Title: 2D, or not 2D? Investigating Vertical Signal Integrity of Tissue Slices

Authors and affiliations

Sebastian Tiesmeyer[1,2], Niklas Müller-Bötticher[1,2], Alexander Malt[1,2], Brian Long[3], Sergio Marco-Salas[4], Paul Kiessling[5], Paul Horn[6], Adrien Guillot[6], Louis B Kuemmerle [7,8], Leyao Ma[2], Frank Tacke[6], Fabian Theis[7], Christoph Kuppe[5], Mats Nillson[4], Roland Eils[1,2,10], Naveed Ishaque[1]

[1] Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Center of Digital Health, Berlin, Germany

[2] Freie Universität Berlin, Department of Mathematics and Computer Science, Berlin, Germany

[3] Allen Institute for Brain Science, Seattle, WA 98109, USA.

[4] Science for Life Laboratory, Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden

[5] Department of Nephrology, Rheumatology, and Clinical Immunology, University Hospital RWTH Aachen, Aachen, Germany.

[6] Department of Hepatology & Gastroenterology, Charité - Universitätsmedizin Berlin, Campus Virchow-Klinikum and Campus Charité Mitte, Berlin, Germany.

[7] Institute of Computational Biology, Computational Health Center, Helmholtz Munich

[8] Institute for Tissue Engineering and Regenerative Medicine (iTERM), Helmholtz Zentrum München, Munich, Germany

[9] Department of Nephrology, Rheumatology, and Clinical Immunology, University Hospital RWTH Aachen, Aachen, Germany.

[10] Health Data Science Unit, Heidelberg University Hospital and BioQuant, University of Heidelberg, Germany

Correspondence

Correspondence should be sent to: naveed.ishaque@bih-charite.de

Keywords

Spatial transcriptomics, ISH, ISS, spatial gene expression, 3D, spatial biology, imaging-based spatial transcriptomics, cell segmentation, bioinformatics, doublets, quality control

Abstract

Imaging-based spatially resolved transcriptomics can localise transcripts within cells in 3D. Cell segmentation precedes assignment of transcripts to cells and annotation of cell function. However, cell segmentation is usually performed in 2D, thus unable to deal with spatial doublets arising from overlapping cells, resulting in segmented cells containing transcripts originating from multiple cell-types. Here we present a computational tool called ovrlpy that identifies overlapping cells, tissue folds and inaccurate cell-segmentation.

Main

Spatially resolved transcriptomics (SRT) has revolutionized the study of cellular architecture within tissues[1]. However, despite profiling 5–15 µm thick tissue sections, most available SRT studies ignore the vertical dimension and interpret their data in a completely flat 2D coordinate space[2–4]. This common practice overlooks the vertical complexity inherent in the tissue, potentially conflating signals from overlapping cells and generating multi-cell 'doublet' artifacts[5]. These spatial doublets may result from biological cell associations or technical artifacts, such as tissue folding or smearing, posing challenges in accurate cell-typing and downstream analysis[6,7]. While sequencing-based SRT methods that involved tissue dissociation measure gene expression in 2D, imaging-based SRT methods can resolve transcript locations in 3D depending on the image acquisition device[8].

We developed geometric noise models to explore the impact of collapsing 3D spatial information into 2D (**Figure 1, Supplementary Figure 1**). By simulating spherical cells within dense tissue sections of various thicknesses, we calculated the "signal-to-total" ratio, representing the proportion of signal originating from the target cell versus vertically adjacent structures. For a 10 µm diameter cell in a 10 µm slice section, the model predicts an average signal-to-total ratio of 0.4, while a thinner 5 µm section raises the signal-to-total to 0.57. These relatively low ratios question the practice of treating imaging-based SRT data as 2D, instead suggesting that SRT signal will routinely be contaminated by its vertical environment.

To explore the 3D-to-2D projection effect in tissue sections, we virtually sub-sliced a 10 µm thick 3D mouse brain coronal section dataset profiled on the Xenium platform[9]. Based on their local vertical position, transcripts were divided into 'top' and 'bottom' virtual sub-slices (**Figure 1d**). We annotated the cell segments in the top and bottom sub-slices independently using MapMyCells[10]. Across 162,033 segmented cells, only 79% of cells were consistently assigned the same cell type between top and bottom slices (**Supplementary Figures 2**). This was in line with the other Xenium replicates (82% and 82%). Notably, the top-bottom mismatches implicated that glial cells like astrocytes, oligodendrocytes and vascular cells are especially prone to doublet formation (**Supplementary Figures 3a**) , consistent with the notion that glia act as architects of the central nervous system[11]. We further investigated the hippocampal CA1 region, which is a homogenous, well-structured tissue section consisting mostly of a dense matrix of glutamatergic excitatory neurons interspersed with sparse GABAergic inhibitory neurons and glial cells[12] (**Figure 1b**). Among 182 detected inhibitory/glial cells, less than 25% showed consistent cell-type assignment between top and bottom sub-slices, while the majority formed spatial doublets with neighbouring glutamatergic neurons (**Figure 1c, Supplementary Figure 3a**).

