nature portfolio

Peer Review File

Comparison of imaging based single-cell resolution spatial transcriptomics profiling platforms using formalin-fixed, paraffin-embedded tumor samples

Corresponding Author: Dr Luisa Solis Soto

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
In this manuscript, Lermi et al. conducted a thorough comparison of state-of-the-art imaging-based spatial transcriptomics platforms in the context of gene coverage, assay technical performance, and cell segmentation and phenotyping. They explored the assay differences in both immune-hot and -cold tumor samples across different time points and leveraged an impressive multi-pronged strategy involving direct and indirect comparisons, quasi-orthogonal validation, and pathologist scoring to evaluate relative performance. As ST technologies are rapidly evolving and their comparative performance is of broad interest, this study addresses an important and timely question in the field and considers key methodological aspects of high practical relevance for researchers selecting an ST platform. While the study presents a valuable comparison, its novelty relative to prior work is not explicitly articulated, making it difficult to assess its unique contributions. Moreover, although significant obstacles are inherently expected when comparing commercial platforms with proprietary components, the study could have benefited from a better design that resolves at least some of the technical and analytical differences otherwise hindering comparability and a more thorough statistical analysis.

Major concerns:
• The authors largely adhered to manufacturer-recommended workflows for each platform, thus maximizing ultimate individual performance, yet they also attempted step-by-step comparisons in matters like cell segmentation and annotation which require standardization across platforms for a meaningful interpretation. It would have been more sensible to either compare ultimate output under manufacturer-determined ideal conditions (e.g. final cell annotation against a ground truth) thus reflecting task-specific bottom-line performance, or stepwise output under standardized conditions (e.g. cell segmentation with the same algorithm) thus reflecting technical physical performance of the assay in general. Specific points below.
• The comparability issue arises across multiple tenets of the manuscript:
◦ Lines 92-96: The study was designed in a manner that confounds tissue type and age. The older samples are both of lung adenocarcinomic origin while the more recent ones are both of pleural mesotheliomic origin.
◦ Lines 117-122: Cell filtering was imposed with different parameters (minimum transcript count and maximum cell size) for CosMx compared to MERFISH and Xenium.
◦ Lines 123-124: The CosMx, MERFISH, and Xenium platforms were implemented as 1000-, 500-, and 339-plex assays. The comparison of transcript counts and uniquely expressed genes across platforms is incommensurable and does not provide meaningful insight. A valid comparison would require normalizing for gene panel differences. Similarly, in lines 220-222, manual clustering was conducted using lineage markers that were selected based on the available gene panels which indicates that clustering results may differ simply due to a difference in the marker genes chosen.
◦ Other issues include the use of both negative controls and blank probes to compute FDR and the need for field of view selection with CosMx. While these are not easily resolvable and are inherent to the platforms, they should at least be mentioned in the limitations statement.
• The authors make a number of claims based on the results but did not conduct the requisite statistical testing. A few examples include:
◦ Lines 123-130 (Fig. 2a, b): No statistical testing can be discerned in the figure or text, to justify the claim that transcript count and uniquely expressed genes are different across the platforms.

◦ Lines 145-149 (Fig. 2d): No statistical tests to justify claims about FDR across platforms.
◦ Lines 163-168 (Fig. 3b): No statistical tests to justify claims about cell area size and variance.
◦ Lines 185-191 (Suppl. Table 3): No statistical tests to justify claims about co-expression of disjoint genes or the relationship to cell size. While statistical testing is not necessary for the latter, it is quantifiable.

