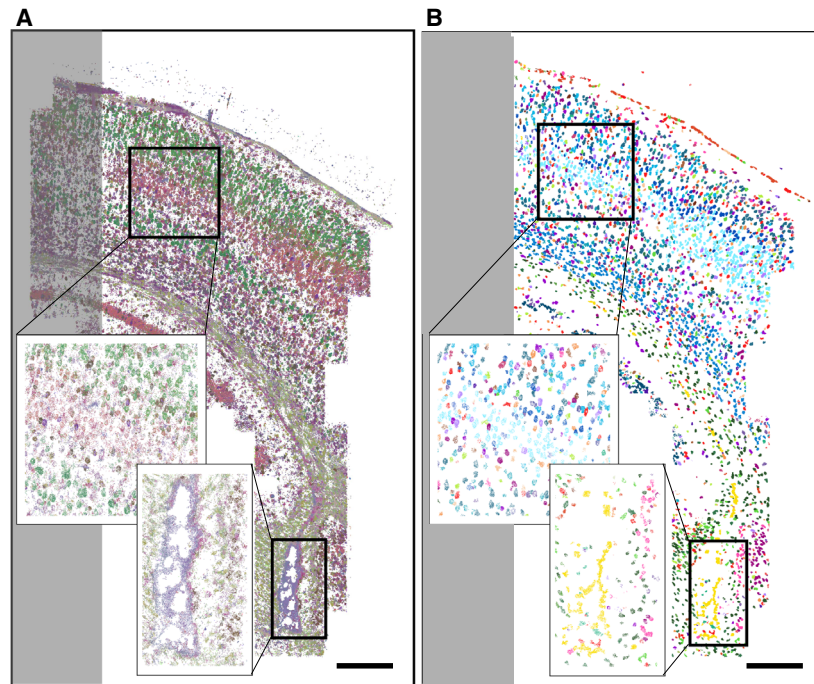
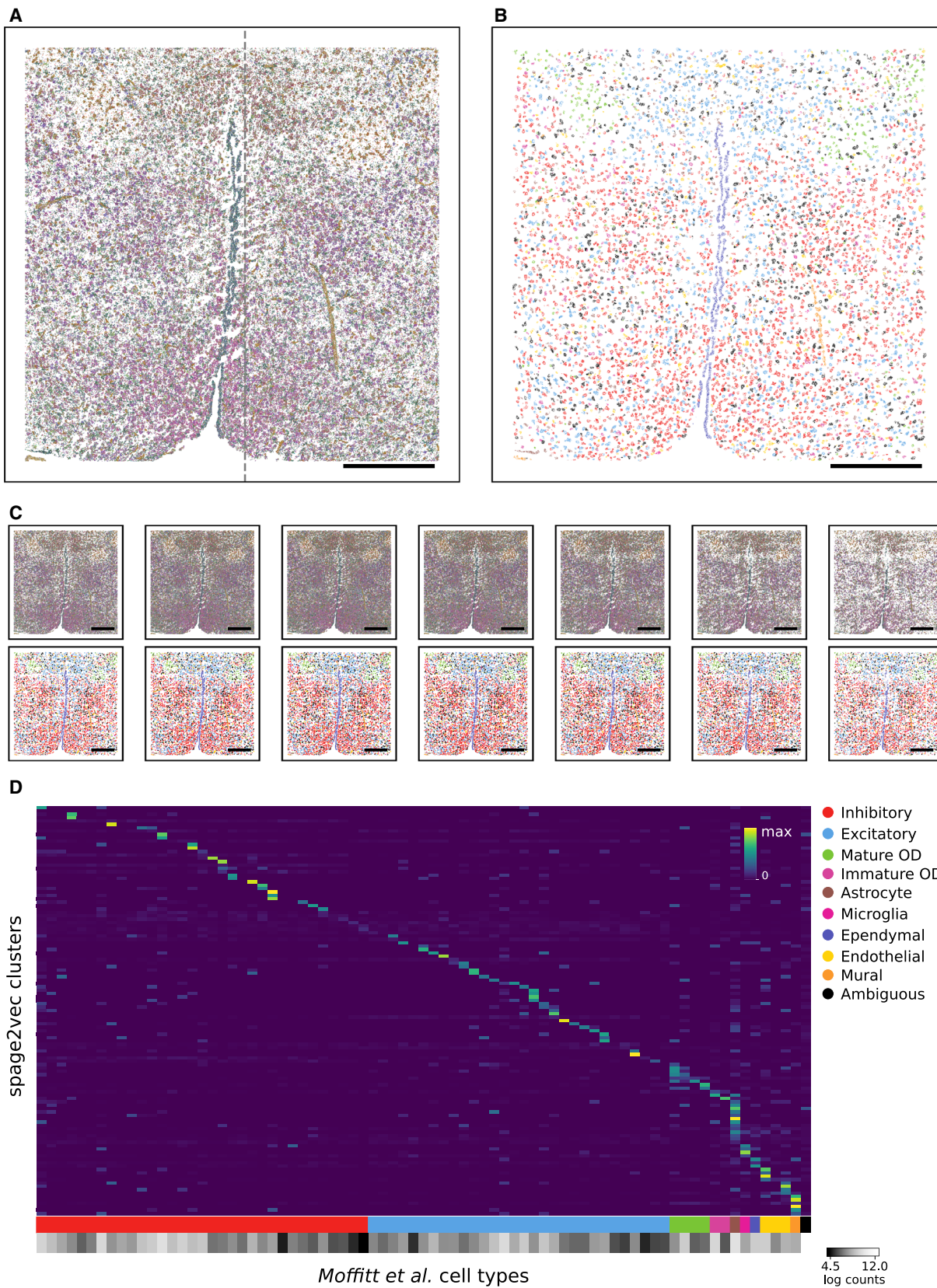


