2DE: A PROBABILISTIC METHOD FOR DIFFERENTIAL EXPRESSION ACROSS NICHES IN SPATIAL TRANSCRIPTOMICS DATA

Nathan Levy[1] **Florian Ingelfinger**[1,2] **Artemii Bakulin**[1] **Giacomo Cinnirella**[1]

Pierre Boyeau[3] **Can Ergen**[3,4,*] **Nir Yosef**[1,3,4*]

[1] Department of Systems Immunology, Weizmann Institute of Science
[2]Department of Internal Medicine I, Medical Center-University of Freiburg
[3] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley
[4] Center for Computational Biology, University of California, Berkeley

ABSTRACT

Spatial transcriptomics enables studying cellular interactions by measuring gene expression in situ while preserving tissue context. Within tissues, distinct cellular niches define micro-environments that influence cell states and function. A fundamental task in spatial transcriptomics is identifying differentially expressed genes within a specific cell type across different niches to quantify context-dependent cell state variation. Despite advances in cell segmentation algorithms, the persisting problem of the wrong assignment of molecules to cells can obscure the analysis by introducing spurious differentially expressed genes that originate from neighboring cells rather than the group of interest. Here, we introduce 2DE, a probabilistic framework designed to refine spatial differential expression analyses by filtering out genes that are over-expressed due to local contamination rather than true cell-intrinsic expression. 2DE operates downstream of any differential expression method, filtering irrelevant genes by considering gene over-expression relative to the expression in the neighborhood and returning marker confidence scores. In a study of human breast cancer, we demonstrate that 2DE improves the precision of the discoveries. 2DE is available as open source software at YosefLab/2DE

1 INTRODUCTION

Single-cell Spatial Transcriptomics (ST) technologies provide a powerful means to study tissue organization by capturing the spatial location and transcriptomic profiles of individual cells (Bressan et al., 2023). Despite significant advances (Petukhov et al., 2022; Jones et al., 2024), molecular quantification and cell segmentation errors during ST protocols can lead to inaccurate expression estimates, such as the artificial co-expression of gene markers from different lineages (Ergen & Yosef, 2025). A key task in ST analysis is identifying differentially expressed genes within a specific cell type across distinct tissue environments to quantify context-dependent cell state variation (Chen et al., 2024). However, this task is confounded by the misassignment of molecules to cells. Here, we introduce 2DE, a probabilistic method designed to refine Differential Expression (DE) analyses in single-cell ST by filtering out gene signals originating from neighboring cells. Operating downstream of any DE method, 2DE enhances the precision of detected gene expression changes. We demonstrate its effectiveness by applying it to spatially confined endothelial populations in human breast cancer, showcasing its ability to improve the precision of discoveries across different DE approaches.

*correspondence to cergen@berkeley.edu, nir.yosef@weizmann.ac.il, contributed equally to this work.

2 THE 2DE METHOD

We consider a single-cell resolved spatial transcriptomics dataset, providing a $N \times G$ gene expression matrix, cell center coordinates, and cell type annotations $\boldsymbol{c}$. We assume that we have access to a Differential Expression (DE) method $\mathcal{M}$, which takes as input two groups of cells and returns Log-Fold Changes (LFCs). LFCs are ratios of expression levels between two groups, transformed into a logarithmic scale, where positive values indicate upregulation and negative values indicate downregulation. Considering two groups *G1* and *G2* corresponding to different spatial contexts (for instance, astrocytes in two brain regions), our goal is to determine which genes have reproducibly different expression levels between the two groups. The naive approach consists in applying $\mathcal{M}$ to $\{G1,\ G2\}$ only. However, empirically this approach returns spurious discoveries due to errors in molecule assignment leading to detection of genes expressed by neighboring cells. Therefore, we should also consider the neighborhood gene expression.
We introduce the group spatial neighborhoods *N1* and *N2*, which are the spatial nearest neighbors of a different type than the cells in *G1*. For instance, if *G1* consists of astrocytes, we want to exclude astrocyte neighbors from the neighborhood *N1*. We proceed in the same way for *N2*.
To formalize this, let us consider a tissue slide *b* consisting of N_b cells. We compute the slide adjacency matrix $\boldsymbol{A} \in [0,1]^{N_b \times N_b}$, by defining either a radius neighbors graph or a nearest neighbors graph. Then, we multiply this matrix with a cell type mask to get an adjusted matrix $A'_{ij} = A_{ij}\delta_{c_i \neq c_j}$ where $\delta_{c_i \neq c_j} = 1$ if $c_i \neq c_j$ and $\delta_{c_i = c_j} = 0$ if $c_i = c_j$ with c_i, c_j being the types of cells i and j, respectively. If the dataset consists of multiple slides, we construct a block diagonal adjacency matrix containing all cells N. We then extract the rows corresponding to the group of interest and gather all the non-zero indices appearing in these rows. For instance, we call this list I_1 for *G1*. Finally, we define $N1 = \{i | i \in I_1\}$ as the set of unique indices appearing in I_1.