Minor concerns:
• Lines 55-56: The authors cite other current studies comparing imaging-based ST assays (refs 9-11) but they don't state how this study provides new insights beyond those existing comparisons. It would strengthen the manuscript to explicitly highlight any novel aspects or methods that distinguish it from prior works. Of note, a certain degree of redundancy in such an emerging field is welcomed, but originality remains an important component of scientific discourse.
• Line 52-55, 68-69: The introduction could benefit from a slightly more detailed outlining of the fundamental assay principles of the ST platforms in question and how they differ from one another (e.g. difference between negative control probes and blank probes).
• Line 72: "mutually exclusive" should be softened with hedging language (e.g. largely or predominantly). CD8A and CD4 are often co-expressed in malignant cases but also in a small subset of lymphocytes under physiological conditions (10.1189/jlb.1RU0814-382, 10.1097/MD.0000000000022786, and PMID 9140307).
• Line 550-552: In general, the use of bulk RNA-seq, scRNA-seq, and DSP WTA in this study is insufficiently described in terms of both conceptual and technical methodology and is rather confusing. Points that need elaboration include the definition of "same cohort," how the ST data was aggregated when compared against bulk RNA-seq, whether a batch correction was done, etc.
• Line 81: The term 'platform-agnostic' suggests a standardized approach across all ST platforms, but several analyses in the study were platform-dependent (e.g. segmentation methods, varying gene panels, and distinct annotation strategies). Additionally, some analyses were omitted for specific platforms. I suggest either clarifying the actual platform-agnostic component of the approach or rephrasing.
• Line 84: Unclear what a 'high-complexity' setting is.
• Line 90-91: The authors may consider to explicitly confirm that there were no differences in the fixation, embedding, and dehydration protocols for all the samples.
• Line 137: To enhance readability, it would help to clarify the exact meaning of 'overlapping' here (e.g. target transcript count < negative control count?)
• Line 159-161: If the optimization of these pixel intensity and expansion distance parameters was 'by-eye,' it should be noted.
• Lines 174-179: This paragraph was difficult to follow and could be restructured and rewritten for clarity.
• Lines 180-183: It is not clear from this text which genes were paired together and I had to refer to the supplementary table for this information. I recommend that this information is written in the text as well as there aren't many pairs.
• Lines 261-269: It is unclear what 'matched' means exactly, how the scRNA-seq data was annotated to begin with, and what 'performed better' entails.
• Two figures are denoted 'Supplementary Fig. 2.'
• Lines 271-273: Unless I am misunderstanding this text, it is unclear what insight this adds to the study as correlations between modalities being compared themselves appears logically circular especially when they have a common source of labels transferred.
• Lines 281-285, 560-575: In general, the strategy for pathologist scoring could benefit from a more clear description. E.g. It completely escapes me what 'grouped several of the clusters of the following cell types into morphologically valid features' means; what are morphologically valid features? What does 'placed next to' mean? My main concern here is that it appears the data does provide the opportunity to produce a more quantitative scoring: based on the methods sections, raw images and instance segmentation masks are available and these could have been used to enable scoring on a per-instance (cell) basis for both cell identity and segmentation performance. This would not only provide a more granular assessment (and the computing of metrics like Jaccard, homogeneity, adjusted mutual information scores, etc.) but also enable statistical testing.
• Lines 343-344: while the cell boundary expansion algorithm is clear, I'm unsure how morphology markers were used for the other algorithms. A brief description of these algorithms would enhance readability.
• Line 357-358: It is unclear how noise eliminated 'distinct highly expressed' genes as that would be expected for lowly expressed genes.
• Line 371-372: It is again unclear how scRNA-seq data increased the annotation accuracy.
• Line 537-539: As a suggestion, the more recent Leiden algorithm improves on Louvain both in computational speed and partitioning quality (10.1038/s41598-019-41695-z). Also, as the study deals with feature spaces of varying dimensionality, I suggest fine-tuning the clustering hyperparameters (e.g. k in the k-NN graph generation step as well as the resolution term in the modularity function).

Reviewer #2

(Remarks to the Author)
The manuscript titled "Comparison of Imaging-Based Single-Cell Resolution Spatial Transcriptomics Profiling Platforms Using Formalin-Fixed, Paraffin-Embedded Tumor Samples" by Nejla Ozirmak Lermi etal., is well-written and highly relevant to the rapidly evolving field of spatial biology. Given the growing interest in tissue transcriptomics profiling technologies, this platform-to-platform comparison on the same tissue sections is both timely and well-justified for publication.
The study evaluates three imaging-based spatial transcriptomics (ST) platforms—CosMx (1,000 genes), MERFISH (500

