POINTS2REGIONS: FAST INTERACTIVE CLUSTERING OF *IN SITU* TRANSCRIPTOMICS DATA

Axel Andersson, Andrea Behanova*, Christophe Avenel, Carolina Wählby, Filip Malmberg*

Department of IT and SciLifeLab BioImage Informatics Facility, Uppsala University, Sweden

ABSTRACT

Investigation of the spatial cellular architecture in tissue is crucial for furthering our understanding of biological processes and diseases. The location of different cell types can be revealed by multiplexed in situ RNA detection. Most of the analysis algorithms for in situ transcriptomics (IST) rely on cell segmentation or prior biological knowledge from complementary single-cell sequencing experiments. Here, we present Points2Regions, a fast and simple tool for interactive clustering and exploration of IST data. This tool comes as a plugin in TissUUmaps 3, making it accessible to an audience without programming experience. Points2Regions defines regions with similar composition of mRNA-markers. The regions are directly displayed on top of large-scale image data in TissUUmaps 3, making visual exploration easy. The tool is computationally efficient and can run without the need of any specialized hardware.

Index Terms— clustering, in situ transcriptomics, tissue analysis, interactive tool

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

1. INTRODUCTION

Over the past years, a large number of techniques for spatially multiplexed in situ transcriptomics (IST) have been developed [1, 2, 3, 4, 5, 6, 7]. These techniques enable mapping of hundreds of different mRNA molecules directly within tissue samples, allowing the dissection and analysis of cell-type heterogeneity while preserving spatial information. Output from these techniques is usually Gigapixel-sized images of tissue sections accompanied by millions of different spatial biomarkers for various mRNA molecules. Researchers are progressively advancing these techniques by improving their detection efficiency, experiment throughput, and multiplexing capabilities, resulting in even larger abundances of data. However, the large quantities and the vast number of unique types of markers obtained in an IST experiment make visual exploration and analysis challenging. To alleviate this, a range of tools have been developed for grouping the marker data into distinct instances, such as cells, cell-types or tissue-level compartments.

1.1. Tools for analyzing IST data

The classical approach for analysing IST data usually starts by segmenting cells using the distance-transform Watershed approach. Detected mRNA molecules are then assigned to cells based on the segmentation. The molecular composition for each cell is subsequentially used to define different cell-types, usually by using clustering analysis [4]. This approach, however, has fundamental limitations and thus several alternative methods have been proposed. For example, the SSAM framework by Park et al. [8] identifies various cell-types in the IST data without relying on any type of cell segmentation. Instead, cell-types are identified by clustering of molecular signatures found using density estimations. To reach a broader audience, the tool was further extended into the more efficient and easy-to-use SSAM-lite [9]. Prabhakaran et al [10] proposed an iterative method for assigning markers to nearby nuclei. TACCO, by Mages et al [11]., solves an optimal transport problem to infer annotations from single cell RNA sequencing data to the IST data. Petukhov et al. provide an extensive tool for segmenting mRNA markers into cell instances [12]. Likewise, the method by He et al. [13] identifies biologically meaningful structures, such as cells, cell bodies, and tissue regions by exploiting the density and composition of mRNA markers. Similarly, Partel et al. [14] proposed an unsupervised clustering technique to automatically define biological compartments within a mouse brain. Tengt et al. [15] also provides a probabilistic model. Here, cells are clustered into semantic groups corresponding to various cell types. SpaceX, by Acharyya [16] et al., is a Bayesian model for identifying co-expression in IST data. Recently, graph neural networks have been employed to learn meaningful representations of molecular neighborhoods [17, 18, 19]. Lastly, there are extensive toolboxes available in Python and R for the assessment of IST data [20, 21, 22].

Although there are plenty of useful tools, we noticed two strands of limitations:

Firstly, most tools are directed towards people with computational experience (with the exception here being SSAM-lite [9]). Large data-volumes also put a demand on specialized software and hardware. Secondly, most available tools do not provide solutions for interactive exploration of the data. Visually exploring the results, such as identified cells, cell-types, molecular patterns and regions, in their spatial setting within the tissue is fundamental for both constructing and validating hypotheses. Here, the richness of information and complexity of the data makes a practical challenge: There is simply too much information at hand to be practically explored in a static visualization. Moreover, biological patterns are often available at different scales. For example, tissue-level structures are explored on a millimeter-scale, whereas cellular and sub-cellular mRNA patterns are explored on a micrometer-scale. Practical interactive solutions are therefore essential for exploring the data.

