**CosMx Data QC and Processing**

Each field of view (FOV) was reviewed, and *Cellpose[[1]](#footnote-1)* cell segmentation was performed, using NanoString’s preset configurations A or E for the “cyto2” model, based on which appeared to perform best. Next, cells that did not exhibit the following characteristics were discarded: total RNA counts greater than 20 and less than 2000, cell area greater than 25 µm2 and less than 1000 µm2, proportion of total counts from negative control probes less than 0.1, and complexity score less than 2. Complexity score was defined as log10[(total RNA counts) / (positive RNA genes)]. Thus, cells with 100 times as many counts as positive genes were considered erroneous and were discarded. After filtering poor quality cells, FOVs with more than 25% of their cells deemed poor quality were themselves entirely discarded to preserve general contiguity of the tissue structure for downstream spatial analysis. Core 22 (**Fig S2b)** was removed from TMA2 due to high levels of necrosis and outlier-level expression of spike-in probes. 91.6% and 83.6% of cells from TMA1 and TMA2 were retained, respectively.

Cells from each TMA were then processed separately, using *Seurat[[2]](#footnote-2)*. First, expression was normalized to log2(counts per thousand + 1) and z-score scaled for all genes. This scaled expression matrix was used to generate 50 principal components, 25 of which were used to generate a UMAP embedding for each TMA. After cell-typing was performed for each TMA, all data was merged. This merged data set was then subject to the same processing scheme. No batch effect was observed between the two TMAs.

**CosMx Data Cell-typing**

First, tumor cells and erythrocytes were identified manually for each TMA. To identify tumor cells, a tumor meta score was generated for each cell, by inputting the following genes into *Seurat*’s *AddModuleScore* function: APOA1, GC, VTN, ARG1, FGG, and GPC3. Louvain clustering was performed with resolution parameters of 0.5 and 0.7 for TMA1 and TMA2 respectively. By visualizing the resulting clusters and the tumor meta score on the UMAP embedding, tumor clusters were found. To identify erythrocytes, all genes with 3 blue bits present in their reporter probe barcode were passed to *AddModuleScore* to produce an autofluorescence meta score. As with tumor cells, using Louvain clustering and visualization, erythrocyte clusters were found. (**Fig 2a**).

Next, using liver[[3]](#footnote-3) and immune-oncology reference profiles, supervised clustering was performed using *InSituType[[4]](#footnote-4)* on the remaining cells to identify final cell types. Spatial nearest neighbor expression produced from each cell’s 50 nearest spatial neighbors was converted to 10 principal components. These components, along with each cell’s immunofluorescent staining data and cell area was passed to *InSituType* as cohort data to improve clustering. When clustering with the immune-oncology profile, genes that were likely sources of spatial contamination were excluded.

**CosMx Data Cell-typing Validation**

After cell-typing, the data from both TMAs was merged, and marker genes were found for each cell type cluster, using two methods. First, an algorithmic approach[[5]](#footnote-5) identified genes that are uniquely highly expressed in each cluster. First, the mean counts per cell was found for each gene in each cluster. For each cluster, all genes with a counts per cell value greater than both 0.1 and the second-largest counts per cell value in any other cluster were identified. These genes were then ranked according to how much larger their counts per cell value was in each cluster, compared to their next largest counts per cell value in any other cluster. Finally, for each cluster, the top 1-5 ranked genes were selected as marker genes for visualization (**Fig 2c)**.

Second, a modeling approach identified significant differentially expressed genes that characterize each cluster. Specifically, after pseudo-bulking, the *presto-GLMM* R package was used to fit a Poisson GLMM for each gene’s counts with the following formula: y\_g ~ 1 + (1|celltype) + (1|celltype:patient) + (1|patient) + offset(logUMI). Linear contrasts representing each cluster versus the rest were tested for each gene’s model to find logFC values and corresponding p-values via a Wald test. These p-values were FDR-corrected. For each cluster, genes with logFC greater than 1.5 and adjusted p-value less than 0.05 were considered marker genes (**Fig S4)**. Review of each cluster’s marker genes demonstrated that high fidelity annotations had been made.