genes), and Xenium (339 genes)—using tissue microarrays (TMA) from mesothelioma (MES, n=22) and non-small cell lung cancer (ICON, n=22). Due to inherent differences between platforms—such as tissue capture area, probe numbers, and inclusion of positive and negative controls—there are expected variations in precision and sensitivity. Despite these challenges, the authors have done a commendable job of systematically comparing assay performance across read counts, unique transcript counts, false discovery rate (FDR), cell segmentation, and annotation accuracy. A key comparison was made between ST-derived transcriptomic data and bulk RNA sequencing (Pearson correlation coefficient), as shown in Figure 4, where an average R = 0.515 for ICON and R = 0.63 for MESO samples was observed.
Key Findings:
1. Poor cell segmentation across platforms & variations in mean cell size (Figure 3a and b).
2. Significant differences in the F1-score of cell-type annotations, with Xenium-MM outperforming other platforms across stromal, immune, and tumor cells:
o Xenium-MM: >80%
o CosMx: ~60%
o MERFISH: ~40% (poor performance)
3. CosMx demonstrated poor cell-type annotation sensitivity in ICON samples, despite having the largest probe set (1,000 genes). However, it performed exceptionally well in MESO samples, suggesting that tissue quality significantly impacts ST performance
4. Extended dataset comparisons with NanoString GeoMx (same company as CosMx) reveal variable correlation quality, largely dependent on tissue block quality.

Minors:
1. Please normalized data for the read counts and unique gene counts in Figure 2 based on the probe numbers since CoxMx has 1, 000 while MERFISH at 500 and Xenium with 339.
2. A summary table with QC matrices such as total tissue area (more or less equal across the 3 platforms?), mean single cell size (Fig3b), gene transcripts/um2 will be helpful for the readers rather than all of them presented in each graph.

Conclusion:
This study presents a fair and thorough comparison of three ST platforms, offering valuable insights for researchers in spatial transcriptomics. The study provides a balanced, data-driven evaluation of ST platform performance using comprehensive QC metrics. The findings will be highly beneficial to scientists evaluating and selecting ST technologies for their research.

Reviewer #3

(Remarks to the Author)
Lermi et al., presents a systematic, head-to-head comparison of three commercially available imaging-based single-cell spatial transcriptomics platforms, CosMx, MERFISH, and Xenium (with both UM and MM segmentation approaches) using formalin-fixed, paraffin-embedded tumor samples from lung adenocarcinoma and pleural mesothelioma tissue microarrays. The study is comprehensive in scope and addressing several technical aspects including panel design differences, transcript counts, cell segmentation accuracy, noise levels and concordance with bulk RNA-seq and DSP data. In addition, the authors incorporate pathologists' evaluations to assess the morphological concordance of the cell-type annotations, thereby integrating both molecular and histopathological insights.
However, these are questions that the authors might consider to address:

1. The manuscript has detailed descriptions of sample preparation, data processing, and analytical pipelines. The authors might consider a summary table outlining key differences (e.g., panel sizes, negative control strategies, segmentation algorithms) across platforms could help readers quickly grasp the technical variances and their implications on downstream results.
2. The comparisons reveal significant differences in transcript counts, cell segmentation accuracy, and noise levels (FDR). In particular, the CosMx platform shows higher transcript counts but also higher background signal, while MERFISH appears to suffer from challenges in cell segmentation. It would strengthen the manuscript to discuss in greater depth how differences in panel design and segmentation strategy might independently contribute to these discrepancies. Are there additional controls or normalization strategies that could help disentangle these effects?
3. Although the authors provide extensive correlation analyses with bulk RNA-seq and DSP data, some of the correlation coefficients vary notably between TMAs (e.g., ICON vs. MESO sets). Further discussion on how sample age and tissue processing variability might influence these reproducibility metrics would be helpful. Moreover, details on replicates and the statistical power of these comparisons should be expanded.
4. the manuscript would benefit from more context regarding the scoring system (e.g., justification for the 1–5 scale) and how inter-observer variability was managed. Additionally, clarifying the calculation of F1-scores (which combine precision and recall based on pathologists' evaluations) will assist readers in understanding the robustness of the cell segmentation and annotation assessments.
5. A discussion of cost, ease of use, throughput, for tumor samples would be valuable. Furthermore, suggestions for future improvements or potential integration strategies to combine strengths of multiple platforms could be explored if possible.
6. Several figures contain dense information. Consider revising figure legends to be more concise and including schematic summaries or flowcharts to depict experimental design and analysis pipelines.
7. There are minor typographical errors and occasional formatting inconsistencies (e.g., spacing around units like "µm" and "µm²")
8. Although this is not peer-reviewed, the authors might want to discuss this pre-print in the discussion https://www.biorxiv.org/content/10.1101/2024.12.23.630033v1

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I appreciate the authors' thorough responses and revisions. All my concerns have been adequately addressed, and I have no further comments.