Fig. 3. Application of spage2vec to osmFISH data from the mouse brain somatosensory cortex. (A) Visualization of functional variation of spatial gene expression at subcellular resolution, where mRNAs are color-coded based on node embedding projection in RGB color space, and (B) spatial gene expression with mRNAs color-coded based on cell-type annotations defined from Codeluppi *et al.* cell segmentation. Shaded areas correspond to regions excluded in the original cell-type analysis. (C) Heat map showing normalized gene counts having specific spage2vec cluster labels and cell-type annotations from Codeluppi *et al.* (marked with different colors), and cell-type legend. The grayscale at the bottom of the heat map shows total gene counts per cell subtype as found by Codeluppi *et al.* Scale bars ~ 300 μm .

Fig. 4. Application of spage2vec to MERFISH data of the mouse brain hypothalamic preoptic region. (A) Visualization of functional variation of spatial gene expression at subcellular resolution, where mRNAs are color-coded based on their node embedding projections in RGB color space. The gray dashed line defines regions of the sample used for training (left) and for testing (right). (B) Spatial gene expression with mRNAs color-coded based on cell-type annotations defined from Moffitt *et al.* cell segmentation. (C) Spatial distribution of node embedding projections in RGB color space (upper row) and cell-type annotations (bottom row) from Moffitt *et al.* across the whole section. (D) Heat map showing normalized gene counts having specific spage2vec cluster labels and cell annotations from Moffitt *et al.* (marked with different colors), and cell-type legend. The grayscale at the bottom of the heat map shows total gene counts per cell subtype as found by Moffitt *et al.* Scale bars ~ 400 μm .