To investigate the effect of tissue thickness on overlap abundance we then analysed a thinner 5 µm mouse brain coronal section profiled on the MERSCOPE platform. We observed a vertical assignment

consistency of 84% over the entire tissue, and 61% for non-neuronal cells in the hippocampal CA1 region (**Supplementary Figures 3-5**). The higher overall vertical consistency compared to the Xenium dataset of a thicker tissue section was apparent for other MERSCOPE replicates (81% and 88%). This matches our model's assumption that thinner tissue sections result in fewer observed cell overlaps, however, we noticed tears in the tissues, possibly due to the physical stresses of cutting thin sections (**Supplementary Figure 6**).

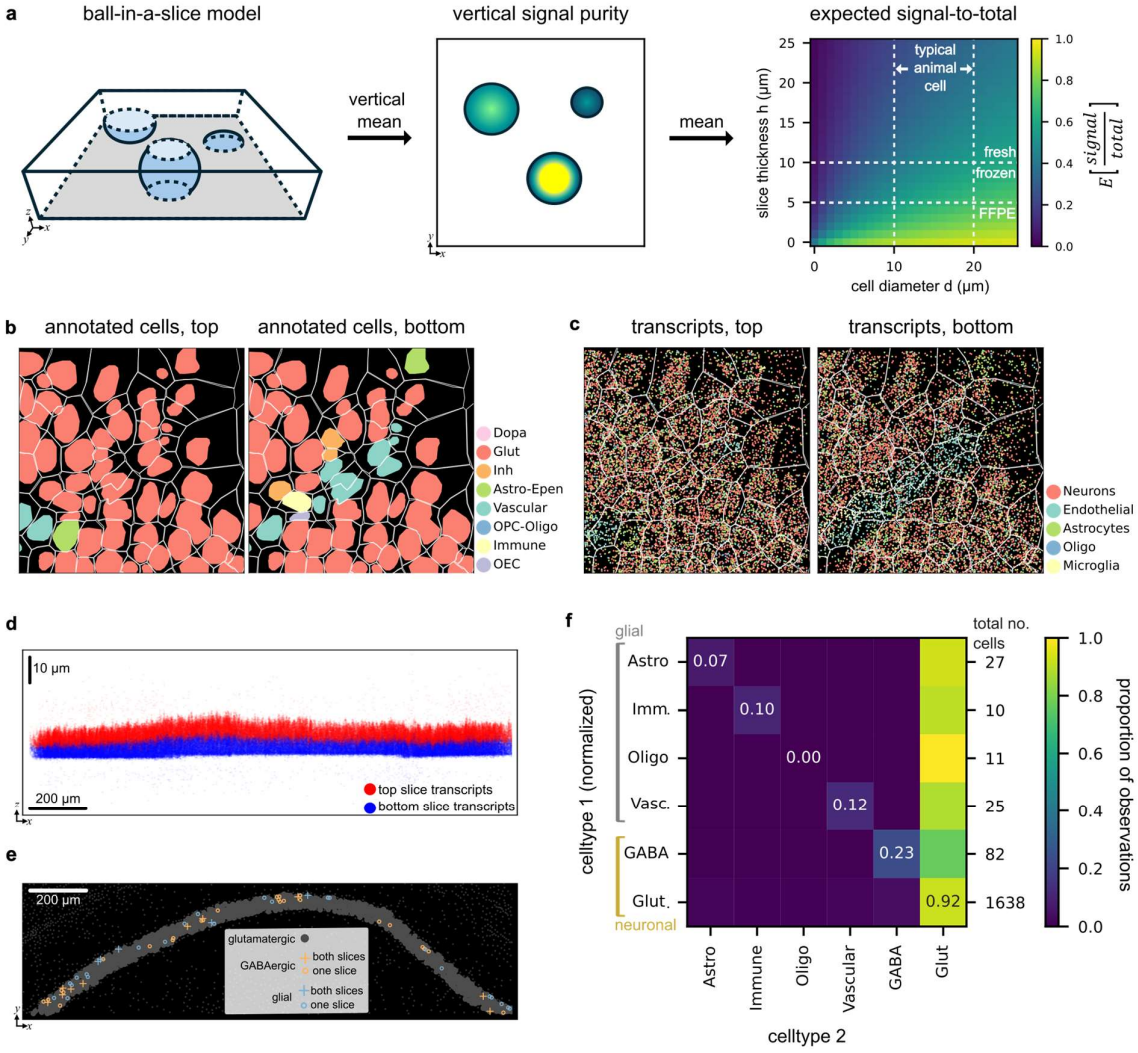

Figure 1: A model for vertical consistency in tissue sections. (A) Geometric model of signal integrity after 3D-to-2D collapse. (B) MapMyCell annotation of cell segments in the top and bottom virtual sub-slices in the CA1 region. Nuclei are colours, and white lines depict cell borders. (C). Transcript coloured by cell-type classes in the top and bottom virtual sub-slices corresponding to panel B. (D) Virtual top-bottom split of the hippocampus. (E) Consistency of 'GABA' and 'glial' assignment of segments across the vertical split. (F) Combinatorial matrix of observation pairs in virtual hippocampus sub-slices.

To address vertical spatial doublets in imaging-based SRT data generated from tissue sections, we implemented *ovrlpy*. Ovrlpy combines the vertical sub-slicing strategy with an unsupervised and segmentation-free SRT analysis algorithm[13,14]. Ovrlpy generates tissue maps of vertical signal integrity (VSI) by calculating local gene expression similarity between top and bottom virtual sub-slices. We established a 0.7 threshold for identifying spatial doublets, based on theoretical and practical validation (**Supplementary Figure 7a-b**). Ovrlpy can identify and visualise regions of interest (ROI) for further inspection (**Figure 2**). The transcriptome signal is embedded both in 2D-UMAP-space and in RGB-colour space. Subsequently, both 3D and 2D views of transcripts in the ROI are rendered.

