#584 Multimodal cell segmentation for spatial in-situ imaging of tissues

Lidan Wu[1], Aster Wardhani[1], Patrick Danaher[1], Ruchir Bhatt[1], Joseph Phan[1], Youngmi Kim[1], Shanshan He[1], Zachary Lewis[1], Carl Brown[1], Rustem Khafizov[1], Dwayne Dunaway[1], Michael Rhodes[1], Joseph Beechem[1]

1. Nanostring Technologies, Seattle WA 98109

nanoString
NanoString Technologies
530 Fairview Avenue North, Seattle, WA 98109

Abstract

Spatial transcriptomics technologies, which produce single-cell gene expression data paired with cell locations, have opened a new frontier in biology. Accurate cell segmentation that assigns transcripts to cell locations is critical to data quality, but very challenging for tissue sections where cells are tightly packaged with shared, 3D boundaries and uneven morphology staining. To address this gap, we have developed a multimodal cell segmentation pipeline that automatically does **image preprocessing, machine-learning-augmented cell segmentation and transcript-based segmentation refinement**. We demonstrate our pipeline on spatial transcriptomics datasets from various FFPE tissue sections.

SMI is for research use only and not for use in diagnostic procedure.

Overview of multimodal cell segmentation pipeline

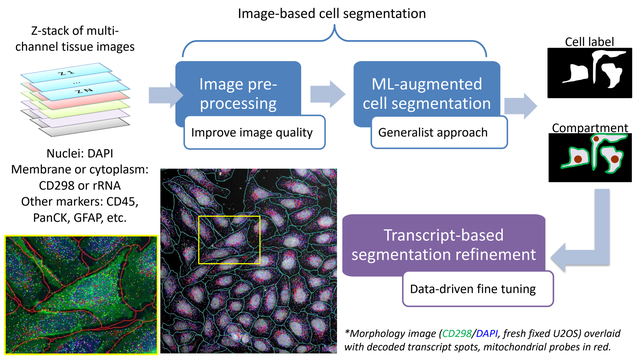
The need for universal pipeline despite the complexity and variability in tissue images

Optical variance:
- focus plane, 3D stacked cells;
- Illumination uniformity, high dynamic range of intensity

Sample variance:
- staining uniformity, loose vs. dense packed tissue,
- auto-fluorescence, necrosis

Morphology variance:
- Mixture of various size and shapes for cells in tissues

Spatial profile of transcripts
- Density? → might not work well for low plex
- Transcript content? → very slow



*Morphology image (CD298/DAPI, fresh fixed U2OS) overlaid with decoded transcript spots, mitochondrial probes in red.

Part I + II: Image-based cell segmentation

Firstly, our pipeline takes tissue images stained with both nuclear and membrane markers (DAPI, CD298/PanCK/CD45) to perform rescaling, normalization and boundary enhancement. The preprocessed images are fed into pre-trained Cellpose neural network models for both nuclear segmentation using nuclear channel only and cytoplasm segmentation using combined nuclear and membrane channels. Outputs of the two segmentation methods are combined by analyzing the intersection-over-union scores to: (a) mitigate issue of non-uniform staining in either channel; and (b) enable subcellular compartment analysis of the spatial transcriptomics data.

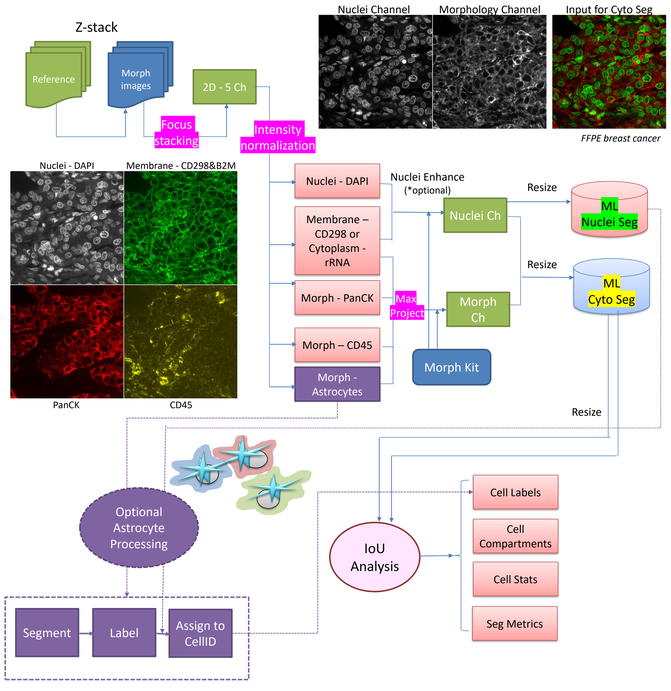
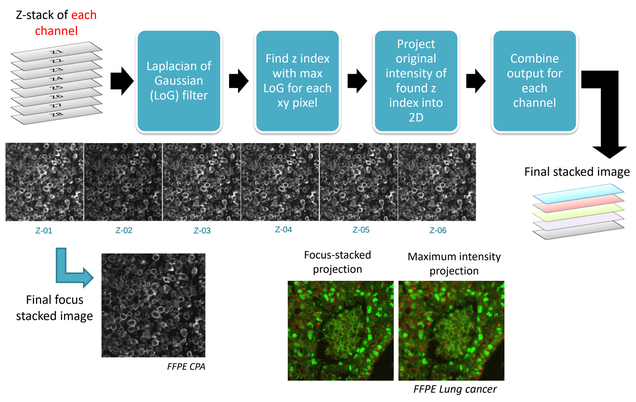