1.2. Our contribution

We propose a simple interactive tool that allows users to automatically cluster IST data into coherent regions with similar mRNA compositions. The clustering is fast, memory efficient, and does not require any specialized hardware nor knowledge of languages such as Python or R. The clustering can be performed at various scales,

* These authors contributed equally to this work

enabling automatic identification of large-scale tissue structures, as well as identification of cell types. Crucially, the clustering is done interactively within the software allowing the user to immediately explore the resulting cluster alongside mRNA markers and Gigapixel-sized tissue images. The tool can be found at `https://tissuumaps.github.io/TissUUmaps/plugins/` - Points2Regions.

2. METHODOLOGY

We chose to implement our interactive clustering method as a Python plugin in the open-source software TissUUmaps 3 [23] as it allows us to efficiently display large quantities of markers and regions on-top of Gigapixel sized images. Moreover, TissUUmaps 3 is easy to install, and comes with a practical GUI, making it easy to use for non-programmer users. The first essential step of the method is to compute a set of descriptive composition vectors (CVs), describing the composition of mRNA markers within small local regions. To do so, we start off by introducing our set of mRNA markers as $\mathcal{S}$, where N is the total number of markers. Each marker is associated with two attributes: A spatial position $r_i \in \mathbb{R}^D$ and a one-hot-encoded label $e_i \in \{0,1\}^M$ indicating the type of the mRNA marker (M being the total number of unique mRNA markers). Here, D refers to the number of spatial dimensions, usually 2 or 3. To simplify our notation, we will for now assume that D = 1 although our method supports higher dimensionalities as well. Composition vectors at different query locations, r', can thus be extracted by evaluating the function:

$$x(r') = \sum_{(r_j,e_j)\in\mathcal{S}} h(\|r' - r_j\|_2)e_j. \tag{1}$$

Here, $h(\cdot)$ is a weighting function depending on the distance between the query location of the CV and the surrounding mRNA markers. The function h varies between implementations, but common choices are Gaussian functions [8, 9], Boxcar functions [14], or functions proportional to the inverse distance [10]. We chose to work with a Gaussian function (truncated at three σ). To quickly evaluate Eq 1 we employ a discretization trick similar to that of [9]. In brief, the position of each observed mRNA molecule is approximated onto a regular grid with K number of bins, each bin separated by a distance Δ. Specifically, $r_j \approx k_j\Delta$ where $k_j = \lfloor \frac{r_j}{\Delta} \rceil$ is the index of the bin. With this approximation, Eq 1 can be conveniently evaluated at every bin location using standard linear algebra. Specifically, the CV at every grid point, $X = [x_1, x_2, \ldots, x_K]^T$, is given by the matrix product

$$X = HAE, \tag{2}$$

where $H \in \mathbb{R}_+^{K\times K}$ denotes a discrete convolution with a Gaussian filter (standard deviation set to $\frac{\Delta}{\sigma}$), $A \in \{0,1\}^{K\times N}$ is the bin-assignment matrix, indicating which mRNA marker falls into what bin, and $E = [e_1, e_2, \ldots, e_N]^T$ is the $N \times M$ matrix indicating the type of each mRNA-marker. Importantly, the above matrices are sparse, and the product is best computed using sparse matrix data structures. CVs computed from regions with low number of detected mRNAs are sub-sequentially filtered out based on their L_1 norm. Regions with similar composition of mRNA markers are thereafter obtained by clustering the CVs. We use a mini-batch KMeans algorithm, available in the Python library scikit-learn, for clustering. This algorithm was chosen due to its computational efficiency and scalability to very large sample sizes. The clustered CVs are finally labeled using connected component labeling based on spatial proximity. Each connected component is converted into a set of polygons and visualized as regions in TissUUmaps 3. The regions can be exported as polygons in geoJSON format.