To determine the upregulated genes of *G1 vs G2*, we compute DE between $\{G1,\ G2\}$, $\{N1,\ G2\}$ and $\{G1,\ N1\}$, using the method $\mathcal{M}$. The upregulated genes for $\{G1,\ N1\}$ define a set of local cell type markers, denoted $\mathcal{S}_1$. Conversely, if a gene is both higher expressed in *N1* compared to *G1* and *G1* compared to *G2*, it is likely that the increased expression in *G1* is spurious.
We argue that the probability of a gene being a *local marker* could be a relevant score to filter spurious genes. To compute this score, we considered the upregulation of a gene in one group relative to the upregulation in its neighborhood: a local marker g should verify

$$LFC^{\ g}_{G1\ vs\ G2} > LFC^{\ g}_{N1\ vs\ G2}, \tag{1}$$

which means that the signal comes from cells in *G1* rather than their neighbors *N1*.
We select genes for which $LFC_{G1\ vs\ G2} > 0$ and use the genes $\mathcal{S}_1$ as truely differentially expressed. We also define $\mathcal{N}_1 = \{g | LFC^{\ g}_{G1\ vs\ G2} > 0,\ g \notin \mathcal{S}_1\}$. We train a Gaussian process classifier on $\boldsymbol{X} = [LFC_{G1\ vs\ G2}\ ,\ LFC_{N1\ vs\ G2}]$ to classify between the *local markers* $\mathcal{S}_1$ and the *neighborhood genes* $\mathcal{N}_1$. Once fitted, the classifier returns a local marker probability $p_g = p(g \in \mathcal{S}_1 | \boldsymbol{X})$ for each gene g, that we can compare to a given threshold τ to filter the neighborhood genes.

3 2DE REVEALS NEO-VASCULARIZATION IN TUMOR REGIONS OF BREAST CANCER

We applied 2DE to an in-situ sequencing dataset of human breast cancer sections, generated with 10X Xenium (Janesick et al., 2023). We found the original cell segmentation to contain a high amount of erroneously assigned transcripts. We resegmented the cells using Proseg (Jones et al., 2024), which led to improved cell-type delineation (Figure 1B). Unless specified, we will report results from the Proseg-segmented data in the following.
We focused our analysis on endothelial cells, which can play crucial roles in the tumor microenvironment, either by favoring the recruitment of anti-tumor effector T and B cells or triggering tumor-supportive states (Harris et al., 2024). In order to retrieve region-specific gene modules, we applied Hotspot (DeTomaso & Yosef, 2021) on spatial coordinates, enabling the grouping of genes into five different modules(Figure S1), among which one highly associated with the invasive tumor region (Figure 1C-left) and another one with the tumor stroma (Figure 1C-right). We then assigned cells to *G1* for endothelial cells in the invasive niche and *G2* for endothelial cells in the stroma (Figure 1D-left). We also defined *N1* and *N2*, the non-endothelial spatial nearest neighbors of *G1* and *G2*, respectively (Figure 1D-right).

Table 1: Precision of gene discoveries for the task of finding upregulated genes in *G1 vs G2*. lvm-DE: latent-variable model DE. Details in Appendix C.

	Original segmentation		Proseg segmentation	
	t-test	lvm-DE	t-test	lvm-DE
Naive	0.27	0.29	0.41	0.47
2DE	0.69	0.75	0.67	**0.88**

To characterize endothelial gene expression in invasive and stromal regions, we performed differential expression (DE) analysis between *G1* and *G2*, comparing two methods $\mathcal{M}$: a latent-variable model DE (lvm-DE) (Boyeau et al., 2023), implemented in scvi-tools, and a t-test, implemented in Scanpy (Wolf et al., 2018). We first investigated genes enriched within endothelial cells of the invasive region compared to the tumor stroma. To assess the reliability of these discoveries, we compared the results of both methods against an scRNA-seq reference dataset (Appendix C). The naive approach identified genes that were downregulated in endothelial cells within the scRNA-seq reference (Figure 1E), suggesting contamination from neighboring cell types. We therefore filtered genes by setting $\tau = 0.9$. To quantitatively compare the refined DE results from 2DE with the naive approach, we evaluated precision using endothelial-upregulated genes in the scRNA-seq reference as ground truth markers. Across both DE methods and segmentations, we observed a consistent increase in precision (Table 1). As expected, we find a stronger improvement in the lower quality original segmentation compared to the improved Proseg segmentation validating our benchmarking strategy.

Among the confident genes, we found *ESM1*, described to be upregulated in invasive breast cancer tissue (Zeng et al., 2023). Also enriched were *SNAI1* and *ZEB1*. *SNAI1* is a key regulator of dysfunctional blood vessels in cancer (Hoffmann et al., 2024). Upon being activated during endothelial to mesenchymal transition, SNAI1 induces the expression of multiple other transcription factors such as *ZEB1*, resulting in neoangiogenesis and promoting cancer growth (Youssef & Nieto, 2024). In addition, we found increased expression of *KDR*, encoding vascular endothelial growth factor receptor 2 (VEGFR2), which is essential for angiogenesis as well as increased permeability (Pérez-Gutiérrez & Ferrara, 2023). Spatial plots of *ESM1*, *SNAI1* and *KDR* show spatial specificity in endothelial cells (Figure 1F-top) and limited expression in non-endothelial cells (Figure 1F-bottom). Taken together, *ESM1*, *KDR* and *ZEB1* are critical for angiogenesis in invasive cancer and we thereby identify spatially confined tumor-promoting endothelial cells (Motzer et al., 2020; Li et al., 2022).
We compared the results obtained with both segmentations in Figure S2. In the original segmentation, the naive approach identifies genes, such as *FOXA1* (Balsalobre & Drouin, 2022; Janesick et al., 2023; Bhat-Nakshatri et al., 2024)) and *KRT7* (Elmentaite et al., 2022) that are actually expressed by cancer cells and therefore spurious signal. Spatial plots of these two genes show high expression in non-endothelial cells (Figure 1G-bottom, Figure S3B-bottom). While we find a drastically reduced frequency of spurious upregulated genes using improved segmentation, 2DE identifies a similar set of marker genes. We computed Jaccard indices between sets of local markers and neighborhood genes, $Jaccard(\mathcal{S}_1^{original}, \mathcal{S}_1^{proseg})$ and $Jaccard(\mathcal{N}_1^{original}, \mathcal{N}_1^{proseg})$, and found a low overlap between neighborhood genes but a high overlap between markers (Figure S2).