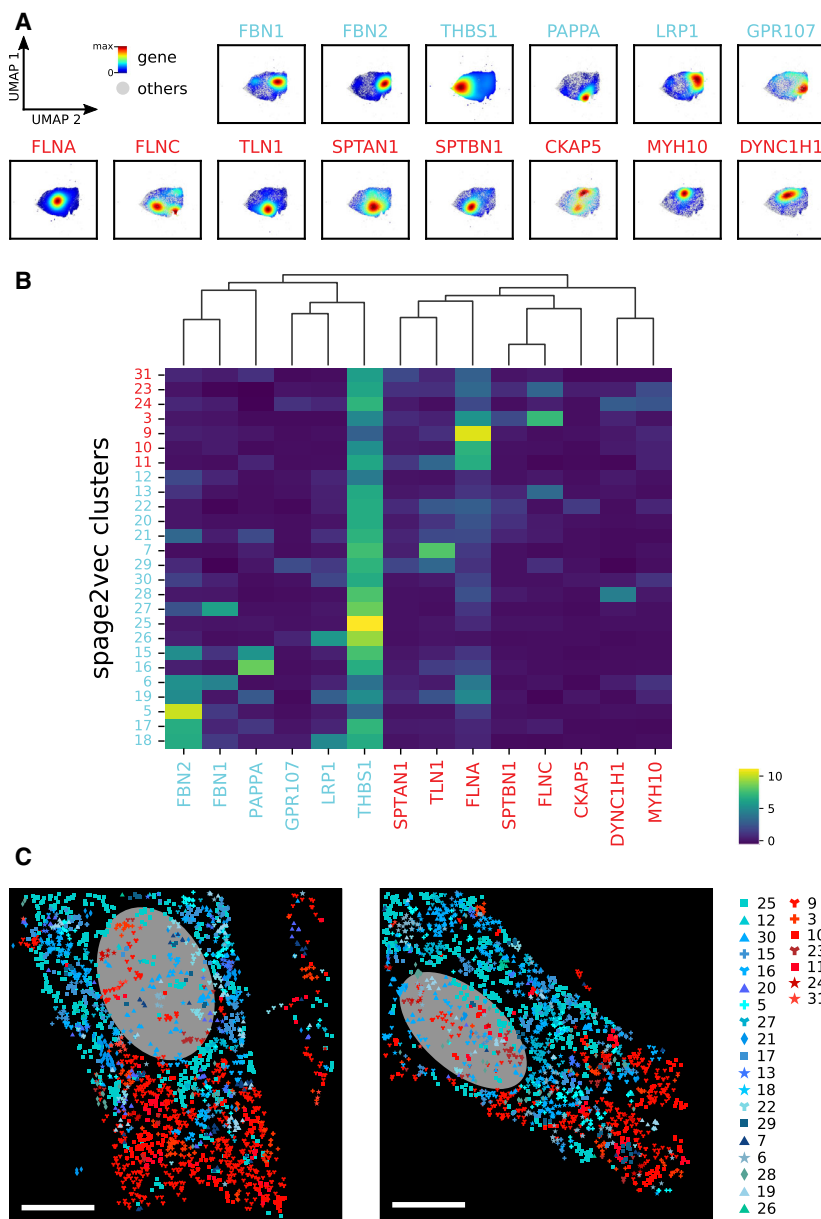


Fig. 5. Application of spage2vec to MERFISH data of human fibroblast cells (IMR90). (A) UMAP 2D projection of spage2vec embedding of genes identified in *Chen et al.* as forming two distinct spatial gene expression patterns (respectively, group I marked in light blue, and group II marked in red). Each plot shows the mRNAs of each gene color-coded based on their density profile estimated using a Gaussian kernel and all other genes in gray. Markers and axis legends are shown on the top left. (B) Heat map showing the obtained spage2vec cluster labels of each mRNA with respect to its gene label for group I and group II (normalized per spage2vec cluster on full dataset). (C) Spatial distributions of all identified clusters in two example cells, where each mRNA is displayed by a marker with color and symbol according to the cluster it belongs to (legend on the right). Cell nuclei have been approximated from Fig. 4C in *Chen et al.* with semitransparent ovals. Clusters have been divided into two groups based on correlation of their gene expression profiles (Fig. S4) and color-coded, respectively, with shades of blue and red. Scale bars ~ 10 μ m.

This is an indication that the learnt spage2vec embedding not only represents localized gene expression signatures, but it also manages to capture the global structures of the spatial gene expression data. Moreover, to confirm the biological significance of the identified clusters, we perform gene ontology (GO) enrichment analysis of the highly expressed genes that characterized each cluster (Table S1). Most of the identified clusters were significantly enriched with multiple GO terms. Recurrent terms were related to location terms like extracellular region, and extracellular matrix, and BP involved in cell signaling and motility.