Reviewer #2

(Remarks to the Author)
The authors have thoroughly addressed major and minor reviewer concerns with scientifically sound justifications, appropriate new analyses, and improved manuscript clarity, especially on the cell segmentation issues for the platforms without the cell membrane staining.
While certain limitations on the platform-intrinsic biases) remain inherent to the study design, they are now well acknowledged and appropriately contextualized.

Only minor issue:
On page 7, line 150, the authors state: “We normalized transcripts and gene counts per cell with panel size.” However, on page 19, line 424-426, they write: “Observation of the highest transcript counts and highest number of uniquely expressed genes per cell with CosMx may be due to the higher number of genes in the panel.”
Could the statement on page 19 be interpreted as a conclusion drawn from the normalized transcript and gene counts per cell, as described on page 7?

Reviewer #3

(Remarks to the Author)
The authors have thoughtfully addressed my previous comments and substantially improved the manuscript.

Response to Reviewers

We thank the editor and reviewers for their thoughtful comments. We appreciate that all reviewers recognized the importance of comparison imaging-based single cell spatial transcriptomics platforms. The primary critiques from Reviewers 1, 2 and 3 relate to aspects of our analyses and discussion. We recognize that our first submission needed more detail in several areas, and that multiple points lacked adequate explanation. This revision makes substantial changes to improve the rigor and explanation throughout the manuscript, clarify the narrative, and format the text, figures, and tables based on the guidelines of Nature Communications.

Reviewer #1 (Remarks to the Author):

Major concerns:

Point 1: The authors largely adhered to manufacturer-recommended workflows for each platform, thus maximizing ultimate individual performance, yet they also attempted step-by-step comparisons in matters like cell segmentation and annotation which require standardization across platforms for a meaningful interpretation. It would have been more sensible to either compare ultimate output under manufacturer-determined ideal conditions (e.g. final cell annotation against a ground truth) thus reflecting task-specific bottom-line performance, or stepwise output under standardized conditions (e.g. cell segmentation with the same algorithm) thus reflecting technical physical performance of the assay in general. Specific points below.

Response to Point 1: Thank you for the valuable feedback. The primary goal of our study is to provide an objective comparison under manufacturer-determined ideal conditions as users of these platforms. In addition, we also think that it is informative to compare stepwise output provided by the manufacturers. These platforms employed different strategies for each step of data generation (e.g.: the biomarkers for cell segmentation are different for each of them, some has cell membrane staining, others do not), therefore utilizing the same algorithms to all platforms or performing additional customization can introduce additional biases that can reduce the objectivity of our study.

Point 2: The comparability issue arises across multiple tenets of the manuscript:
Lines 92-96: The study was designed in a manner that confounds tissue type and age. The older samples are both of lung adenocarcinomic origin while the more recent ones are both of pleural mesotheliomic origin.

Response to Point 2: Thank you for pointing this out. Most of our comparison across platforms are stratified by tissue types and ages. For comparison involving both tissue types and ages, we just reported the trends we observed without claiming statistical significance, i.e., Xenium appears to have more uniform quality on both old and recent tissues. Performing an age-matched comparison on each histology would be impractical to do, as there are often very limited materials. We included this as limitation of the study in lines 545-551.

Point 3: Lines 117-122: Cell filtering was imposed with different parameters (minimum transcript count and maximum cell size) for CosMx compared to MERFISH and Xenium.

Response to Point 3: Thank you for the valuable feedback. Although it would have been ideal utilizing the same parameters to all platforms, the platforms have significant differences in the number of transcripts among them. CosMx had the largest panel size, thus there were more cells with 1 transcript counts, which affected the accuracy of cell type annotations.
Here, we attached UMAPs of CosMx ICON data with removing cells based on transcript counts with a threshold of 30 and maximum cell size (reviewer Figure 1a), and with removing cells based on transcript counts set to a minimum of 10 only (reviewer Figure 1b). We lost 2454 cells with more aggressive data filtering process with CosMx data, however, the number of unknown cells that expressed multiple distinct cell markers together increased in the final cell type annotation with transcript counts set to a minimum of 10 only. Thus, we followed the aggressive threshold with CosMx to better perform cell type annotation step.