Endothelial cells located in the stroma express higher levels of canonical markers of endothelial cells including *EDN1* encoding endothelin a potent vasoconstrictor (Geldhof et al., 2022), *CAVIN2*, key for the formation of caveolae and regulating nitric oxide production, which is a potent vasodilatator (Aitken et al., 2023; Boopathy et al., 2017) as well as the canonical marker of endothelial cells *CLDN5* (Reed et al., 2024) (Figure S3C). Spatial plots of *EDN1*, *CAVIN2* and *CLDN5*, showing spatial specificity in endothelial cells, are displayed in Figure S3D-E. Higher expression of these canonical markers of endothelial cells highlights more mature functional endothelial cells in the stroma compared to those in the invasive cancer region, expressing markers of angiogenesis as well as endothelial-to-mesenchymal transition and, thereby, the loss of markers of differentiated endothelial cells. To summarize, endothelial cells promoting angiogenesis are spatially tightly confined to invasive cancer and likely provide a tissue environment promoting cancer growth.

4 DISCUSSION

Differential expression analysis in spatial transcriptomics data is confounded by spurious signals originating from neighboring cells, and thereby wrong gene discoveries. To address this issue, we introduced 2DE, a method that assigns confidence scores to genes by assessing their upregulation relative to the neighborhood. Applying 2DE to a breast cancer dataset, we demonstrated its utility as a post-processing step for multiple DE methods, significantly improving the precision of detected markers compared to a naive approach. Furthermore, we showed that 2DE identifies a consistent set of confident marker genes, regardless of segmentation quality.

MEANINGFULNESS STATEMENT

Biological systems rely on processing information between individual cells and their environment to exert higher-order tissue functions, such as the synchronized conduction of the heart muscle, cognition in the central nervous system, or the immune response to pathogens. Meaningful representations of life should reflect these interactions and spatial omics set the ground for it. However, technical artifacts still obscure representations derived from this data. We introduced a broadly applicable method to help researchers interpret spatial data better and derive meaningful biological insights from it.