3. RESULTS

We test our method on an osmFISH dataset from a mouse brain somatosensory cortex [4]. We compare Points2Regions clusters to other methods and qualitatively evaluate the visual differences. The dataset consists of roughly 1,8M markers. Running Points2Regions takes approximately 5 seconds until regions are displayed (on a laptop computer with Intel(R) Core(TM) i7-7500U CPU @ 2.70GHz processor). Figure 1 shows the clustering results of our method and the comparison to other techniques. In subfigure 1**a**, we present a visual quality assessment of the clustering. We compare our proposed Points2Regions to two other methods: The graph convolutional neural network powered method spage2vec [17], and the original clustering published together with the data by Codeluppi et al. [4]. We can see that in the case of the first cropped patch, Codeluppi performs very well, apart from the fact that it removes all the markers outside the nuclei since it considers only markers on top of the segmentation mask created from the DAPI image (stained nuclei). That is also the reason why the markers in the clusters are so consistent. Similarly, spage2vec performs well in the nuclei areas, however, the clusters are very noisy in the cytoplasm regions. Points2Regions clustering results are comparable with the Codeluppi results, although, only using IST marker data. In the cropped patch number two, we can see that Codeluppi excludes a lot of IST markers (a pink region called Excluded). This is because the quality of the DAPI image makes it difficult to segment the cell nuclei correctly. Again, spage2vec results are noisy and do not result in distinguished cells, while Points2Regions finds several cluster regions containing a number of individual cells. In the third cropped patch we can see that all markers were excluded when using the Codeluppi method because of the DAPI quality, while spage2vec could find several individual clusters. Also Points2Regions finds individual clusters.

The correspondence between clusters obtained using the separate methods is further shown as contingency tables in subfigure 1**b**. For example, the element n_{ij} in the first heatmap corresponds to the total number of mRNA-markers that are clustered into the respective clusters i and j by two separate methods Points2Region and spage2vec. The tables are normalized row-wise to enhance contrast and to display how the methods correlate. A perfect correspondence between the methods would lead to a bright diagonal stripe (value 1) and the rest would be dark (value 0).

All these examples have been cropped from a larger dataset to illustrated different aspects of the data. The full dataset with classification results from all the mentioned techniques can be found here: `https://tissuumaps.scilifelab.se/osmFISH.tmap?path=private/Andrea/osmFISH`

Here, the clustering was performed using a bin width of $\Delta = \frac{R}{4}$ and $\sigma = 2\Delta$, where R is approximately the radius of a cell nucleus. With these parameters the clustering would reveal structures on a cellular scale. However, Points2Regions can also be used to define larger-scale structures. For example, we run Point2Regions on the data generated from an in situ transcriptomics experiment on a mouse brain coronal section [14]. The clustering was done with $\Delta = 42\mu$m and $\sigma = 84$ μm. We used the same number of clusters as in the original manuscript [14]. The identified regions, along with the regions from the original manuscripts, are shown in Figure 2 together with a contingency table. The clear diagonal pattern in Figure 2 b indicate that the clustering results of the two methods are strongly correlated, and thus Points2Regions can also be used to