Image pre-processing improves the robustness of segmentation pipeline

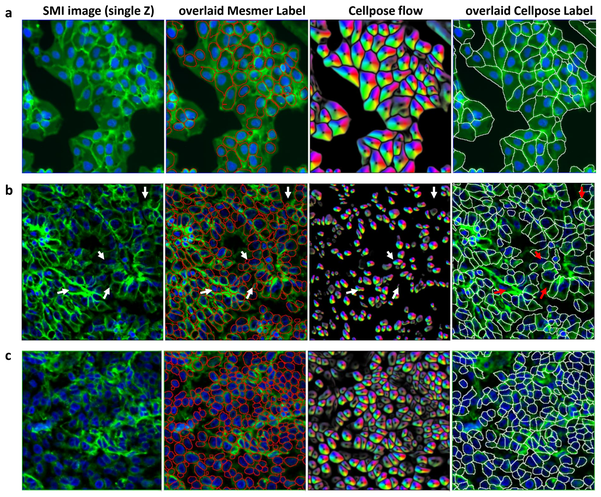
Focus stacking: bring objects in z-stack into same focus plane



Leverage publicly available pre-trained ML based cell segmentation module

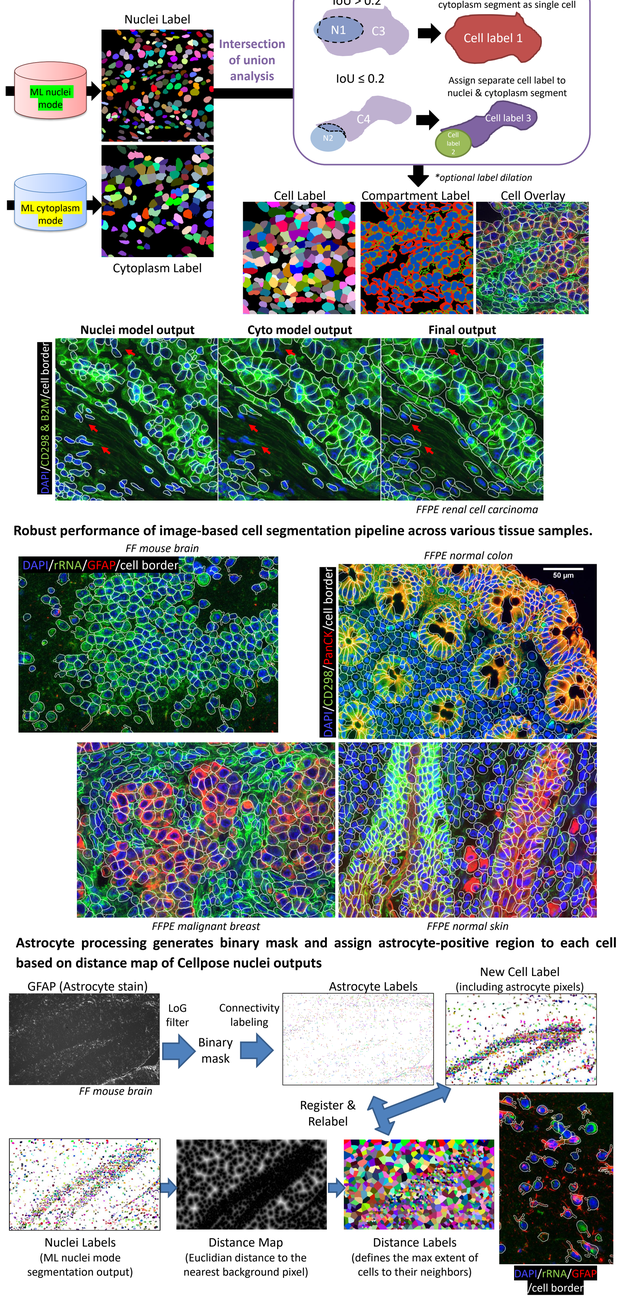
Cellpose, a generalist approach of neuron network-based cell segmentation[2,3]