Discussion

We showed that spage2vec can learn low-dimensional embeddings encoding important topological and functional information of local gene expression. This rich low-dimensional space can be used for downstream clustering analysis in order to detect biologically meaningful re-occurring gene expression signatures that correlate well with subcellular and cellular domains. The embedding, found by unsupervised training, has an inductive property to generalize over unseen nodes. This means that it can be applied to a new unseen dataset, as long as the new dataset has the same

feature set (i.e., consists of gene expression data from the same gene panel). This is especially useful to predict embeddings for new spatial gene expression datasets and map them to a common lower dimensional space. The fact that `spage2vec` is a fully unsupervised approach triggers the possibility to explore cellular heterogeneity in situ without the need of scRNA sequencing data-driven analysis.

The presented approach is completely independent of cell segmentation, and equally applicable to 2D and 3D data, meaning that dense gene expression datasets such as those from MERFISH can be analyzed without relying on the accuracy of cell segmentation. In fact, most cell segmentation approaches are based on identifying cell nuclei and then approximating gene-to-cell assignment by shortest distance to the closest nucleus. This can very often introduce noise as cells may vary very much in shape, and the nucleus of a given cell may not even be present in the same tissue section as the bulk of the cell. Furthermore, the presented segmentation-free `spage2vec` approach enables detection of biologically significant cellular and subcellular components as well as subcellular gene expression signatures representing functional domains located far away from a cell nucleus.

Materials and methods

Building a spatial gene expression graph

Spatially resolved gene expression data consist of gene expression information and coordinates describing spatial location (in 2D or 3D) in a tissue sample. This information can be represented as a graph by saying that a node in the graph is a single mRNA that has a categorical feature representing the gene it belongs to. Next, connections are drawn between each mRNA and all its local neighbors within a maximum spatial distance d_{\max} . We automatically define the distance d_{\max} such that $\geq 97\%$ of all nodes are connected to at least one neighbor, automatically adjusting for the spatial resolution of the dataset and providing a good balance between global and local features in the representation. Connected components with less than three nodes representing spurious expressions are removed from the graph before further processing (Fig. 1A). Note that the same graph representation works in both 2D and 3D.

Neural network model and training

Next, `spage2vec` strives to transform the spatial gene expression graph into an embedding where similar localized gene expression signatures are assigned similar vectors using a neural network model. The neural network model

consists of an unsupervised GraphSAGE [27] model implemented with the open source machine learning python library StellarGraph (<https://github.com/stellargraph/stellargraph>). The model learns embeddings of unlabeled graph nodes by combining the node's own feature with features sampled and aggregated from the node's local neighborhood. Specifically, node embeddings are learnt by solving a binary node classification task that predicts whether arbitrary node pairs are likely to co-occur in a random walk performed on the graph. For this task, the training set consists of *positive* node pairs, pairs that co-occur within walks of length 2 on the graph, and *negative* pairs of nodes uniformly randomly selected from the graph. Through training this binary node pair classifier, the model automatically learns an inductive mapping from a high-dimensional feature space (i.e., spatial gene expression) to a lower dimensional node embedding space, preserving important topological and structural features of the nodes. The model architecture consists of two identical GraphSAGE encoder networks sharing weights, taking as input a pair of nodes together with the graph structure and producing as output a pair of node embeddings. Thereafter, a binary classification layer with a sigmoid activation function learns to predict how likely it is that a pair will occur at a random position in the graph. Model parameters are optimized by minimizing binary cross-entropy between the predicted node pair labels and the true labels, without supervision.

Neural network hyperparameters

The proposed `spage2vec` model architecture used for all experiments presented here consists of two GraphSAGE layers with 50 hidden units, a bias term, l2 normalization, and l1 kernel regularization, using attentional aggregator function [31] with LeakyReLU [32]. Each GraphSAGE encoder embeds each node's neighborhood with a 2-hop node aggregation strategy, sampling, respectively, 20 and 10 nodes for the first and the second hops. The model is trained with on-the-fly batch generation with batch size equal to 50, using Adam [33] as optimizer with learning rate equal to $0.5e-4$. The output of `spage2vec` will thus be one vector of length 50 per spatial gene expression position. All details and settings are provided as Python notebooks (<https://github.com/wahlby-lab/spage2vec>).

