Segmentation Error in Spatial Transcriptomics

December 2025

Abstract

Cell segmentation in spatial transcriptomics is typically framed as accurately delineating cell boundaries. Alternatively, segmentation errors can be viewed as transcripts that map to locations other than their cell of origin. Under this framing, RNA diffusion prior to probe hybridization becomes a significant source of error, potentially affecting 15–22% of assigned transcripts.

1. Reframing Cell Segmentation Error

Cell segmentation algorithms for spatial transcriptomics typically aim to identify the physical boundaries of cells using nuclear staining (DAPI), membrane markers, or transcript density patterns. Performance is usually evaluated by comparing segmented boundaries to ground truth annotations such as staining for cytoplasmic membrane proteins.

However, for most biological applications what matters is not whether a segmentation boundary precisely follows the cell membrane, but whether the transcripts assigned to a given cell actually originated from that cell. In this view, segmentation error can be viewed as transcripts assigned to locations other than their true cell of origin. Under this definition, RNA diffusion—not boundary detection—emerges as a significant error mode.

2. RNA Diffusion in Spatial Transcriptomics

Imaging-based spatial transcriptomics methods such as Xenium, MERSCOPE, and CosMx detect transcripts through in situ hybridization of fluorescent probes followed by imaging. The recorded position of each transcript reflects its location at the time of probe hybridization, not necessarily its location within the detected cell. The segmented cell is defined by a detected nucleus, either by H&E or DAPI staining and a distance from the centroid of that detected nucleus. Between tissue sectioning and probe binding, RNA molecules may diffuse from their original positions.

Lateral Diffusion

Lateral (xy-plane) diffusion displaces transcripts mainly horizontally within the tissue section. In fresh frozen tissue, RNA molecules are not crosslinked and may diffuse freely in the aqueous environment during sample processing. FFPE fixation creates protein-RNA crosslinks that may limit but not eliminate this diffusion. Lateral diffusion causes transcripts to move from their source cell into neighboring cells or into the interstitial space between cells.

Analysis of transcript density as a function of distance from the tissue periphery provides insight into the range of lateral diffusion. At the tissue edge, there are no neighboring cells to contribute contaminating signal, yet we observe gradual increase in transcript density moving inward from the edge. This gradient extends approximately 10–20 μm before reaching the bulk tissue density, suggesting that lateral diffusion typically displaces transcripts by this distance.

Independent support comes from the Salas et al. analysis, which found that transcripts more than 10.7 μm from cell centroids correlate more strongly with domain-specific background than with cell-type-specific nuclear signatures. Given typical nuclear radii of approximately 5 μm, this implies that the biologically meaningful signal extends only about 5–6 μm beyond the nuclear boundary.

Vertical Diffusion

Vertical (z-axis) diffusion occurs through the thickness of the tissue section. Standard Xenium sections are 5 μm (FFPE) or 10 μm (fresh frozen), while cells typically have diameters of 10–20 μm. A single tissue section therefore captures only a portion of each cell's volume. The section may contain overlapping portions of multiple cells stacked in z, and transcripts may diffuse vertically within the section. The contribution of vertical diffusion is expected to be greater in thicker sections. The combined effect of lateral and vertical diffusion is that a transcript detected at position (x, y) may have originated from a cell whose centroid is located some distance away. If this distance exceeds the cell boundary, the transcript will not be assigned to the correct cell.

Non-Specific Binding and Technical Noise

In addition to RNA diffusion, a minor source of transcript mislocalization arises from non-specific probe binding and interactions with the glass substrate. Xenium and similar platforms include negative control probes (NCPs) that target non-biological sequences and should produce no signal. These controls measure the rate of off-target binding events independent of diffusion.

In well-performing assays, the false signal from non-specific binding is typically less than 1 - 3% of detected transcripts, substantially lower than the 15–22% contamination estimated from RNA diffusion. Furthermore, negative control signals are typically concentrated in off-tissue regions and do not show the spatial gradient from tissue edges that characterizes diffusion. This distinction confirms that the interstitial signal we observe is predominantly diffused RNA rather than non-specific probe binding to glass or other substrates.

3. Interstitial Density as a Diffusion Metric

Direct measurement of RNA diffusion in tissue sections is technically challenging. However, the interstitial space - regions within the tissue that are not occupied by segmented cells, and RNAs detected outside the tissue perimeter (“off tissue”) - provides an indirect measurement opportunity. Transcripts detected in interstitial regions cannot have originated there as there are no cells present to produce them. These transcripts must have diffused from adjacent cells.

If RNA diffusion is approximately isotropic (uniform in all directions), then the rate at which transcripts diffuse into interstitial space should equal the rate at which they diffuse into neighboring cells. Under this assumption, interstitial transcript density provides an estimate of the diffusion-based contamination rate within cells.