- Cellpose constructs an intermediate representation that forms smooth topological basin and thus transforms noisy intensity distribution into a smooth one.
  - Ground-truth of manual annotation is transformed to a vector flow representation via simulated heat diffusion.
  - Train a neural network to predict spatial gradients (horizontal + vertical = vector fields)
  - Create binary mask for ROI via gradient tracking which route cell pixel to center.
- Cellpose pre-trained models are well trained on diverse datasets.
  - Nuclei dataset: fluorescence nuclei, H&E stain
  - Cytoplasm dataset: Cytoplasm stain, membrane stain, non-fluorescence cells, non-cell microscopy images, non-microscopy images.



Cell segmentation with pre-trained cellpose model (3rd column for the predicted vector field of spatial gradients, 4th column for the predicted cell boundaries in white overlay on top of input images) performs better than alternative ML-based cell segmentation pre-trained model, Mesmer[4] (2nd column for the predicted cell boundaries in red overlaid on top of input images) on cell images at single z-plane (1st column). Cell or tissue samples were stained with DAPI (blue) and anti-CD298 (green) to visualize the nuclei and plasma membrane. (a) Fresh fixed culture U2OS cells where cells have shared boundaries and pointed shapes; (b) FFPE kidney cancer section where cells have big variance in size, density and morphology within single field of view; (c) FFPE melanoma tissue section where cells were packed in 3D with many cells out of focus or have weak or blurry signal (arrows). In general, cellpose pre-trained model gives more accurate cell boundaries across various tissues, but tends to miss objects that are blurry, out-of-plane or of weak signal-to-noise ratio (SNR). Image preprocessing module that could bring objects in z-stack into focus plane and enhance SNR help reduce the false negative detection rate of cellpose model.

Combining outputs of multiple models into final cell segmentation results

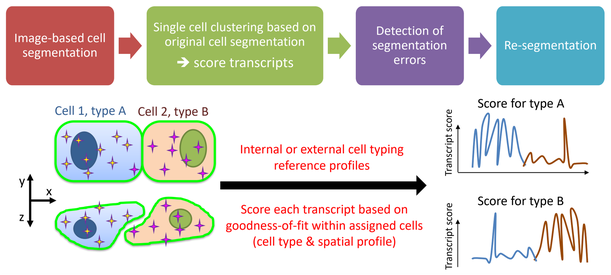


Robust performance of image-based cell segmentation pipeline across various tissue samples.

Astrocyte processing generates binary mask and assign astrocyte-positive region to each cell based on distance map of Cellpose nuclei outputs

Part III: Transcriptional profile-based segmentation refinement

Secondly, our pipeline exploits the spatial profile of transcripts on the same tissue section for rapid segmentation refinement. The image-based cell segmentation enables single-cell typing/clustering of transcript profiles. Individual transcripts are then scored for goodness-of-fit within their respective cells, based on the probability of each gene belonging to each cell type and the spatial dependency of transcript score. As confirmed by the membrane-stained images, cells with boundary errors at the junction of different cell types, exhibit strong spatial dependency in their transcript score profile and thus can be easily identified. Our pipeline further identifies the spatially connected groups of transcripts with low goodness-of-fit within incorrectly segmented cells. A set of heuristic rules on neighborhood cell typing and transcript number are then applied to the identified transcript groups to decide on the re-segmentation actions, like merging, splitting and trimming. The re-segmented results show no significant spatial dependency on transcript score of individual cells, suggesting the successful correction of poorly segmented cells.



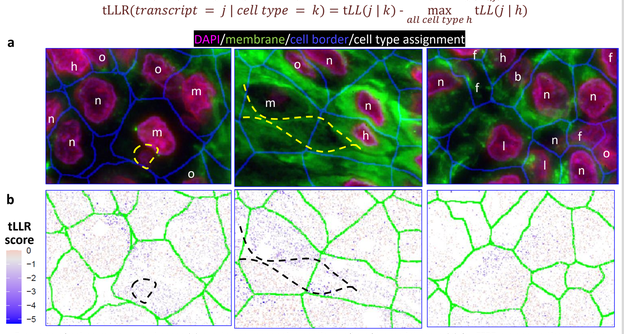
Identify segmentation boundary errors based on transcript profiles

Spatial pattern of transcript score (tLLR, log-likelihood ratio) indicates the presence of area from different cell types in same segment entity.

tLLR score, transcript log likelihood ratio: the difference between a transcript's log-likelihood under cell type of query and the cell type with highest log-likelihood across all cell types.