An example application to the Xenium mouse brain data set of Figure 1 revealed different types of low-VSI artefacts (**Figure 2b-f, Supplementary Figure 8**). The unsupervised colour assignment of transcripts corresponded well to different cell-types (**Supplementary Figure 9**). Low-VSI peaks appear throughout the entire tissue implicating widespread cell overlaps, alongside systematically low VSI for tissue types like the meninges, choroid plexus and capillaries. Many of these observations were consistent with the previous segmentation-based analysis (**Supplementary Figure 10**). The ventral brain region on the bottom contains several prominent structures, which had low VSI and visual consistency with sample preparation tissue folds (**Supplementary Figure 8**). Furthermore, ovrlpy confirmed doublets involving glial and inhibitory cells in the CA1 (**Figure 2e**, **Supplementary Figure 11**). These observations were also found when applying ovrlpy to the MERSCOPE mouse brain dataset (**Supplementary Figure 12**).

To investigate the generalisability of ovrlpy across tissues, we analysed a publicly available mouse liver dataset profiled on the MERSCOPE platform **(Supplementary Figure 13,14)** and were able to generate unsupervised gene expression embeddings consistent with the liver cellular composition[8]. Patterning of the VSI map as well as instance visualizations suggest an increased tendency for liver endothelial cells and stromal cells to form spatial doublets with their immediate environment (typically hepatocytes as well as, in the case of endothelial cells, other building blocks of multilayered membranes, **Supplementary Figure 13,14**).