Algorithm

The following procedure was used to estimate contamination from Xenium output files:

1. Define the tissue boundary using a convex hull of cell centroids, dilated by 15 μm to

account for peripheral cells.
2. Calculate total tissue area and total segmented cell area (sum of individual cell areas).
3. Compute interstitial area as tissue area minus cell area.
4. Classify each transcript (Q≥20) as: assigned to a cell, interstitial (within tissue but not assigned), or off-tissue.
5. Calculate interstitial transcript density (transcripts per μm²).
6. Estimate per-cell contamination as interstitial density multiplied by cell area.
7. Express contamination as a fraction of total transcripts assigned to each cell.

Briefly, the level of contaminating RNA signal was estimated to be uniform and equal to the density of signal observed over regions of the tissue section that contained no cells.

4. Results from Public Xenium Datasets

We applied this analysis to four Xenium Prime 5K public datasets representing different tissue types and preparation methods:

Fresh frozen samples exhibited higher interstitial transcript density compared to FFPE samples, consistent with the greater section thickness (10 μm vs 5 μm) in fresh frozen preparations. Estimates of contaminating RNAs in cells ranged from 15.8% to 21.6%. The mouse brain dataset showed the highest estimated contamination (21.6%), though this value may be inflated by the presence of neuropil - the dense network of axons, dendrites, and synapses that legitimately contain RNA but is not captured by nuclear-based segmentation.

Table 1. Contamination estimates across four Xenium Prime 5K datasets.

Dataset	Tissue	Fixation	Interstitial Density	Est. Contamination
Lymph Node	Human	FFPE	4.31/1000μm²	17.8%
Prostate	Human	FFPE	3.52/1000μm²	15.8%
Mouse Brain	Mouse	FF	6.48/1000μm²	21.6%
Ovary	Human	FF	5.02/1000μm²	18.3%

Dataset Sources: All datasets are available from the 10x Genomics public datasets portal:

- Human Lymph Node FFPE
- Human Prostate FFPE
- Mouse Brain Fresh Frozen
- Human Ovary Fresh Frozen

Estimated Contamination by Expansion Distance

The following table provides estimated contamination rates at different expansion distances for Xenium and VisiumHD. For Xenium, estimates are derived from our interstitial density analysis and the Salas et al. correlation analysis. For VisiumHD, estimates are modeled based on the same diffusion physics, where contamination scales with the area captured beyond the true cytoplasmic boundary (approximately 5 μm from the nuclear edge for most cell types).

Table 2. Estimated contamination by expansion distance from nuclear boundary.

Expansion	Xenium	VisiumHD	Notes
0 μm (nuclear only)	<5%	<5%	Highest specificity
5 μm / +2 bins	~8%	~6%	Salas recommended limit
10 μm / +5 bins	~14%	~12%	Diminishing returns
15 μm / 8×8 μm bin	~18%	~15%	Default/standard

Note: VisiumHD estimates assume 2 μm bin size. The 8×8 μm standard binning captures approximately 15 μm diameter, comparable to Xenium default expansion. VisiumHD shows slightly lower contamination at equivalent expansion because transcripts are captured in fixed bins rather than precise probe locations, potentially limiting the detection of the most distal diffused molecules.

5. Supporting Evidence from Independent Analysis

Salas et al. recently published a comprehensive benchmarking study of Xenium data quality (Nature Methods, 2025). Their analysis employed a different methodology but reached conclusions consistent with the findings above.

The authors developed a metric called Negative Marker Purity (NMP), which quantifies specificity by measuring what fraction of transcripts are assigned to cell types that should express those genes, based on a single-cell RNA-seq reference. Cell types that contain transcripts of genes they should not express (according to the reference) are presumed to have received those transcripts through contamination.

Critically, the authors performed a correlation analysis examining how transcript signatures change as a function of distance from cell centroids. They found that transcripts located more than 10.7 μm from the cell centroid showed higher correlation with domain-specific background signatures than with cell-type-specific nuclear signatures. Given an average nuclear diameter of approximately 5 μm, this suggests that the biologically meaningful signal extends only about 5–6 μm beyond the nuclear boundary.

This finding has implications for the default Xenium segmentation, which expands nuclear masks by 15 μm. Under the Salas et al. analysis, approximately 9 μm of this expansion may capture signal that correlates more strongly with background than with the cell's true expression profile. Their recommendation to use Baysor segmentation with nuclear priors, rather than fixed radial expansion, is consistent with our interpretation that pericellular signal may be substantially contaminated.

A recent preprint by Bilous et al. (bioRxiv, 2025) provides complementary analysis of Xenium data quality across 40+ breast and lung tumor sections. Their study confirms that transcript

diffusion is a widespread and platform-agnostic challenge, and advocates for probabilistic segmentation approaches such as Baysor to mitigate misassignment. While their work offers valuable computational strategies, it does not employ a direct empirical quantification of diffusion magnitude. Our interstitial density approach provides a tissue-agnostic metric that can be applied to any Xenium or VisiumHD dataset without requiring a matched single-cell reference, offering a practical complement to reference-based methods.