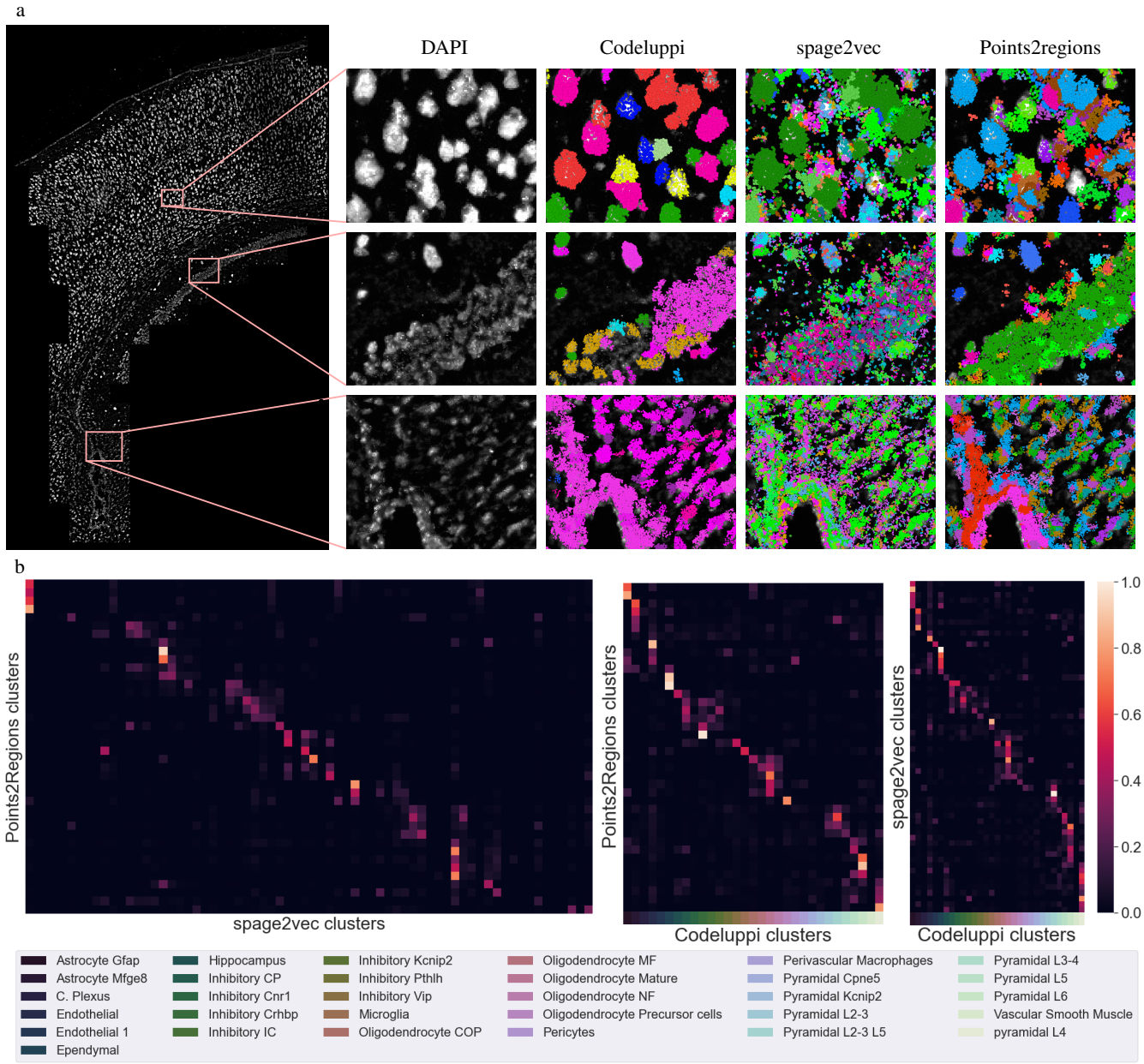

Fig. 1. Clustering results of different methods applied on osmFISH data from the mouse brain somatosensory cortex. **a** Tissue sections with cropped-out areas show the clustering results of three different techniques; Codeluppi, spage2vec, and Points2Regions. **b** Correspondence between clustering methods is shown as contingency tables. Each element in the table indicates the total number of mRNA markers that were clustered into different clusters by the different methods. The tables are normalized row-wise to enhance contrast. The first table is comparing the Points2Regions clusters with the spage2vec clusters. The second table is comparing the Points2Regions clusters with the Codeluppi clusters (marked with different colors) and the third table is comparing the spage2vec clusters with the Codeluppi clusters (marked with different colors) and cell-type annotations. The diagonal pattern in the tables indicate that the identified clusters are correlated.

define biologically relevant regions on a compartment level.

4. DISCUSSION

We have presented an interactive clustering method for in situ transcriptomics packaged as a plugin in TissUUmaps 3. Codeluppi clustering has very neat results, however, it is dependent on the DAPI stain and excludes points that are not associated with a cell nucleus, limiting analysis to those cells that are fully captured in the tissue section. Spage2vec, on the other hand, does not rely on any auxiliary nuclei segmentation, but instead leverages graph neural network to learn meaningful representations of localized mRNA markers, which in turn are used to identify cell-types. Althoguh spage2vec can learn meaningful representations, trainig it on a large dataset (such as osm-

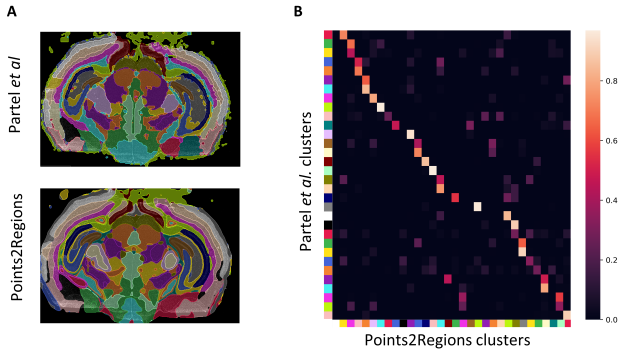

Fig. 2. **A** Anatomical compartments automatically defined using two methods: Partel et al. [14], as well as the Points2Regions. **B** Correspondence between the two clustering methods is shown as a contingency table. Entry n_{ij} in the table indicate the total number of mRNA markers that were clustered into cluster i by Points2Region and cluster j Partel et al. [17]. The table is normalized row-wise to enhance contrast.