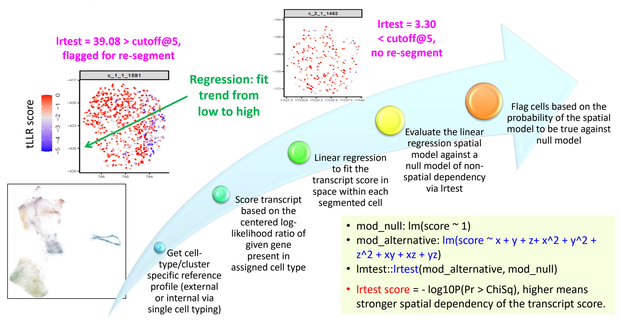
$$tLL(j \mid k) = loglik\ (transcript\ =\ j \mid cell\ type\ =\ k) = \log(\mu_{k,j})$$

$$\text{tLLR}(transcript\ =\ j \mid cell\ type\ =\ k) = tLL(j \mid k) - \max_{all\ cell\ type\ h} tLL(j \mid h)$$



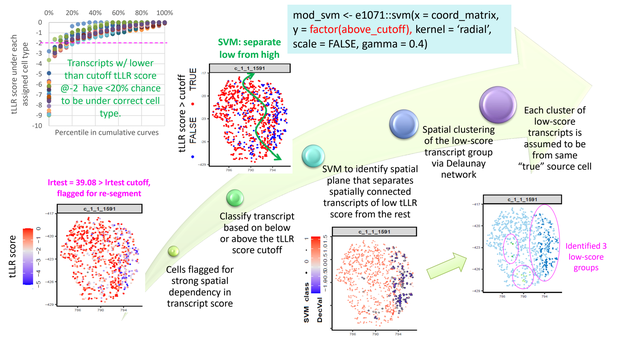
Examples of paired morphology images (a) and the corresponding spatial profile of transcript score (b) for FFPE melanoma tissue section stained for DAPI, anti-CD298 antibodies. Unsupervised cell typing/clustering (white fonts in a) was performed given the original cell segmentation via conventional thresholding method (blue in a and green in b).

Cells with boundary errors at the junction of different cell types would have spatial dependency on the cell type-specific score for transcripts of the query cell.



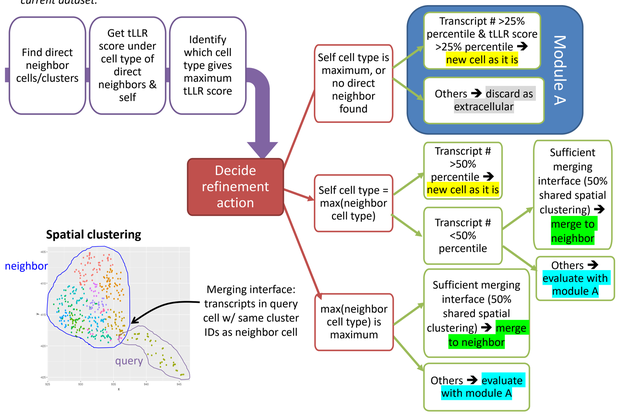
Segmentation refinement on low-score transcript groups

Splitting transcript groups based on cell-type-specific transcript score & spatial clustering

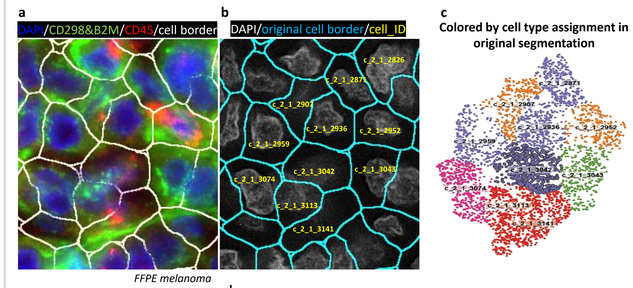


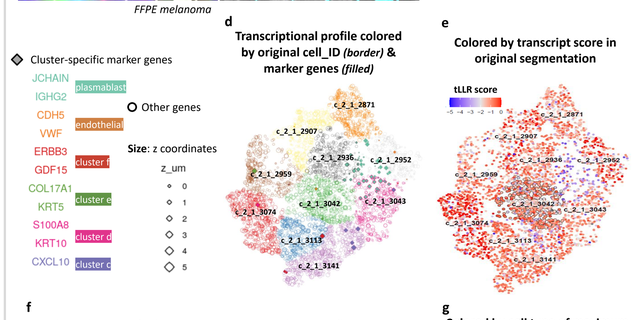
Evaluating neighborhood of each group of low-score transcripts to decide on segmentation refinement actions based on heuristic rules