Depending on the number of mRNAs in the dataset and the size of the gene panel, we suggest a different dimensionality for the `spage2vec` embedding such that it can capture meaningful variation in the data but also produces not too sparse representations that would negatively influence posterior clustering performances. We recommend using more hidden nodes and higher embedding dimensionality for spatial gene expression datasets with a larger gene panel and higher number of mRNAs, so that the complexity of the data can better be captured by the GNN. The number of hidden layers is strictly related to the number of

hops, or search depth, used in the node aggregation strategy, where an increasing number of hops aggregate information further away from a given node and consequently better capture global features of the input spatial gene expression graph. Nevertheless, it has been shown that using more than two hops gave only marginal returns in performances while consistently increasing training time [27]. Instead, the number of nodes sampled in each hop is used to uniformly sample a fixed-size set of neighbors in order to have fixed memory and computational footprint at each batch. In Ref. [27], the authors show that GraphSAGE can achieve generally good performance using two hops with a total sampling size ≤ 500 . We set these two parameters in order to preserve low variance and relatively low training time. Specifically, we look at average node degree of the spatial gene expression graph (e.g., average node degree equal to 8) setting a slightly higher value for sampling in the second hop (e.g., number of nodes sampled in the second hop equal to 10), and a doubly larger value for the first hop in order to have a lower variance for the closest neighbors (e.g., number of nodes sampled in the first hop equal to 20).

Visualization of node embeddings

To visualize the extracted spatial gene expression embeddings created by `spage2vec`, we reduced the embedding dimensionality to three dimensions with UMAP [34]. This allowed us to present the localized gene expression signatures as data points in a 3D RGB color space. Mapping the new color-coding back to tissue space shows that many of the transcripts not only cluster in space but also seem to recur and correlate with cellular and subcellular spatial domains (Fig. 1D).

Identification of localized gene expression signatures

For further comparing the `spage2vec` output with approaches aimed at identifying cell types, we hypothesize that recurring localized gene expression signatures are spatial functional domains that may be cell-type-specific, or represent processes shared among different cell types. We therefore cluster the 50-dimensional `spage2vec` output using the Leiden clustering algorithm [35,36] followed by column-wise Z -score normalization of the cluster expression matrix (genes \times clusters). Clusters where gene expression profiles have a correlation greater than 95% are merged by summing their expression counts, and the merged cluster expression matrix is renormalized with Z -score normalization, leading to a final set of clusters. Note that the trained model has an inductive property, meaning that it can generalize and find embeddings for previously unseen localized gene expression signatures.

Gene ontology analysis

In order to examine the identified clusters for enrichments of GO terms in Chen *et al.* dataset, we extract for each cluster highly expressed genes that have Z -scores higher than 1. We search for enrichments in location terms (CC), molecular-level activities (MF), and BP with `goatools` [37] using the most recent human annotations (<https://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz>). For each selected gene set of each cluster, we query both the relative annotated GO terms and terms immediately upstream, against a background list composed by all the genes present in the panel. Terms found to be statistically significant with P -values smaller than 0.005 were reported (Table S1).

Datasets

We apply `spage2vec` to three publicly available published mouse brain tissue datasets obtained by three different spatial transcriptomics assays: (a) ISS of left and right hippocampal area CA1 [11]; https://tissuumsmaps.research.it.uu.se/demo/ISS_Qian_et_al.html, with a resolution of 0.325 μm per px and a total of 84 880 detections of 99 different mRNAs; (b) an `osmFISH` analysis of the somatosensory cortex [7]; https://tissuumsmaps.research.it.uu.se/demo/osmFISH_Codeluppi_et_al.html, comprising a tissue section of 3.8 mm^2 , with a resolution of 0.065 μm per pixel, and a total of 1 802 589 detections of 33 different mRNAs; and (c) a `MERFISH` analysis of the hypothalamic preoptic region [6]; https://tissuumsmaps.research.it.uu.se/demo/MERFISH_Moffitt_et_al.html, comprising a 3D tissue section 10 μm thick of 1.8 by 1.8 mm and a total of 3 728 169 detections targeting 135 different genes.