FISH data presented here) takes hours and may require specialized hardware. On the other hand, Points2Regions does not rely on any DAPI image, only the markers themselves, and still provides consistent clusters almost instantly. The very low computation time of Points2Regions has allowed us to implement in as an interactive tool, suitable for users without programming knowledge. The user can select the number of clusters, and the size of the neighborhood and get the results instantly, and importantly explore the results directly on top of images of the tissue. This allows for freedom in terms of testing several values for a number of clusters and neighborhood sizes.

5. COMPLIANCE WITH ETHICAL STANDARDS

This is a numerical simulation study based on previously published, publicly available data.

6. CONFLICTS OF INTEREST

This research was funded by the European Research Council via ERC Consolidator grant CoG 682810 to C.W. and the SciLifeLab BioImage Informatics Facility. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

7. REFERENCES

[1] Sheel Shah et al., "In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus," *Neuron*, vol. 92, no. 2, pp. 342–357, oct 2016.

[2] Xiao Wang et al., "Three-dimensional intact-tissue sequencing of single-cell transcriptional states," *Science*, vol. 361, no. 6400, July 2018.

[3] Jeffrey R. Moffitt et al., "Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region," *Science*, vol. 362, no. 6416, Nov. 2018.

[4] Simone Codeluppi et al., "Spatial organization of the somatosensory cortex revealed by osmFISH," *Nature Methods*, vol. 15, no. 11, pp. 932–935, Oct. 2018.

[5] Chee-Huat Linus Eng et al., "Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH," *Nature*, vol. 568, no. 7751, pp. 235–239, Mar. 2019.

[6] Hower Lee et al., "Direct rna targeted in situ sequencing for transcriptomic profiling in tissue," *Scientific reports*, vol. 12, no. 1, pp. 1–9, 2022.

[7] Amanda Janesick et al., "High resolution mapping of the breast cancer tumor microenvironment using integrated single cell, spatial and in situ analysis of ffpe tissue," *bioRxiv*, 2022.

[8] Jeongbin Park et al., "Cell segmentation-free inference of cell types from in situ transcriptomics data," *Nature Communications*, vol. 12, no. 1, June 2021.

[9] Sebastian Tiesmeyer et al., "SSAM-lite: A light-weight web app for rapid analysis of spatially resolved transcriptomics data," *Frontiers in Genetics*, vol. 13, Feb. 2022.

[10] Sandhya Prabhakaran, "Sparcle: assigning transcripts to cells in multiplexed images," *Bioinformatics Advances*, vol. 2, no. 1, pp. vbac048, 2022.

[11] Simon Mages et al., "TACCO: Unified annotation transfer and decomposition of cell identities for single-cell and spatial omics," Oct. 2022.

[12] Viktor Petukhov et al., "Bayesian segmentation of spatially resolved transcriptomics data," Oct. 2020.

[13] Yichun He et al., "ClusterMap for multi-scale clustering analysis of spatial gene expression," *Nature Communications*, vol. 12, no. 1, Oct. 2021.

[14] Gabriele Partel et al., "Automated identification of the mouse brain's spatial compartments from in situ sequencing data," *BMC biology*, vol. 18, no. 1, pp. 1–14, 2020.

[15] Haotian Teng et al., "Clustering spatial transcriptomics data," *Bioinformatics*, vol. 38, no. 4, pp. 997–1004, 2022.

[16] Satwik Acharyya et al., "Spacex: Gene co-expression network estimation for spatial transcriptomics," *bioRxiv*, 2021.

[17] Gabriele Partel et al., "Spage2vec: Unsupervised representation of localized spatial gene expression signatures," *The FEBS Journal*, vol. 288, no. 6, pp. 1859–1870, Oct. 2020.

[18] Jiachen Li et al., "Ccst: Cell clustering for spatial transcriptomics data with graph neural network," 2021.

[19] Jian Hu et al., "Spagcn: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network," *Nature methods*, vol. 18, no. 11, pp. 1342–1351, 2021.

[20] Giovanni Palla et al., "Squidpy: a scalable framework for spatial omics analysis," *Nature methods*, vol. 19, no. 2, pp. 171–178, 2022.

[21] Alexander F. Wolf et al., "Scanpy: large-scale single-cell gene expression data analysis," *Genome biology*, vol. 19, no. 1, pp. 1–5, 2018.

[22] Yuhan Hao et al., "Integrated analysis of multimodal single-cell data," *Cell*, vol. 184, no. 13, pp. 3573–3587, 2021.

[23] Nicolas Pielawski et al., "Tissuumaps 3: Interactive visualization and quality assessment of large-scale spatial omics data," *bioRxiv*, 2022.