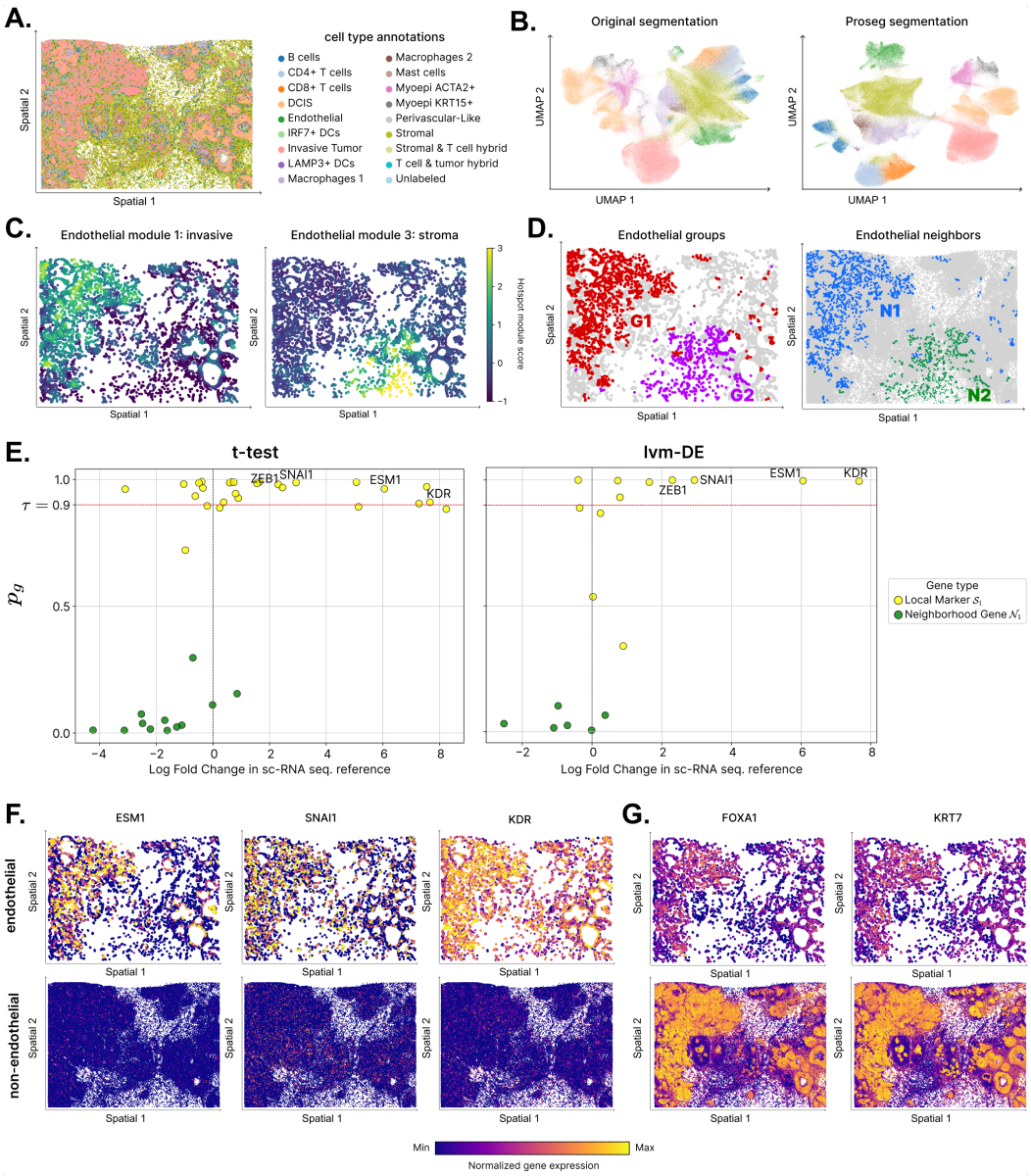

Figure 1: 2DE enables the characterization of spatially confined endothelial populations in breast cancer samples. **A.** Tissue slice colored by cell type labels. **B.** UMAP of embeddings obtained with the NicheVI model (Levy et al., 2024), using the original cell segmentation (left) and after resegmenting the cells using Proseg (Jones et al., 2024) (right). Cells are colored by cell type labels. **C.** On the resegmented data, endothelial gene modules from Hotspot applied on spatial coordinates. One module co-localizes with the invasive tumor region, and the other with the tumor stroma. Cells are colored by module scores. **D.** We identify endothelial cells in the invasive niche *G1* and endothelial cells in the stroma *G2*. *N1*, *N2* refers to the nearest neighbors in space of *G1* and *G2* while ignoring endothelial cells. **E.** After Proseg segmentation and for two different DE methods, we visualize p_g as a function of the log-fold change from an endothelial-vs-all comparison in scRNA-seq data of breast cancer. All genes identified by the naive DE approach are shown, colored by their marker label. Notably, neighborhood genes are downregulated in the reference dataset, suggesting contamination from neighboring cells in spatial data. Genes with $p_g > 0.9$ are kept as confident markers. **F.** We display the spatial gene expression of marker genes upregulated in both segmentations, subset to endothelial cells (top) and subset to non-endothelial cells (bottom). These genes are higher expressed in endothelial cells. **G.** Similar to F. These genes are higher expressed in non-endothelial cells. Here, we display the observed expression after Proseg segmentation.

A METHODS

A 2DE DETAILS

2DE takes as input two defined groups of cells $\{G1, G2\}$ as well as their spatial coordinates and cell types. To determine the upregulated genes of *G1 vs G2*, the first step is to compute the neighborhood $N1$. We computed a nearest neighbors graph with $k_{DE} = 6$ neighbors, which led to an average number of neighbors $k_{adj} = 2.7$ after removing cell-type connections. Then, we compute DE between $\{G1, G2\}$, $\{N1, G2\}$ and $\{G1, N1\}$, using the method $\mathcal{M}$. We considered two methods; the first one, lvm-DE (Boyeau et al., 2023), leverages a fitted generative model to estimate log-fold changes between conditions from the normalized expression distribution. We used the NicheVI (Levy et al., 2024) deep generative model using default architecture and training parameters, then ran lvm-DE with the following parameters: a pseudo-count $\epsilon = 10^{-4}$, a LFC cutoff $\delta = 0.03$ for all comparisons except $\{G1, N1\}$ for which $\delta = 0.15$, $n_{samples} = 10^5$ samples from the posterior. We also ran a t-test with Scanpy, using the `scanpy.tl.rank_genes_groups` method on the log-transformed gene expression.

On the upregulated genes of *G1 vs G2*, we then trained a Gaussian process classifier to distinguish between marker and neighborhood genes. We used the Scikit-learn (Pedregosa et al., 2011) implementation and defined the kernel as the product of a constant kernel C with a rational quadratic kernel $\mathcal{K}(l, a)$. We tuned the Gaussian process hyperparameters by sampling 20 combinations within given bounds. We set $C \in [10^{-3}, 10^3]$, $l \in [10, 10^2]$ and $a \in [10^{-3}, 1]$.

To determine the upregulated genes of *G2 vs G1*, we follow the same steps: we first retrieve *N2*, then compute DE between $\{G2, G1\}$, $\{N2, G1\}$ and $\{G2, N2\}$.

For the spatial plots of gene expression, we processed the counts as follows: we applied a `scanpy.pp.log1p` transformation followed by min-max scaling, then selected the 99.9th percentile as the upper limit of the color scale.