We further validated `spage2vec` on a `MERFISH` dataset of 421 human fibroblast cells (IMR90) [29]; https://tissuumsmaps.research.it.uu.se/demo/MERFISH_Chen_et_al_2015.html imaged with 1.45 NA, 100 \times oil immersion objective. The gene panel consisted of 130 genes, with 740 043 decoded transcripts.

Code availability

All software was developed in Python 3 using open source libraries. The processing pipeline and the source code used to generate figures and analysis results presented in this paper are available as Python notebooks at <http://www.doi.org/10.5281/zenodo.4030404>, or on our GitHub repository (<https://github.com/wahlby-lab/spage2vec>).

Acknowledgements

We thank Mats Nilsson, Sten Linnarsson, and Xiaowei Zhuang for making their datasets publicly available. We also thank Leslie Solorzano for providing

support in visualization of the results with the TisUMaps viewer. This research was funded by the European Research Council via ERC Consolidator grant 682810 to CW and Swedish Foundation for Strategic Research (grant BD150008).

Conflict of interest

The authors declare no conflict of interest.

Author contributions

GP conceived the method and performed data analysis. CW supervised the study and provided critical comments and discussions. All authors provided critical feedback and helped shape the research, analysis, and manuscript.

Peer Review

The peer review history for this article is available at <https://publons.com/publon/10.1111/febs.15572>.

References

- Svensson V, Vento-Tormo R & Teichmann SA (2018) Exponential scaling of single-cell RNA-seq in the past decade. *Nat Protoc* **13**, 599–604.
- Grün D & van Oudenaarden A (2015) Design and analysis of single-cell sequencing experiments. *Cell* **163**, 799–810.
- Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Clatworthy M *et al.* (2017) Science forum: the human cell atlas. *Elife* **6**, e27041.
- Shah S, Lubeck E, Zhou W & Cai L (2016) In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron* **92**, 342–357.
- Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, Evans K, Liu C, Ramakrishnan C, Liu J *et al.* (2018) Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691.
- Moffitt JR, Bambach-Mukku D, Eichhorn SW, Vaughn E, Shekhar K, Perez JD, Rubinstein ND, Hao J, Regev A, Dulac C *et al.* (2018) Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, eaau5324.
- Codeluppi S, Borm LE, Zeisel A, La Manno G, van Lunteren JA, Svensson CI & Linnarsson S (2018) Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat Methods* **15**, 932–935.
- Eng CH, Lawson M, Zhu Q, Dries R, Koulina N, Takei Y, Yun J, Cronin C, Karp C, Yuan GC *et al.* (2019) Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* **568**, 235–239.
- Park J, Choi W, Tiesmeyer S, Long B, Borm LE, Garren E, Nguyen TN, Codeluppi S, Schlesner M, Tasic B *et al.* (2019) Segmentation-free inference of cell types from in situ transcriptomics data. *bioRxiv* 800748. [PREPRINT]
- Parzen E (1962) On estimation of a probability density function and mode. *Ann Math Stat* **33**, 1065–1076.
- Qian X, Harris KD, Hauling T, Nicoloutsopoulos D, Muñoz-Manchado AB, Skene N, Hjerling-Leffler J & Nilsson M (2020) Probabilistic cell typing enables fine mapping of closely related cell types in situ. *Nat Methods* **17**, 101–106.
- Zhu Q, Shah S, Dries R, Cai L & Yuan GC (2018) Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. *Nat Biotech* **36**, 1183.
- Quail DF & Joyce JA (2013) Microenvironmental regulation of tumor progression and metastasis. *Nat Med* **19**, 1423.
- Riquelme PA, Drapeau E & Doetsch F (2008) Brain micro-ecologies: neural stem cell niches in the adult mammalian brain. *Philos Trans Royal Soc B Biol Sci* **363**, 123–137.
- Swain PS, Elowitz MB & Siggia ED (2002) Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc Natl Acad Sci USA* **99**, 12795–12800.
- Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, Rotem A, Rodman C, Lian C, Murphy G *et al.* (2016) Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196.
- Zhang J & Li L (2008) Stem cell niche: microenvironment and beyond. *J Biol Chem* **283**, 9499–9503.
- Buxbaum AR, Haimovich G & Singer RH (2015) In the right place at the right time: visualizing and understanding mRNA localization. *Nat Rev Mol Cell Biol* **16**, 95–109.
- Cajigas IJ, Tushev G, Will TJ, tom Dieck S, Fuerst N & Schuman EM (2012) The local transcriptome in the synaptic neuropil revealed by deep sequencing and high-resolution imaging. *Neuron* **74**, 453–466.
- Besse F & Ephrussi A (2008) Translational control of localized mRNAs: restricting protein synthesis in space and time. *Nat Rev Mol Cell Biol* **9**, 971–980.
- Holt CE & Bullock SL (2009) Subcellular mRNA localization in animal cells and why it matters. *Science* **326**, 1212–1216.
- Das S, Singer RH & Yoon YJ (2019) The travels of mRNAs in neurons: do they know where they are going? *Curr Opin Neurobiol* **57**, 110–116.