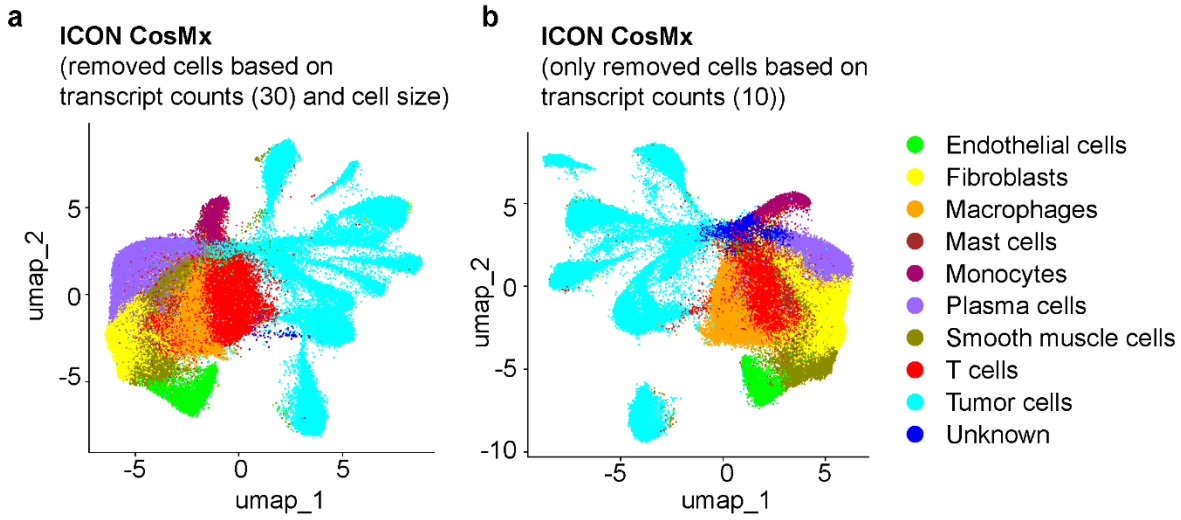

Figure 1: UMAP of CosMx ICON with different filtering parameters. a, b, Each color represents a cell type.

Point 4: Lines 123-124: The CosMx, MERFISH, and Xenium platforms were implemented as 1000-, 500-, and 339-plex assays. The comparison of transcript counts and uniquely expressed genes across platforms is incommensurable and does not provide meaningful insight. A valid comparison would require normalizing for gene panel differences.

Response to Point 4: Thank you for your suggestion to normalize transcript and gene counts per cell based on panel size. We normalized transcript counts and uniquely expressed genes per cell data based on panel size and included Mann-Whitney U test results to show significant differences between platforms in each TMA. We replaced Figure 2 a and b with normalized data in the revised manuscript.

Point 5: Similarly, in lines 220-222, manual clustering was conducted using lineage markers that were selected based on the available gene panels which indicates that clustering results may differ simply due to a difference in the marker genes chosen.

Response to Point 5: Thank you for the valuable feedback. We performed Louvain unsupervised clustering algorithm to cluster cells in each platform and annotated clusters with a cell type name based on highly expressed genes in each cluster. We believe that clustering results should not differ based on markers genes chosen for annotations. Our annotation results appear very robust, depending little on marker gene selection which are used for labeling specific cell types. Therefore, while cell type annotations of clusters may differ in the selection of marker genes, multiple gene markers should allow the detection of specific cell types; thus the differences between panel design don't affect cell type annotation drastically unless cell types that are not covered in the panel. For example, B cells expresses both *CD19* (https://www.proteinatlas.org/ENSG00000177455-CD19), *MS4A1* (https://www.proteinatlas.org/ENSG00000156738-MS4A1/single+cell) and *BANK1* (https://www.proteinatlas.org/ENSG00000153064-BANK1) genes, and they all can be a cell marker to identify B cells. When the platforms include multiple cell markers per cell type, there will be no issue to use different cell markers to identify a cell type in each platform. While CosMx and MERFISH included *CD19* and *MS4A1*, Xenium added *BANK1* in addition to *CD19* and *MS4A1* to their predesigned panels. Thus, we want to use additional genes in the manual cell type annotation strategy to improve the cell type annotation success. Our data indicate that the clustering results differ based on the platform's ability to detect gene expression that can allow appropriate clusters and cell type annotations, not in the marker genes chosen for labeling. We restructured the sentence in the revised version of text lines 290-297 to clarify the data analyses steps.

Point 6: Other issues include the use of both negative controls and blank probes to compute FDR and the need for field of view selection with CosMx. While these are not easily resolvable and are inherent to the platforms, they should at least be mentioned in the limitations statement.

