Quantifying spatial resolution and assessing impact on tertiary analysis for ex-situ spatial transcriptomics technologies

Shaobo Zhang, Sujai Chari, Yao Xiao, Darren Segale, Kareem Ahmad, Asli Yildirim, Joel Sher, Samuel Clamons, Christopher Edlund, Cande Rogert, Emily Parker, James Han, Kevin Thai, Matteo Luberti, Mark Wang, Joachim Schmid

ABSTRACT

Spatial Transcriptomics enables a wide range of biological discoveries by capturing gene expression profiles across diverse tissue microenvironments. Common approaches to analyze the transcriptomics data include clustering, cell typing, cell to cell interaction, and pathway enrichment analysis. Achieving high performance with these methods requires technologies that combine high sensitivity along with fine spatial resolution. Illumina Spatial Technology (IST) delivers the combination of whole transcriptome coverage, high sensitivity, and cellular resolution over large sections of tissue samples enabling discoveries. IST achieves cellular resolution through the combination of high density surface features, accurate cell segmentation, and low diffusion in the assay.

Accurate cell segmentation requires identification of cells in tissue images as well as precise alignment of tissue images with transcript data. After using a diverse and curated custom dataset combined with expert annotation of fresh frozen tissue images for training, our AI model achieves high nuclei segmentation accuracy with F1 scores > 0.8. The accuracy of our fiducial based image alignment is validated using feature-based alignment of H&E image and spatial transcript data across various sub-regions where all sub-regions show alignment error of < 1 µm.

To address the lack of systematic methods available to assess spatial resolution across tissue types, we introduce a set of versatile tools for evaluating resolution. In addition to the aforementioned feature-based alignment, we also provide methods to measure nuclear/intronic transcript diffusion and cell marker gene diffusion measurement, which are applicable across tissue types. For example, using IST applied to mouse brain (coronal) tissue sections, we show high transcript localization by finding that up to 80% of the nuclear transcripts are located within the cell bins. Lastly, we highlight a diffusion-model based fractional transcript counting strategy that reveals more accurate and refined clustering results.

Using this toolbox, we also demonstrate that IST has superior resolution compared to other ex-situ technologies. Moreover, whole transcriptome combined with high spatial resolution enables IST to deliver biological findings beyond the capabilities of targeted panel in situ technologies.

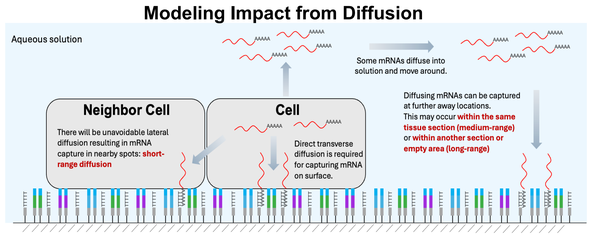
OVERVIEW

Noise sources impacting resolution

Spatial transcriptomics enables highly multiplexed, spatially-localized gene expression analysis from intact tissue. In this assay, tissue sections are placed on a spatially barcoded surface with mRNA capture oligos. Diffusion of mRNA, where transcripts are captured some distance away from their original location, can introduce noise that interferes with data interpretation.

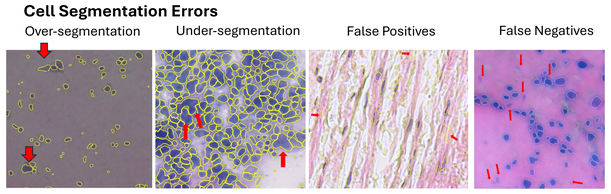
Depending on the distance between the transcript location and the final capture location, diffusion can be categorized as short-range, medium range and long-range diffusion:

Modeling Impact from Diffusion

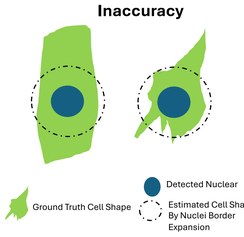


Additional sources of noise include errors in identifying nuclei and estimated cell borders using H&E images as shown below.