- 23 Miller S, Yasuda M, Coats JK, Jones Y, Martone ME & Mayford M (2002) Disruption of dendritic translation of CaMKII α impairs stabilization of synaptic plasticity and memory consolidation. *Neuron* **36**, 507–519.
- 24 Perry RB, Doron-Mandel E, Iavnilovitch E, Rishal I, Dagan SY, Tsoory M, Coppola G, McDonald MK, Gomes C, Geschwind DH *et al.* (2012) Subcellular knockout of importin β 1 perturbs axonal retrograde signaling. *Neuron* **75**, 294–305.
- 25 Yoon BC, Jung H, Dwivedy A, O'Hare CM, Zivraj KH & Holt CE (2012) Local translation of extranuclear lamin B promotes axon maintenance. *Cell* **148**, 752–764.
- 26 Swanger SA & Bassell GJ (2011) Making and breaking synapses through local mRNA regulation. *Curr Opin Genet Dev* **21**, 414–421.
- 27 Hamilton W, Ying Z & Leskovec J (2017) Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, (pp. 1024–1034).
- 28 Wu Z, Pan S, Chen F, Long G, Zhang C & Philip SY (2020) A comprehensive survey on graph neural networks. *IEEE Trans Neural Netw Learn Syst* 1–21. doi: 10.1109/TNNLS.2020.2978386
- 29 Chen KH, Boettiger AN, Moffitt JR, Wang S & Zhuang X (2015) Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090.
- 30 Solorzano L, Partel G & Wählby C (2020) TissUUmapi: interactive visualization of large-scale spatial gene expression and tissue morphology data. *Bioinformatics* **36**, 4363–4365.
- 31 Veličković P, Cucurull G, Casanova A, Romero A, Lio P & Bengio Y (2017) Graph attention networks. *arXiv [Preprint]* arXiv: 1710.10903.
- 32 Maas AL, Hannun AY & Ng AY (2013) Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml* **30**, 3. http://robotics.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf
- 33 Kingma DP & Adam BJ (2014) A method for stochastic optimization. *arXiv [Preprint]*. arXiv:1412.6980.
- 34 McInnes L, Healy J & Melville J (2018) Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv [Preprint]* arXiv:1802.03426.
- 35 Traag VA, Waltman L & van Eck NJ (2019) From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* **9**, 1–2.
- 36 Wolf FA, Angerer P & Theis FJ (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15.
- 37 Klopfenstein DV, Zhang L, Pedersen BS, Ramírez F, Vesztrocy AW, Naldi A, Mungall CJ, Yunes JM, Botvinnik O, Weigel M *et al.* (2018) GOATOOLS: a python library for gene ontology analyses. *Sci Rep* **8**, 1–7.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Data S1. Instructions for visualizing spage2vec clusters in TissUUmapi online viewer.

Fig. S1. Gene expression per detected cluster of the ISS data from Qian X. *et al.*

Fig. S2. Gene expression per detected cluster of the osmFISH data from Codeluppi S. *et al.*

Fig. S3. Gene expression per detected cluster of the MERFISH data from Moffitt J.R. *et al.*

Fig. S4. Gene expression per detected cluster of the MERFISH data from Chen *et al.*

Table S1. GO analysis of spage2vec clusters of MERFISH Chen *et al.* spatial gene expression data.