Response to Point 6: Thank you for pointing this out. As you noted, negative controls and blank probes are prepared by companies to understand background issues of platforms and the selection and inclusion of these probes are under companies' control. Field of view (FOV) selection with CosMx is indeed a limitation of the assay. We have stated this in the results section and showed the disadvantages of FOV selection in Fig. 1c. We also included these differences as limitations of the assays in the limitation paragraph in line 542-548 in the revised version of text.

Point 7: The authors make a number of claims based on the results but did not conduct the requisite statistical testing. A few examples include:
Lines 123-130 (Fig. 2a, b): No statistical testing can be discerned in the figure or text, to justify the claim that transcript count and uniquely expressed genes are different across the platforms.

Response to Point 7: Thank you for suggestion. We included Mann Whitney U test to revised version of Fig. 2a, b.

Point 8: Lines 145-149 (Fig. 2d): No statistical tests to justify claims about FDR across platforms.

Response to Point 8: Thank you for suggestion. We rephrased the paragraph by giving FDRs and p values in text lines 180-189.

Point 9:Lines 163-168 (Fig. 3b): No statistical tests to justify claims about cell area size and variance.

Response to Point 9: Thank you for suggestion. We included statistical test results between cell area size in text lines 208-209 and rephrased the range assumptions.

Point 10: Lines 185-191 (Suppl. Table 3): No statistical tests to justify claims about co-expression of disjoint genes or the relationship to cell size. While statistical testing is not necessary for the latter, it is quantifiable.

Response to Point 10: We revised the Supplementary Table 3 (new version is Supplementary Table 5) by calculating co-expression of disjoint genes per TMA for each platform to run statistical tests. We included additional figures (revised Supplementary Fig. 1) to show differences between platforms with p values. We also expanded text in the manuscript results lines 239-245 and discussion lines 504-511.

Minor concerns:

Point 11: Lines 55-56: The authors cite other current studies comparing imaging-based ST assays (refs 9-11) but they don't state how this study provides new insights beyond those existing comparisons. It would strengthen the manuscript to explicitly highlight any novel aspects or methods that distinguish it from prior works. Of note, a certain degree of redundancy in such an emerging field is welcomed, but originality remains an important component of scientific discourse.

Response to Point 11: Thank you for suggestion. We expanded the introduction lines and 71-79.

Point 12: Line 52-55, 68-69: The introduction could benefit from a slightly more detailed outlining of the fundamental assay principles of the ST platforms in question and how they differ from one another (e.g. difference between negative control probes and blank probes).

Response to Point 12: Thank you for suggestion. We included them in the introduction lines 56-59. Also, we included a summary table to show differences between ST assays (revised Supplementary Table 1).

Point 13: Line 72: "mutually exclusive" should be softened with hedging language (e.g. largely or predominantly). CD8A and CD4 are often co-expressed in malignant cases but

also in a small subset of lymphocytes under physiological conditions (10.1189/jlb.1RU0814-382, 10.1097/MD.0000000000022786, and PMID 9140307).

Response to Point 13: Thank you for suggestion. We revised the text line 94, and discussion lines 509-511 accordingly.

Point 14: Line 550-552: In general, the use of bulk RNA-seq, scRNA-seq, and DSP WTA in this study is insufficiently described in terms of both conceptual and technical methodology and is rather confusing. Points that need elaboration include the definition of "same cohort," how the ST data was aggregated when compared against bulk RNA-seq, whether a batch correction was done, etc.

Response to Point 14: Thank you for the feedback. We used bulk RNA-seq and DSP WTA to evaluate gene expressions of ST data with concordance of orthogonal data. We used scRNA-seq of pleural mesothelioma to annotate cell types with automated cell type annotation method with MESO TMAs. We have clarified the data pre-processing of bulk RNA-seq, DSP WTA and scRNA-seq data in methods lines 711-723.

Point 15: Line 81: The term 'platform-agnostic' suggests a standardized approach across all ST platforms, but several analyses in the study were platform-dependent (e.g. segmentation methods, varying gene panels, and distinct annotation strategies). Additionally, some analyses were omitted for specific platforms. I suggest either clarifying the actual platform-agnostic component of the approach or rephrasing.

Response to Point 15: Thank you for your comment. We rephrased the sentence in lines 104-105.

Point 16: Line 84: Unclear what a 'high-complexity' setting is.

Response to Point 16: Thank you for your feedback. We have removed this part from text in lines 107-108.

Point 17: Line 90-91: The authors may consider to explicitly confirm that there were no differences in the fixation, embedding, and dehydration protocols for all the samples.