B SPATIAL GENE MODULES COMPUTATION

We use Hotspot (DeTomaso & Yosef, 2021) to uncover spatial signatures of cell states. To generate Figure 1C, we ran Hotspot using the spatial coordinates as similarity metric, following hotspot.readthedocs.io/tutorial. Then we used the module scores to assign cells to modules, by considering the maximal module score for the cell. To account for cells expressing a mixture of gene modules with no clear over-expression of one module, we set a threshold Δ to the score. Formally, for N cells and M modules, we denote $\boldsymbol{H} \in \mathbb{R}^{N \times M}$ the module scores matrix. For any cell n, the module m_n is:

$$m_n = \begin{cases} j = \text{argmax}_{m \in M} \left[H_{nm} \right], & \text{if } H_{nj} > \Delta \\ 'Unassigned' & \text{otherwise.} \end{cases} \tag{2}$$

We set $\Delta = 0.5$. Spatial plots of module assignments m are displayed in Figure S1.

C BENCHMARK WITH NAIVE APPROACH USING SCRNA-SEQ REFERENCE DATA

To assess the biological plausibility of the upregulated genes discovered by a DE method from spatial transcriptomics data, we can compare our results with reference scRNA-seq data. We considered a breast cancer atlas (Wu et al., 2021) of 26 tumors and defined ground truth endothelial markers by computing endothelial-vs-all DE using a t-test and selecting genes with positive LFCs. We selected the intersection of these markers with the Xenium gene panel. We then defined the *precision* of a DE method as the fraction of relevant discoveries (genes that are within the ground truth markers) among all discoveries.

REFERENCES

C. Aitken, V. Mehta, M. A. Schwartz, and E. Tzima. Mechanisms of endothelial flow sensing. *Nature Cardiovascular Research*, 2(6):517–529, Jun 2023. doi: 10.1038/s44161-023-00276-0.

Aurelio Balsalobre and Jacques Drouin. Pioneer factors as master regulators of the epigenome and cell fate. *Nat. Rev. Mol. Cell Biol.*, 23(7):449–464, July 2022.

Poornima Bhat-Nakshatri, Hongyu Gao, Aditi S Khatpe, Adedeji K Adebayo, Patrick C McGuire, Cihat Erdogan, Duojiao Chen, Guanglong Jiang, Felicia New, Rana German, Lydia Emmert, George Sandusky, Anna Maria Storniolo, Yunlong Liu, and Harikrishna Nakshatri. Single-nucleus chromatin accessibility and transcriptomic map of breast tissues of women of diverse genetic ancestry. *Nat. Med.*, 30(12):3482–3494, December 2024.

G. T. K. Boopathy, M. Kulkarni, S. Y. Ho, A. Boey, E. W. M. Chua, V. A. Barathi, T. J. Carney, X. Wang, and W. Hong. Cavin-2 regulates the activity and stability of endothelial nitric-oxide synthase (enos) in angiogenesis. *Journal of Biological Chemistry*, 292(43):17760–17776, Oct 2017. doi: 10.1074/jbc.M117.794743.

Pierre Boyeau, Jeffrey Regier, Adam Gayoso, Michael I Jordan, Romain Lopez, and Nir Yosef. An empirical bayes method for differential expression analysis of single cells with deep generative models. *Proceedings of the National Academy of Sciences*, 120(21):e2209124120, 2023.

Dario Bressan, Giorgia Battistoni, and Gregory J Hannon. The dawn of spatial omics. *Science*, 381 (6657):eabq4964, 2023.

Hao Chen, Young Je Lee, Jose A Ovando-Ricardez, Lorena Rosas, Mauricio Rojas, Ana L Mora, Ziv Bar-Joseph, and Jose Lugo-Martinez. Recovering single-cell expression profiles from spatial transcriptomics with scresolve. *Cell Reports Methods*, 4(10), 2024.

David DeTomaso and Nir Yosef. Hotspot identifies informative gene modules across modalities of single-cell genomics. *Cell systems*, 12(5):446–456, 2021.

Rasa Elmentaite, Cecilia Domínguez Conde, Lu Yang, and Sarah A Teichmann. Single-cell atlases: shared and tissue-specific cell types across human organs. *Nat. Rev. Genet.*, 23(7):395–410, July 2022.

Can Ergen and Nir Yosef. Resolvi-addressing noise and bias in spatial transcriptomics. *bioRxiv*, pp. 2025–01, 2025.

V. Geldhof, L. P. M. H. de Rooij, L. Sokol, J. Amersfoort, M. De Schepper, K. Rohlenova, G. Hoste, A. Vanderstichele, A. M. Delsupehe, E. Isnaldi, N. Dai, F. Taverna, S. Khan, A. K. Truong, L. A. Teuwen, F. Richard, L. Treps, A. Smeets, I. Nevelsteen, B. Weynand, S. Vinckier, L. Schoonjans, J. Kalucka, C. Desmedt, P. Neven, M. Mazzone, G. Floris, K. Punie, M. Dewerchin, G. Eelen, H. Wildiers, X. Li, Y. Luo, and P. Carmeliet. Single cell atlas identifies lipid-processing and immunomodulatory endothelial cells in healthy and malignant breast. *Nature Communications*, 13(1):5511, sep 2022. doi: 10.1038/s41467-022-33052-y.