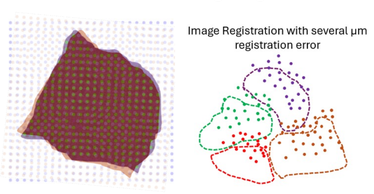
Cell Segmentation Errors



Cell Border Estimation Inaccuracy



Inaccurate Image Registration

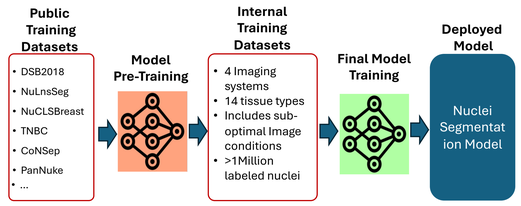


Accurate registration of both H&E image and the transcript data is necessary to ensure accurate assignment of transcripts to cells. The above example depicts the extent of transcript misassignment that could occur with just a few µm of image registration error.

Cell Segmentation Accuracy

Nuclei Identification

Multi-Stage Model Training Process

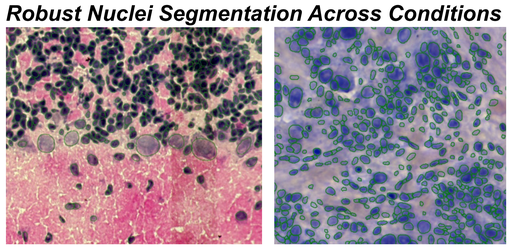


Continuous Model Performance Improvement

	Training Data	F1	Precision	Recall	Accuracy
Initial Model	Public dataset	0.123	0.161	0.08	0.140
V1 Model	Internal annotation dataset	0.60	0.48	0.79	0.42
V2 Model	Large annotation data dataset with optimized training strategy	0.83	0.82	0.84	0.71

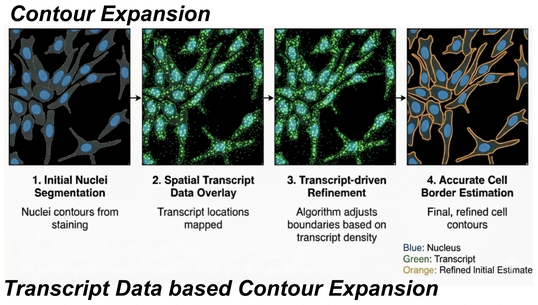
Upcoming V3 model will provide more accurate segmentation across more challenging nuclei shapes and complex cellular environments

Robust Nuclei Segmentation Across Conditions



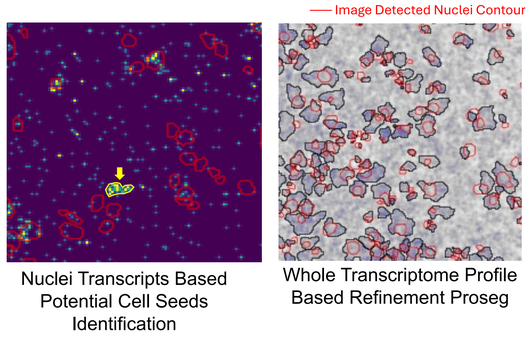
Cell Border Estimation

Contour Expansion



Transcript Data based Contour Expansion

1. Use transcripts data to identify potential cell seeds without nuclei staining
2. Use transcripts profile to smartly expand the nuclei border and estimate cell contour