**CosMx Data Differential Expression Analysis**

To identify LINE1-ORF1 high patients, the total LINE1-ORF1 counts in all tumor cells was quantified for each patient. This measure was converted to counts per million (CPM), and the patients in the upper tercile for LINE1-ORF1 CPM were labeled “high,” while the remaining patients were labeled “low” (**Fig 3a**). Next, each gene’s counts in tumor, endothelial, and macrophage cells were modeled with a Negative Binomial GLMM using the *NEBULA[[6]](#footnote-6)* package and the following formula: y\_g ~ line1orf1\_group + celltype + line1orf1\_group:celltype + (1|patient) + offset(logUMI). A linear contrast representing “high” tumor cells versus “low” tumor cells was tested for the model with an asymptotic Wald test for each gene, and the resulting p-values were FDR-adjusted. Genes that were likely sources of spatial contamination were not modeled or tested (**Fig 3b-c**). This same scheme was applied to endothelial cells as well (**Fig S5a**).

To identify LINE1-ORF1 high tumor cells, within each patient, cells in the upper tercile for LINE1-ORF1 normalized expression were labeled “high” and the remaining cells were labeled “low” (**Fig 3d**). Tumor cells in each group were pseudo-bulked to form “high” and “low” paired tumor samples for each patient. Next, each gene’s counts in tumor samples were modeled with the *DESeq2[[7]](#footnote-7)* package using the following formula: y\_g ~ line1orf1\_group + patient. During this process, size factors were calculated for each sample, using the upper quartile method from *edgeR[[8]](#footnote-8)*. Finally, a linear contrast representing “high” tumor cells versus “low” tumor cells was tested with a likelihood ratio test (**Fig 3e-f**). For this, we set parameters fitType = “glmGamPoi” and reduced = ~patient.

**CosMx Data Gene Set Enrichment Analysis**

Using our pseudo-bulk based differential expression analysis results, we omitted spike-ins and likely contaminated genes, then ranked all remaining genes in descending order of their likelihood ratio test statistic multiplied by their logFC sign. This pre-ranked gene list was passed to the GSEA function from the *clusterProfiler[[9]](#footnote-9)* package, which employs the *fgsea[[10]](#footnote-10)* algorithm. We tested all gene sets in the C8 and GO:MF categories from MSigDB[[11]](#footnote-11) (**Fig 3g**).

**CosMx Data Neighborhood Enrichment Analysis**

We conducted neighborhood enrichment analysis using methods inspired by Varrone et al.[[12]](#footnote-12). Within each FOV, we computed a Delaunay tessellation. For each cell, which we refer to as the index cell, we defined the neighbor cells to be the set of directly adjacent cells on this tessellation. We combined FOV-wise neighborhood definitions to form a global neighborhood graph, trimming all graph edges with an associated distance value greater than 100 µm. Thus, by this scheme, for a cell to be a neighbor of a given index cell, the cell must exist in the same FOV as the index; the cell must be directly adjacent to index, with no cells between them; and the cell must be within 100 µm of the index.

After partitioning tumor cells into LINE1-ORF1 high and low subsets in a patient-wise fashion, as previously described, we aimed to assess high and low tumor cell neighborhoods for comparison. To accomplish this, we did the following for each cell-type, including LINE1 high and low tumor cells: 1) In a patient-wise fashion, for a given cell type, we treated all cells of that type as index cells. 2) We found the proportions of neighbors to all index cells that belonged to each cell type identity. 3) We defined these proportions as the observed proportions of neighbors to the index cell type. 4) We defined the expected proportions of neighbors to the index cell type analytically. The expected proportion of neighbors of a specific cell type identity, which we will denote as cell type X, was defined as the total number of neighborhood graph edges connecting to cells of type X, divided by the total number of edges in the graph. 5) Adding a small pseudo-count of 0.0001 to the observed and expected proportion values, then taking log2 of the (observed/expected) ratio defined the log2 enrichment ratio (log2ER), a measure of colocalization. 6) We compared log2ER profiles between “high” and “low” tumor index cells, using a patient-paired Wilcoxon Rank Sum Test (**Fig 4a, Fig S5a**).