Michael A Harris, Peter Savas, Balaji Virassamy, Megan M R O'Malley, Jasmine Kay, Scott N Mueller, Laura K Mackay, Roberto Salgado, and Sherene Loi. Towards targeting the breast cancer immune microenvironment. *Nat. Rev. Cancer*, 24(8):554–577, August 2024.

Helene Hoffmann, Martin Wartenberg, Sandra Vorlova, Franziska Karl-Schöller, Matthias Kallius, Oliver Reinhardt, Asli Öztürk, Leah S Schuhmair, Verena Burkhardt, Sabine Gätzner, et al. Normalization of snai1-mediated vessel dysfunction increases drug response in cancer. *Oncogene*, 43 (35):2661–2676, 2024.

Amanda Janesick, Robert Shelansky, Andrew D Gottscho, Florian Wagner, Stephen R Williams, Morgane Rouault, Ghezal Beliakoff, Carolyn A Morrison, Michelli F Oliveira, Jordan T Sicherman, et al. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nature Communications*, 14(1):8353, 2023.

Daniel C Jones, Anna E Elz, Azadeh Hadadianpour, Heeju Ryu, David R Glass, and Evan W Newell. Cell simulation as cell segmentation. *bioRxiv*, 2024.

Nathan Levy, Florian Ingelfinger, Can Ergen-Behr, and Boaz Nadler. Nichevi: A probabilistic framework to embed cellular interaction in spatial transcriptomics. In *ICLR 2024 Workshop on Machine Learning for Genomics Explorations*, 2024.

R. Li, J. R. Ferdinand, K. W. Loudon, G. S. Bowyer, S. Laidlaw, F. Muyas, L. Mamanova, J. B. Neves, L. Bolt, E. S. Fasouli, A. R. J. Lawson, M. D. Young, Y. Hooks, T. R. W. Oliver, T. M. Butler, J. N. Armitage, T. Aho, A. C. P. Riddick, V. Gnanapragasam, S. J. Welsh, K. B. Meyer, A. Y. Warren, M. G. B. Tran, G. D. Stewart, I. Cortés-Ciriano, S. Behjati, M. R. Clatworthy, P. J. Campbell, S. A. Teichmann, and T. J. Mitchell. Mapping single-cell transcriptomes in the intra-tumoral and associated territories of kidney cancer. *Cancer Cell*, 40(12):1583–1599.e10, December 2022. doi: 10.1016/j.ccell.2022.11.001.

R. J. Motzer, R. Banchereau, H. Hamidi, T. Powles, D. McDermott, M. B. Atkins, B. Escudier, L. F. Liu, N. Leng, A. R. Abbas, J. Fan, H. Koeppen, J. Lin, S. Carroll, K. Hashimoto, S. Mariathasan, M. Green, D. Tayama, P. S. Hegde, C. Schiff, M. A. Huseni, and B. Rini. Molecular subsets in renal cancer determine outcome to checkpoint and angiogenesis blockade. *Cancer Cell*, 38(6): 803–817.e4, December 2020. doi: 10.1016/j.ccell.2020.10.011.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Viktor Petukhov, Rosalind J Xu, Ruslan A Soldatov, Paolo Cadinu, Konstantin Khodosevich, Jeffrey R Moffitt, and Peter V Kharchenko. Cell segmentation in imaging-based spatial transcriptomics. *Nature biotechnology*, 40(3):345–354, 2022.

L. Pérez-Gutiérrez and N. Ferrara. Biology and therapeutic targeting of vascular endothelial growth factor a. *Nature Reviews Molecular Cell Biology*, 24(11):816–834, November 2023. doi: 10.1038/s41580-023-00631-w.

A.D. Reed, S. Pensa, A. Steif, J. Stenning, D.J. Kunz, L.J. Porter, K. Hua, P. He, A.J. Twigger, A.J.Q. Siu, K. Kania, R. Barrow-McGee, I. Goulding, J.J. Gomm, V. Speirs, J.L. Jones, J.C. Marioni, and W.T. Khaled. A single-cell atlas enables mapping of homeostatic cellular shifts in the adult human breast. *Nature Genetics*, 56(4):652–662, Apr 2024. doi: 10.1038/s41588-024-01688-9. Epub 2024 Mar 28.

F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.

Sunny Z Wu, Ghamdan Al-Eryani, Daniel Lee Roden, Simon Junankar, Kate Harvey, Alma Andersson, Aatish Thennavan, Chenfei Wang, James R Torpy, Nenad Bartonicek, et al. A single-cell and spatially resolved atlas of human breast cancers. *Nature genetics*, 53(9):1334–1347, 2021.

K. K. Youssef and M. A. Nieto. Epithelial-mesenchymal transition in tissue repair and degeneration. *Nature Reviews Molecular Cell Biology*, 25(9):720–739, September 2024. doi: 10.1038/s41580-024-00733-z.

Q. Zeng, M. Mousa, A. S. Nadukkandy, L. Franssens, H. Alnaqbi, F. Y. Alshamsi, H. A. Safar, and P. Carmeliet. Understanding tumour endothelial cell heterogeneity and function from single-cell omics. *Nature Reviews Cancer*, 23(8):544–564, August 2023. doi: 10.1038/s41568-023-00591-5.