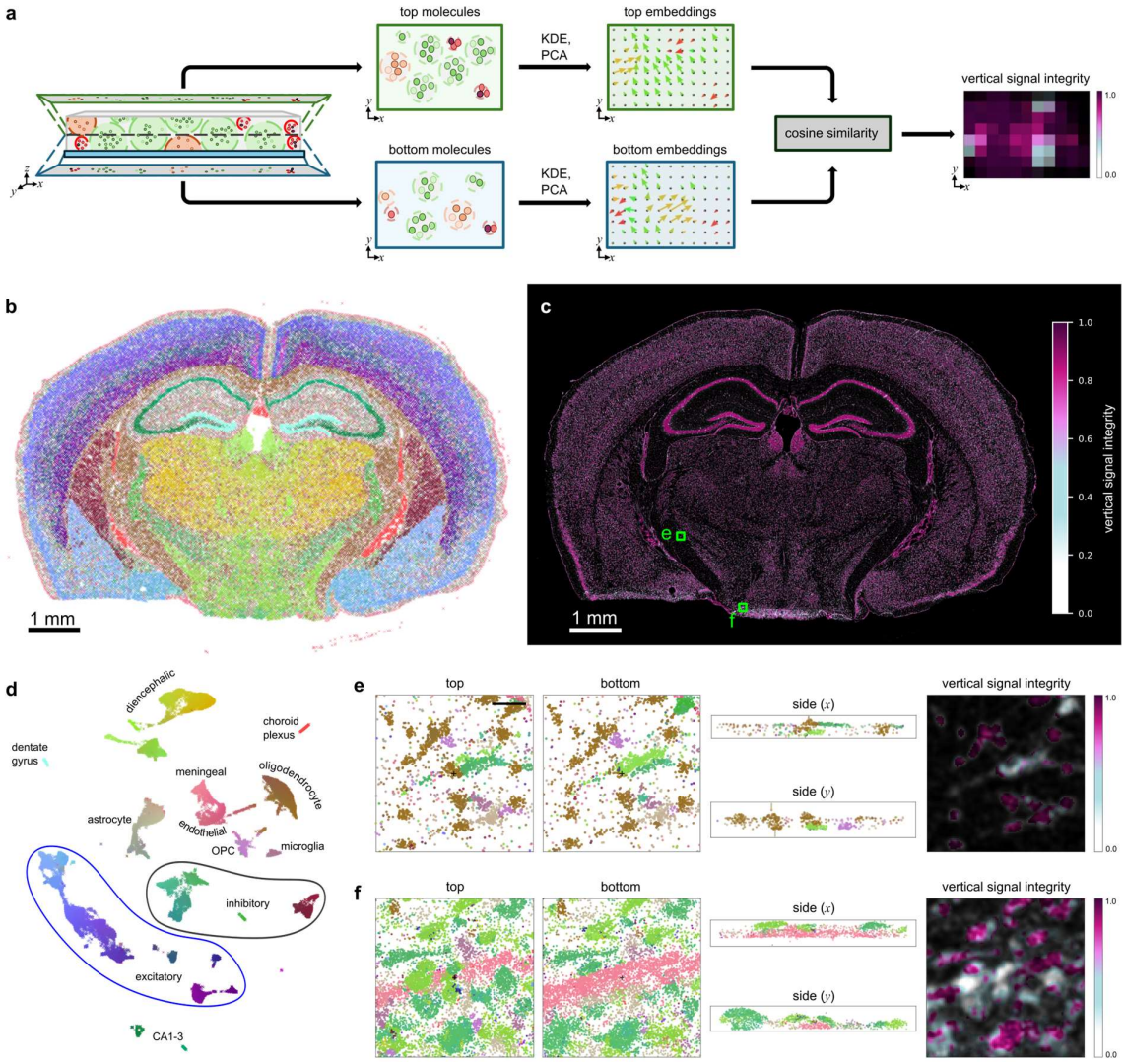

Figure 2: ovrlpy identifies overlapping cells in imaging-based spatially resolved transcriptomics data. (A) A rendition of the ovrlpy signal-incoherence pipeline. (B) Colour-embedded transcriptome signal at local maxima sampling locations. (C) Ovrlpy signal incoherence map. (D) Ovrlpy's segmentation-free, unsupervised transcriptome signal embedding. Scale bar corresponds to 20 µm. (E) Panels depicting of cell overlaps in the CA1 region. (F) Panels depicting of cell overlaps in the tissue fold. Scale is the same as panel E.

3D segmentation methods offer a potential solution to address spatial doublets mainly for imaging-based SRT data but are limited by technical challenges. Reduced microscopic resolution along the z-axis impedes precise boundary delineation, and cells partially outside the tissue slice pose difficulties for standard model priors as in the case of missing DAPI signal or inapt cell size constraints. To demonstrate ovrlpy's utility as a cell-segmentation quality control tool, we investigated two 3D segmentation datasets. First we performed 3D cell segmentation based on transcript locations of the Xenium mouse brain dataset using Baysor[15]. For the obtained segments, we observed a significant correlation of MapMyCells's reported cell typing correlation coefficient and ovrlpy's reported VSI score (**Supplementary Figure 15**). However, Baysor struggled to model the tissue fold artifact in the ventral region, failing to correctly segment a number of leptomeningeal cells along a stretch of the folded meninges in regions where the leptomeningeal markers were clearly present. We then

investigated a publicly available 3D DAPI segmentation dataset of the mouse brain from Vizgen (**Supplementary Figure 16)**. Here, only 7% of the 609 high confidence doublets (VSI<0.5) identified by ovrlpy were detected by Byasor, suggesting limitations in existing 3D cell segmentation models ability to identify vertically overlapping cells effectively. These findings underscore ovrlpy's sensitivity to spatial artifacts that may be missed by existing 3D cell segmentation methods and emphasize the urgent need for robust quality control strategies in SRT analysis.