Nuclei Transcripts Based Potential Cell Seeds Identification

Whole Transcriptome Profile Based Refinement Proseg

Image Registration

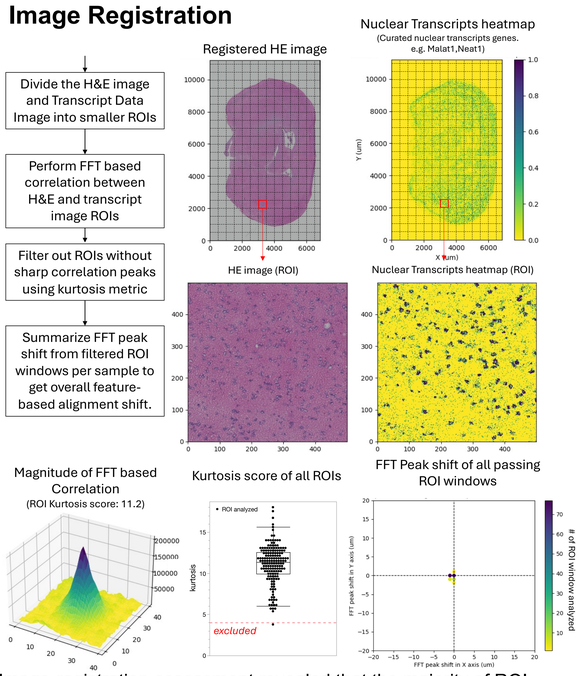
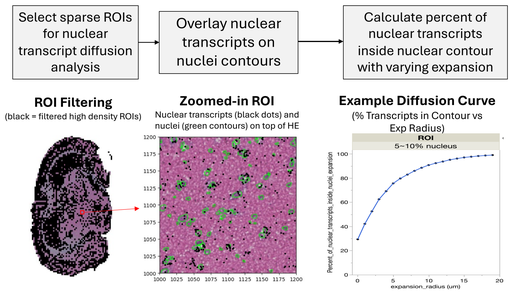


Image-registration assessment revealed that the majority of ROI windows exhibited <= 1um shifts in both X and Y dimensions, reflecting the superior image registration accuracy of Illumina Spatial Technology.

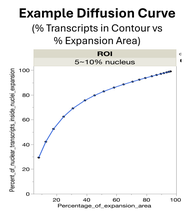
Quantifying Diffusion

Transcript Diffusion Measurement

Nuclear Transcripts Diffusion

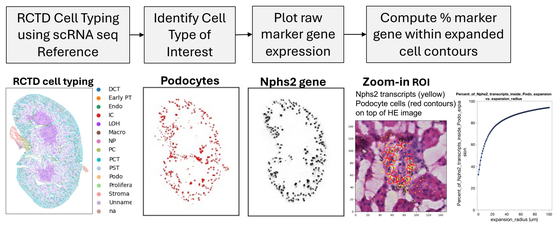


Measuring percent of nuclear transcripts contained within expanded nuclei contours as a function of expansion radius and percent expansion area enables the assessment of the impact of diffusion. Limiting the measurement to sparse ROIs ensures accurate transcript to cell assignment for this analysis.

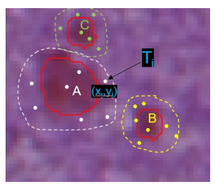


Marker Gene Transcripts Diffusion

The spatial expression of marker genes associated with specific cell types can be used to assess diffusion even in dense ROIs. Cell typing can be used to identify a cell type of interest. The percent of marker genes contained within expanded cell contours can be computed as a function of expansion radii.

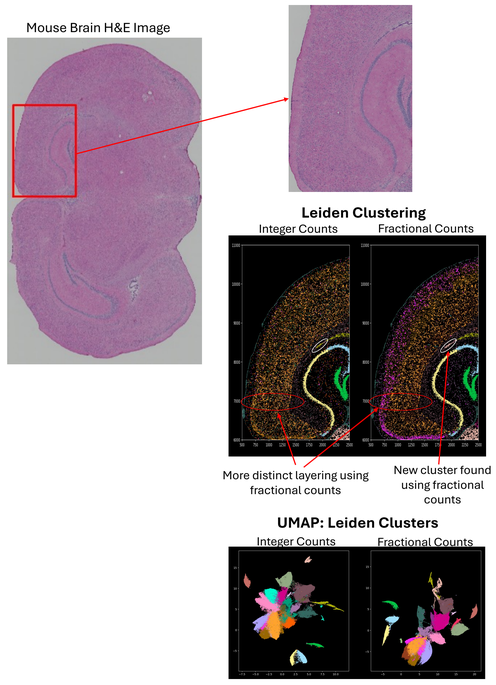


Fractional Transcript Assignment



T_i could be assigned fractionally to cell bins A,B, and C instead of solely being assigned to cell bin A

Instead of assigning a transcript completely to a particular cell based on capture location, a fractional assignment can be made to multiple cells to account for the possibility that the transcript may have diffused from other nearby cells. A diffusion model based fractional transcript assignment could lead to more statistically accurate transcript counts per cell bin.