B SUPPLEMENTARY FIGURES

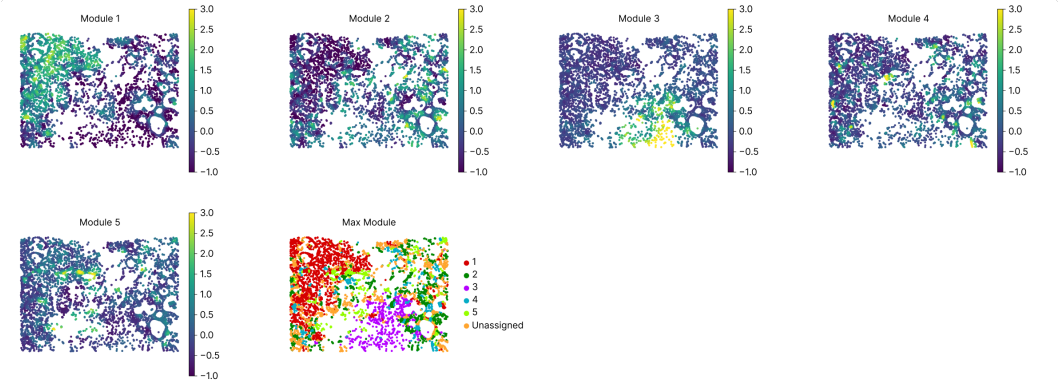

Figure S1: **A.** Spatial plots of module scores, computed with Hotspot on the resegmented data. We assigned each cell to a specific module by taking the argmax over the cell module scores and setting a threshold of 0.5, under which cells are assigned to a 'weak' module class. **B.** Similar to A, but on the original segmentation.

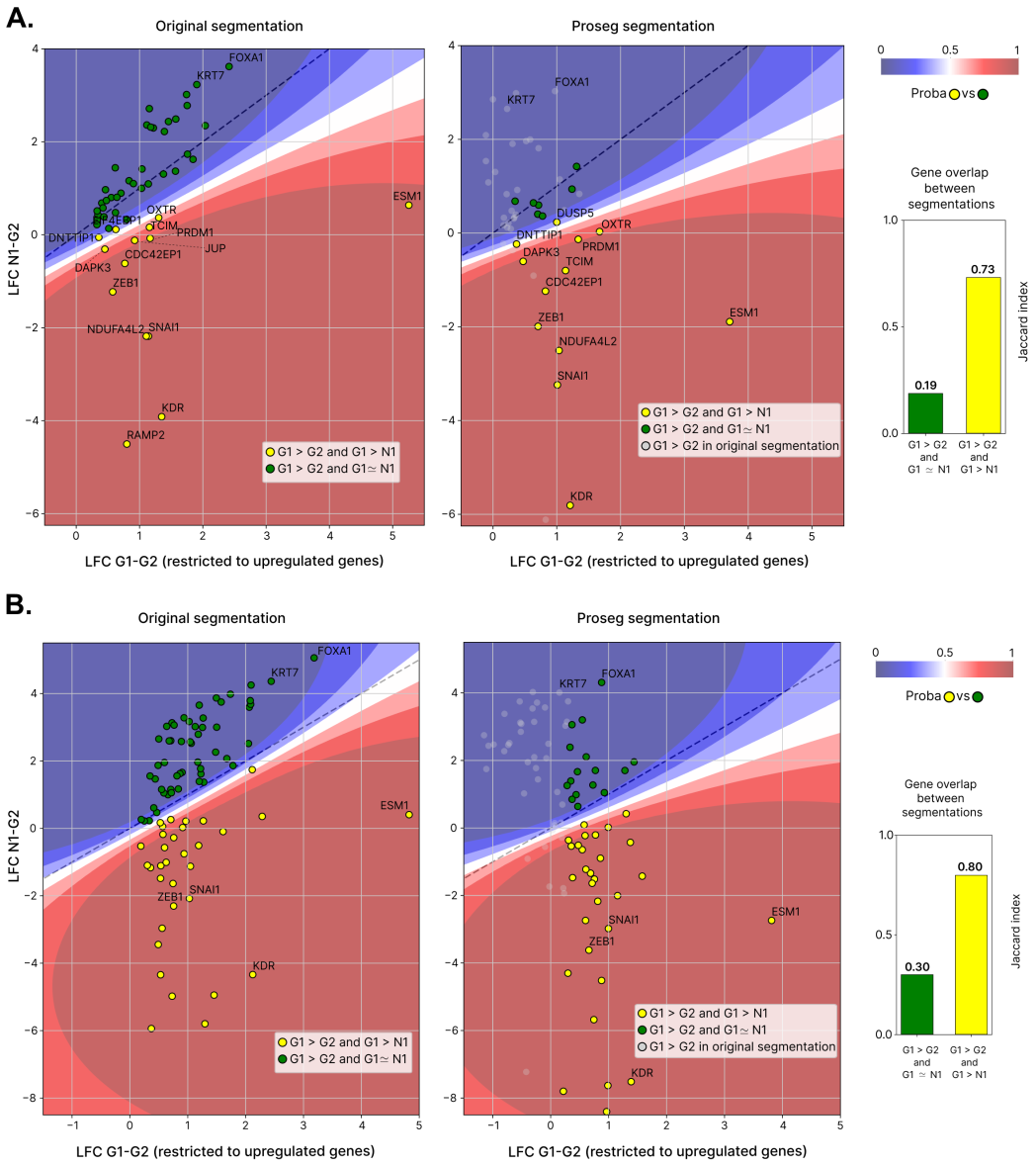

Figure S2: **A.** Using lvm-DE, median Log-Fold Change (LFC) of upregulated genes in *G1* vs *G2*, using the original segmentation (left) and the resegmented data (right) displayed on x-axis, while we compare differential expression computed between *N1* and *G2* on the y-axis. Genes are colored by their marker label $\mathcal{S}_1$(yellow)/$\mathcal{N}_1$(green). We also display the classifier decision boundary. The Jaccard index measures gene overlaps for the markers/ neighborhood gene sets. **B.** Similar to A, with t-test as DE method.