We further investigated the effect of ovrly-defined doublets on downstream clustering analysis (**Supplementary Figure 17**). Mean VSI scores were determined across each cell segment, and segments with a mean VSI below 0.7 were marked as vertical doublets. We then annotated all cell segments and inspected their gene expression UMAP embeddings. We found that the doublets were embedded in between the more distinct singlets, as observed with single-cell RNAseq doublet gene expression embeddings[16]. Removing the vertical doublets resulted in an improved singal-to-noise ratio, demonstrated by a clearer separation of cell-types in both the Xenium and MERSCOPE datasets.

In conclusion, our analysis offers a practical solution to reduce the artefacts introduced by assuming SRT data to be 2D. Researchers should consider the frequent phenomenon of overlapping structures, spatial doublets and partly-out-of-slice cell fragments when planning experiments, interpreting results, or creating computational analysis tools. To the best of our knowledge, ovrlpy is the only tool to mitigate the limitations of 2D SRT analysis, offering both a data augmentation mechanism and an expanded toolkit for exploring VSI. Ovrlpy is an unsupervised algorithm thus making it suitable for both known and unknown tissues and cell types. The identified overlaps are plausible given cell-type specificity of transcripts and known cellular organisation within tissues. As SRT continues to evolve and is increasingly applied to attack real-world biological problems, tools like ovrlpy will be essential for ensuring the integrity and interpretability of high-resolution spatial data. Ovrlpy is compatible with standard python SRT analysis infrastructure, provides tutorials of example applications as well as guidelines for optimal model parameter selection and output interpretation.

References

1. Vandereyken, K., Sifrim, A., Thienpont, B. & Voet, T. Methods and applications for single-cell and spatial multi-omics. *Nat. Rev. Genet.* **24**, 494–515 (2023).
2. Chen, H., Li, D. & Bar-Joseph, Z. SCS: cell segmentation for high-resolution spatial transcriptomics. *Nat. Methods* **20**, 1237–1243 (2023).
3. Fang, S. *et al.* Computational Approaches and Challenges in Spatial Transcriptomics. *Genomics Proteomics Bioinformatics* **21**, 24–47 (2023).
4. Fu, X. *et al.* BIDCell: Biologically-informed self-supervised learning for segmentation of subcellular spatial transcriptomics data. *Nat. Commun.* **15**, 509 (2024).

5. Shu, J., Fu, H., Qiu, G., Kaye, P. & Ilyas, M. Segmenting overlapping cell nuclei in digital histopathology images. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Int. Conf.* **2013**, 5445–5448 (2013).

6. Palokangas, S., Selinummi, J. & Yli-Harja, O. Segmentation of folds in tissue section images. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Int. Conf.* **2007**, 5642–5645 (2007).

7. Taqi, S. A., Sami, S. A., Sami, L. B. & Zaki, S. A. A review of artifacts in histopathology. *J. Oral Maxillofac. Pathol. JOMFP* **22**, 279 (2018).

8. Moses, L. & Pachter, L. Museum of spatial transcriptomics. *Nat. Methods* **19**, 534–546 (2022).

9. Salas, S. M. *et al.* Optimizing Xenium In Situ data utility by quality assessment and best practice analysis workflow. Preprint at https://doi.org/10.1101/2023.02.13.528102 (2023).

10. Yao, Z. *et al.* A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. *Nature* **624**, 317–332 (2023).

11. Allen, N. J. & Lyons, D. A. Glia as architects of central nervous system formation and function. *Science* **362**, 181–185 (2018).

12. Megías, M., Emri, Z., Freund, T. F. & Gulyás, A. I. Total number and distribution of inhibitory and excitatory synapses on hippocampal CA1 pyramidal cells. *Neuroscience* **102**, 527–540 (2001).

13. Park, J. *et al.* Cell segmentation-free inference of cell types from in situ transcriptomics data. *Nat Commun* **12**, 3545 (2021).

14. Müller-Bötticher, N., Tiesmeyer, S., Eils, R. & Ishaque, N. Sainsc: A Computational Tool for Segmentation-Free Analysis of In Situ Capture Data. *Small Methods* **n/a**, 2401123.

15. Petukhov, V. *et al.* Cell segmentation in imaging-based spatial transcriptomics. *Nat. Biotechnol.* **40**, 345–354 (2022).

16. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst.* **8**, 281-291.e9 (2019).