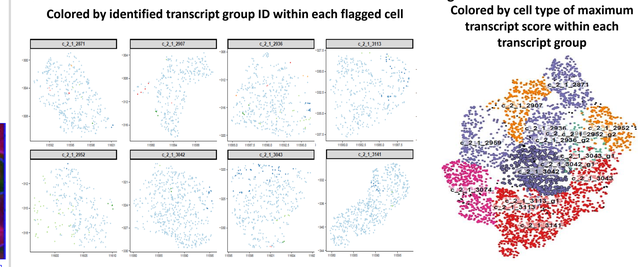
Define the cutoffs for transcript # & tLLR score based on the corresponding baseline distribution under each cell type in current dataset.

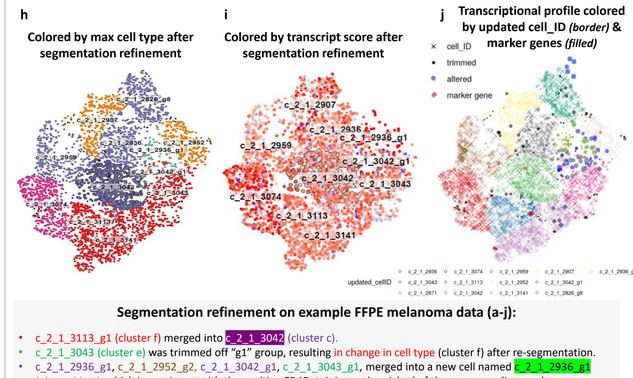


Examples of segmentation refinement outcomes







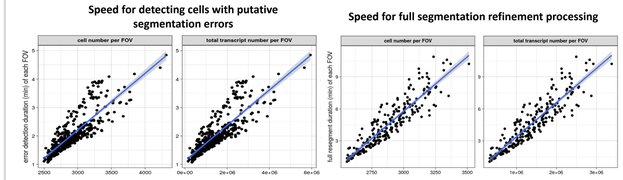


Segmentation refinement on example FFPE melanoma data (a-j):

- c_2_1_3113_g1 (cluster f) merged into c_2_1_3042 (cluster c).
- c_2_1_3043 (cluster e) was trimmed off "g1" group, resulting in change in cell type (cluster f) after re-segmentation.
- c_2_1_2936_g1, c_2_1_2952_g2, c_2_1_3042_g1, c_2_1_3043_g1, merged into a new cell named c_2_1_2936_g1 (plasmablast), which is consistent with the positive CD45 staining and enriched of the corresponding marker genes (IGHG2, JCHAIN).

A rapid algorithm to detect cell segmentation error based on transcriptional spatial profiles:

Processing speed on average PC: ~12.4hours to fully process 164 FOVs of 22GB transcript file size, ~ 75k cells, 230M transcripts.



On server, ~240,000 cells within 75 FOVs took 1.5hr to identify cell segmentation errors, and additional 2.5hr to complete cell segmentation refinement. Further optimization on parallel processing of multiple FOVs/files would further improve the processing speed.

Conclusions

Cell segmentation is critical for data quality in spatial transcriptomics. The pipeline presented here harnesses information from both **morphology images and mRNA locations** to generate an automated and robust cell segmentation algorithm across different tissue types. The technique is computationally tractable with > 1 million cells, making it viable for even the biggest spatial transcriptomics experiments.

Reference

1. He S, Bhatt R, Brown C, et al. High-plex Multiomic Analysis in FFPE at Subcellular Level by Spatial Molecular Imaging. *bioRxiv* 2021.11.03.467020. doi: 10.1101/2021.11.03.467020
2. Stringer C, Wang T, Michaelos M, Pachitariu M. Cellpose: a generalist algorithm for cellular segmentation. *Nat Methods*. 18, 100–106 (2021). doi: 10.1038/s41592-020-01018-x
3. Stringer C, Pachitariu M. Cellpose 2.0: how to train your own model. *bioRxiv* 2022.04.01.486764; doi: 10.1101/2022.04.01.486764
4. Greenwald NF, Miller G., Moen E. et al. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nat Biotechnol* 40, 555–565 (2022). doi: 10.1038/s41587-021-01094-0

Scan here to download or learn more