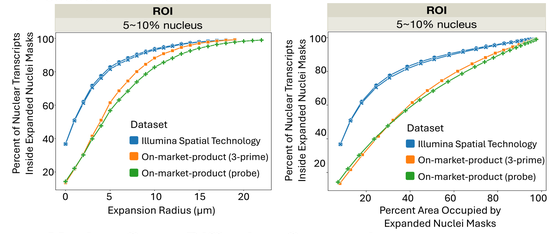
The use of fractional transcript assignment results in more refined Leiden clustering include more the layering in the cortex becoming more distinct. Additionally, the UMAP illustrates more separation among clusters using fractional counts.
The use of fractional transcript assignment to cell bins can be extended to transcripts that are captured even outside estimated cell areas. In this example, over 50% of the transcripts captured outside cell bins were still utilized in the computation of gene expression counts for cell bins.

RESULTS

Resolution Comparison across Spatial Technologies

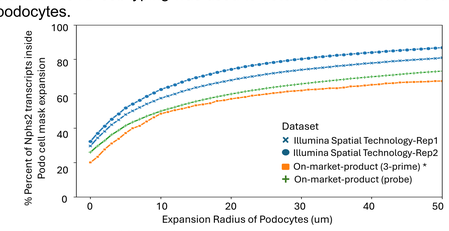
Nuclear Transcript Diffusion Comparison

All data is generated from fresh frozen mouse brain coronal sections.



Marker Gene Diffusion Comparison

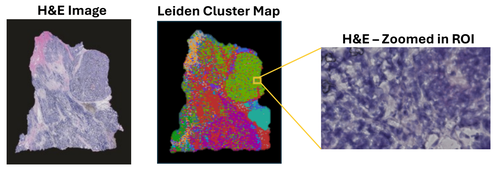
All data is generated from fresh frozen mouse kidney sections. Data for the Illumina Spatial Technology and for On-market probe based technology were obtained using serial sections from the same tissue blocks. RCTD cell typing was used to determine the location of the podocytes.

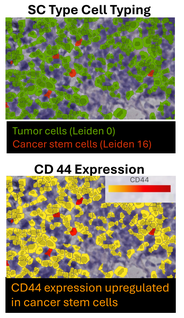


* Based on publicly available online dataset

Resolution Impact on Biological Findings

The combination of high spatial resolution and whole transcriptome coverage offered by IST enables biological findings that are challenging using targeted panel approaches. We demonstrate these benefits using a high grade human breast cancer sample.

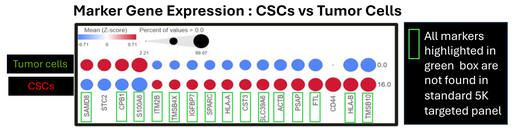




The marker gene based cell typing using SCType highlights the spatial resolution with the Cancer Stem Cells (CSCs) (Leiden cluster 16) being surrounded by tumor cells (Leiden cluster 0). The accuracy is validated by spatial expression of the upregulated marker CD44 in the cancer stem cells.

Furthermore, the marker gene dot plot below illustrates the benefit of whole transcriptome as many of the marker genes shown are not present in the commonly used 5K targeted panels.

Marker Gene Expression : CSCs vs Tumor Cells



CONCLUSIONS

We describe the various noise sources that impact the resolution of spatial transcriptomics data including diffusion of mRNA, cell segmentation errors, and image alignment errors.

Next, we describe a set of tools to quantify the accuracy of the various components that impact spatial resolution. Specifically, we describe a methodology to evaluate the accuracy of nuclei identification and demonstrate the benefits of using a custom trained AI model. We also describe approaches to improve the cell border estimation using the spatial transcript data.

For image alignment, we describe a workflow to evaluate the accuracy of the alignment of microscopy image with the transcript data using feature based alignment.

In order to assess the impact of diffusion, we introduce the monitoring of percent nuclear transcripts and marker genes found within nuclei/cell borders.

Further using prior knowledge of the diffusion model, we demonstrate more refined clustering on a mouse brain sample using cell binned data based on the notion of fractional transcript assignment.

Finally, we demonstrate using the aforementioned resolution tools that the combination of whole transcriptome and high spatial resolution enables Illumina Spatial Technology (IST) to provide superior resolution and increased biological findings compared to other spatial technologies.