6. Detection Sensitivity and Relative Abundance

The RNA contamination problem intersects with a separate but related issue: the differential detectability of nuclear versus cytoplasmic RNA. Understanding this distinction is important for interpreting apparent subcellular localization patterns and for designing segmentation strategies.

Detection Sensitivity

Hybridization-based detection methods, including those used in spatial transcriptomics and single-cell RNA-seq, may preferentially detect nuclear RNA for several reasons. Active transcription sites within the nucleus create high local concentrations of nascent RNA. The nuclear envelope provides partial protection from cytoplasmic RNases that degrade RNA. In FFPE samples, formaldehyde crosslinking may preferentially retain RNA within the nuclear compartment. These factors suggest that nuclear RNA is detected more efficiently than cytoplasmic RNA, independent of true abundance differences.

Relative Abundance

The relative abundance of nuclear versus cytoplasmic RNA varies substantially by gene and cell type. For many genes, mature mRNA is predominantly cytoplasmic, as transcripts are exported from the nucleus following processing. However, some transcripts are retained in the nucleus, and nascent (unspliced) transcripts are by definition nuclear. Single-nucleus RNA-seq studies have demonstrated that nuclear RNA captures 70–80% of the gene diversity observed in whole-cell preparations, suggesting that nuclear RNA provides a reasonably complete picture of cellular transcription.

Implications for Segmentation

The combination of potentially higher detection efficiency for nuclear RNA and the substantial contamination of pericellular regions suggests that nuclear-focused segmentation strategies may sacrifice relatively little true signal while substantially reducing noise. The transcripts lost by restricting analysis to nuclear regions may be predominantly either low-detection-efficiency cytoplasmic transcripts or proportionately enriched for contaminating transcripts from neighboring cells.

7. Recommendations

For Xenium Analysis

The default 15 μm nuclear expansion in Xenium may capture substantial contamination from neighboring cells. Based on our analysis and the Salas et al. findings, we suggest the following approaches:

- Consider reducing nuclear expansion to 5 - 6 μm, or using nuclear-only segmentation

for applications where specificity is prioritized over sensitivity.

- Baysor segmentation with nuclear priors may provide improved specificity by inferring cell boundaries from transcript density patterns rather than arbitrary geometric expansion.
- Report interstitial transcript density or estimated contamination fraction as a quality control metric, particularly when comparing datasets with different preparation methods.
- Exercise caution when interpreting rare cell populations, especially those defined by markers that are highly expressed in neighboring cell types.
- Findings of differential subcellular localization should be interpreted cautiously, as apparent cytoplasmic enrichment may reflect contamination rather than true biology.

For VisiumHD Analysis

VisiumHD data is captured in 2 μm bins, which offers flexibility in defining cell boundaries. Rather than expanding from nuclear centroids by arbitrary distances, we suggest a bin-overlap approach:

1. Identify nuclear masks using standard segmentation of DAPI or hematoxylin or similar nuclear staining.
2. Assign bins to cells based on overlap with nuclear masks (e.g., bins whose centroids fall within a nuclear mask, or bins with >50% area overlap).
3. Optionally include one ring of immediately adjacent bins, providing approximately 2 μm of pericellular signal.

This approach limits maximum expansion to 2–4 μm, compared to the 10 - 15 μm typical of other methods. The tradeoff is a reduction in transcripts per cell (estimated 20–30%) in exchange for improved specificity. For applications such as cell type identification and differential expression analysis, this tradeoff may be favorable.

8. Limitations and Future Directions

Several limitations of this analysis should be noted. The isotropic diffusion assumption may not hold in all tissue contexts; extracellular matrix, collagen fibers, and other structures could create directional diffusion bias. The interstitial density method may overestimate contamination in tissues with extended cellular processes (such as neurons) where legitimate transcripts exist outside the segmented cell body. Our estimates represent tissue-wide averages and do not capture spatial heterogeneity in diffusion rates.

Future work could address these limitations through development of spatially-aware contamination models, experimental validation using controlled systems with known transcript positions, and integration of contamination estimates into cell type annotation algorithms.

9. Conclusion

RNA diffusion is a significant source of noise in spatial transcriptomics data. The analysis of public Xenium datasets for interstitial noise suggests that 15–22% of transcripts may be assigned to incorrect cells, with higher rates in fresh frozen preparations. Independent findings

from Salas et al. support the conclusion that pericellular signal may be substantially contaminated. Nuclear-focused segmentation strategies may offer improved specificity for many applications, though the optimal approach will depend on the specific biological question and acceptable tradeoffs between sensitivity and specificity.

References

1. Marco Salas S, et al. Optimizing Xenium In Situ data utility by quality assessment and best-practice analysis workflows. *Nature Methods* 22, 813–823 (2025).

2. Bilous M, et al. From Transcripts to Cells: Dissecting Sensitivity, Signal Contamination, and Specificity in Xenium Spatial Transcriptomics. *bioRxiv* (2025). https://doi.org/10.1101/2025.04.23.649965