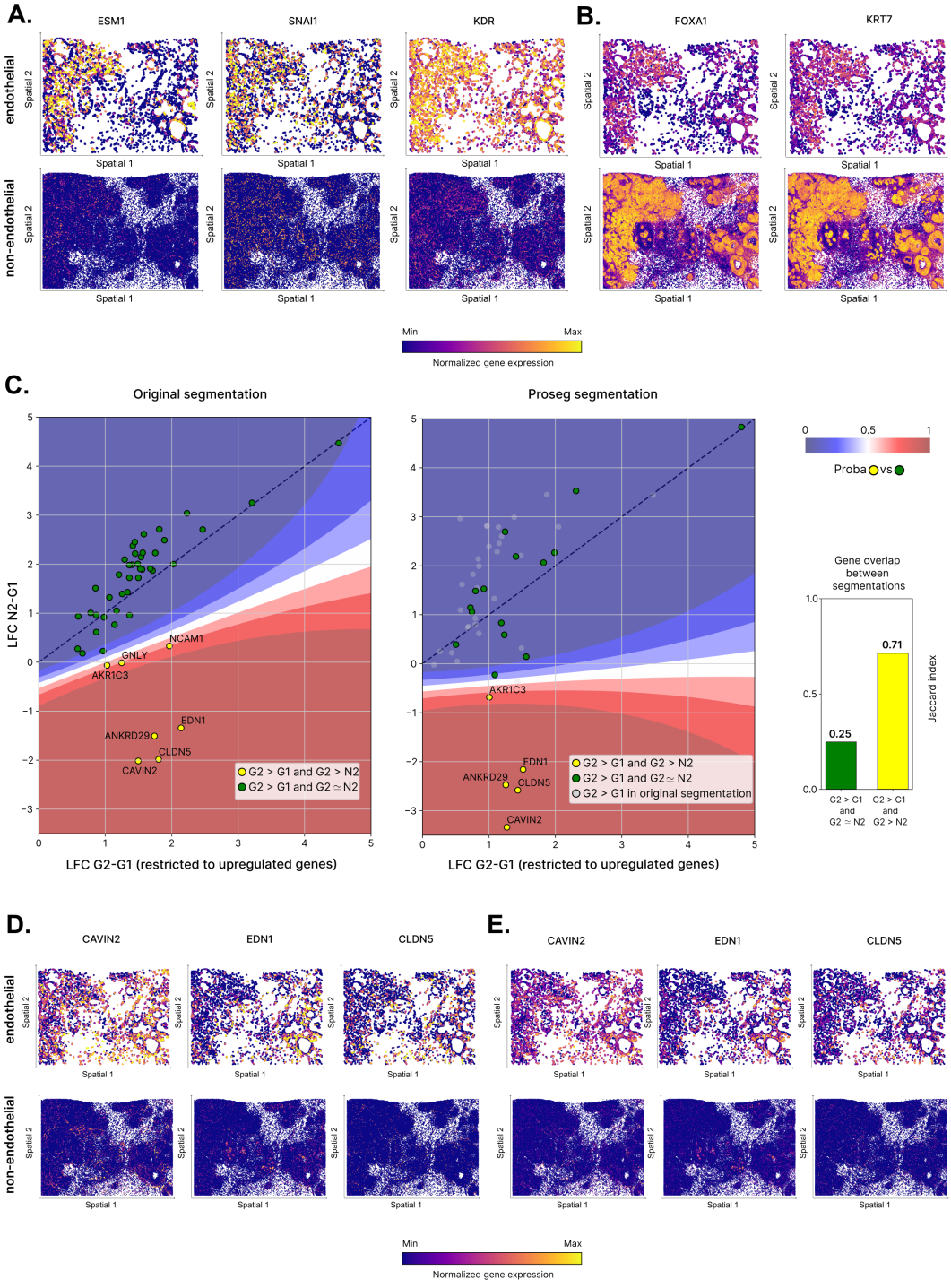

Figure S3: **A-B.** Similar to Figure 1F-G, but in the original segmentation. **C.** Using lvm-DE, median Log-Fold Change (LFC) of upregulated genes in *G2* vs *G1*, using the original segmentation (left) and the resegmented data (right) displayed on the x-axis, while we compare differential expression computed between *N2* and *G1* on the y-axis. Genes are colored by their marker label $\mathcal{S}_1/\mathcal{N}_1$. We also display the classifier decision boundary. The Jaccard index measures gene overlaps for the markers/ neighborhood gene sets. **D.** Spatial plots of genes upregulated in *G2*, in endothelial cells (top) and non-endothelial cells (bottom), showing stromal endothelial specificity. Here, we display the observed expression in the original segmentation. **E.** Similar to D, but after Proseg segmentation.