Second, we aimed to compare the cell co-localization patterns in LINE1-ORF1 high and low patients. To accomplish this, we first split the data set by patient-level LINE1-ORF1 group. Just as previously described, we computed neighborhood graphs for each partition of the data set. Without splitting by individual patient, we computed log2ER values for each (**Fig 4c**). To assess significance for differences in log2ER values across the two partitions, we used simulation. We randomly permuted the patient grouping labels 1000 times. During each permutation, we computed the difference between the log2ER values in the “high” group and the log2ER values in the “low” group. Comparing a given observed difference to the simulated differences allowed us to calculate empirical p-values (**Fig S6b**). To test for a difference in the overall immune cell colocalization pattern across the “high” and “low” patient groups, we defined the sum of all pairwise log2ER values for all immune cell types as the immune colocalization score. Next, we found the difference in this score across the “high” and “low” patient groups, which we defined as the observed value. Recording this same metric during each of the 1000 simulations detailed above allowed us to define an empirical null distribution, which we used to calculate an empirical p-value (**Fig 4d)**.

**RNA-ISH Data Analysis**

RNA-ISH counts were quantified, and cell segmentation was performed, using Indica Labs HALO software. For each patient, the total number of LINE1 counts per µm2 of tissue was quantified. The metric was chosen because the vast number of spots present for LINE1 ISH stains makes resolving each spot’s location as being within a specific cell challenging. All other RNA-ISH targets were quantified in terms of counts per cell for each patient. Patients in the upper tercile for LINE1 counts per µm2 were labeled as LINE1 high, while the remaining patients were labeled as LINE1 low. Kaplan Meier analysis was performed based on this LINE1 split with a log-rank test for significance using the *survival* R package.

**Data and Code Availability**

All data will be provided upon request. Analysis code is available at <https://github.com/ccnawrocki/hcc-tma-cosmx-project>.

1. Stringer, C., Wang, T., Michaelos, M., & Pachitariu, M. (2021). Cellpose: a generalist algorithm for cellular segmentation. Nature methods, 18(1), 100-106. [↑](#footnote-ref-1)
2. Hao, Y., Stuart, T., Kowalski, M.H. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* 42, 293–304 (2024). https://doi.org/10.1038/s41587-023-01767-y [↑](#footnote-ref-2)
3. Karlsson M et al., A single-cell type transcriptomics map of human tissues. *Sci Adv.* (2021) PubMed: [34321199](https://ncbi.nlm.nih.gov/pubmed/34321199)DOI: [10.1126/sciadv.abh2169](http://dx.doi.org/10.1126%2Fsciadv.abh2169) [↑](#footnote-ref-3)
4. https://www.biorxiv.org/content/10.1101/2022.10.19.512902v1 [↑](#footnote-ref-4)
5. Danaher P, Hasle N, Nguyen ED, Hayward K, Rosenwasser N, Alpers CE, Reed RC, Okamura DM, Baxter SK, Jackson SW. Single cell spatial transcriptomic profiling of childhood-onset lupus nephritis reveals complex interactions between kidney stroma and infiltrating immune cells. bioRxiv [Preprint]. 2023 Nov 13:2023.11.09.566503. doi: 10.1101/2023.11.09.566503. Update in: Sci Transl Med. 2024 Nov 27;16(775):eadl1666. doi: 10.1126/scitranslmed.adl1666. PMID: 38014158; PMCID: PMC10680641. [↑](#footnote-ref-5)
6. He, Liang, Jose Davila-Velderrain, Tomokazu S. Sumida, David A. Hafler, Manolis Kellis, and Alexander M. Kulminski. 2021. “NEBULA Is a Fast Negative Binomial Mixed Model for Differential or Co-Expression Analysis of Large-Scale Multi-Subject Single-Cell Data.” Communications Biology, no. 629 (May). <https://doi.org/10.1038/s42003-021-02146-6>. [↑](#footnote-ref-6)
7. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014). https://doi.org/10.1186/s13059-014-0550-8 [↑](#footnote-ref-7)
8. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139-40. doi: 10.1093/bioinformatics/btp616. Epub 2009 Nov 11. PMID: 19910308; PMCID: PMC2796818. [↑](#footnote-ref-8)
9. https://www.sciencedirect.com/science/article/pii/S2666675821000667?via%3Dihub [↑](#footnote-ref-9)
10. https://www.biorxiv.org/content/10.1101/060012v3 [↑](#footnote-ref-10)
11. https://www.pnas.org/doi/10.1073/pnas.0506580102 [↑](#footnote-ref-11)
12. https://pubmed.ncbi.nlm.nih.gov/38066188/ [↑](#footnote-ref-12)