Response to Point 17: Thank you for your comment. All samples were processed into FFPE blocks using UT MD Anderson Cancer Center standard operating procedures at the time of sample collection. We cannot explicitly confirm that there were no differences (even slight differences) in the fixation, embedding, and dehydration protocols for all the samples because tumor blocks were collected retrospectively. However, we understand that exploring the effect of preanalytical variables in spatial transcriptomics is important and deserves systematic studies that are out of scope of the current study.

Point 18: Line 137: To enhance readability, it would help to clarify the exact meaning of 'overlapping' here (e.g. target transcript count < negative control count?)

Response to Point 18: Thank you for the feedback. We restructured the paragraph in lines 162-179.

Point 19: Line 159-161: If the optimization of these pixel intensity and expansion distance parameters was 'by-eye,' it should be noted.

Response to Point 19: Thank you for the feedback. We included this information in the revised version in lines 203-204.

Point 20: Lines 174-179: This paragraph was difficult to follow and could be restructured and rewritten for clarity.

Response to Point 20. Thank you for the feedback. We rephrased the paragraph in lines 216-234.

Point 21: Lines 180-183: It is not clear from this text which genes were paired together and I had to refer to the supplementary table for this information. I recommend that this information is written in the text as well as there aren't many pairs.

Response to Point 21: Thank you for the feedback. We stated these pairs in the text lines 235-236.

Point 22: Lines 261-269: It is unclear what 'matched' means exactly, how the scRNA-seq data was annotated to begin with, and what 'performed better' entails.

Response to Point 22: Thank you for the feedback. We restructured the sentences in lines 337-351. scRNA-seq data was performed with samples from pleural mesothelioma patients that were also in TMAs utilized for ST data generation. We used this scRNA-seq data as reference dataset for automated cell type annotation (Method). We included data preprocessing of scRNA-seq data in methods lines 703-706. Pleural mesothelioma tumor cells have unique biology, thus manual annotation of ST data with mesothelioma samples was challenging. Here, using the scRNA-seq reference dataset from patient samples that were matched with those in the TMAs used for ST data eased the cell type annotation challenges by incorporating a label transferring method from scRNA-seq.

Point 23: Two figures are denoted 'Supplementary Fig. 2.'

Response to Point 23: Thank you for the feedback. We reordered figures, tables and supplementary figures based on formatting guidelines of Nature Communications.

Point 24: Lines 271-273: Unless I am misunderstanding this text, it is unclear what insight this adds to the study as correlations between modalities being compared themselves appears logically circular especially when they have a common source of labels transferred.

Response to Point 24: Thank you for the feedback. Label transferring methods assign cell types to cells based on calculated prediction scores. Here, we would like to show the similarity

of assigned cell type between ST assays based on shared genes. When high correlation occurs between a cell type and multiple cell types in the opposite platform, the platform had noise and probe sensitivity issues. We expanded the paragraph in lines 352-363 to clarify it.

Point 25: Lines 281-285, 560-575: In general, the strategy for pathologist scoring could benefit from a more clear description. E.g. It completely escapes me what 'grouped several of the clusters of the following cell types into morphologically valid features' means; what are morphologically valid features? What does 'placed next to' mean? My main concern here is that it appears the data does provide the opportunity to produce a more quantitative scoring: based on the methods sections, raw images and instance segmentation masks are available and these could have been used to enable scoring on a per-instance (cell) basis for both cell identity and segmentation performance. This would not only provide a more granular assessment (and the computing of metrics like Jaccard, homogeneity, adjusted mutual information scores, etc.) but also enable statistical testing.

Response to Point 25: Thank you for the feedback. We reconstructed related paragraphs in results lines 368-401 and methods lines 738-743 to provide a better description of the pathologist scoring used in this study. We also agree with the reviewer that scoring per cell could provide a more quantitative data for statistical testing, however this method requires also a development of an accurate and validated computational pathology pipeline. Further studies that can allow a more accurate quantitative assessment are currently ongoing using pathology-supervised computational pathology and artificial intelligence approaches; however, this is out of the scope of this study.

Point 26: Lines 343-344: while the cell boundary expansion algorithm is clear, I'm unsure how morphology markers were used for the other algorithms. A brief description of these algorithms would enhance readability.

Response to Point 26: Thank you for the feedback. We stated the description of algorithms in the methods section lines 661-665 and added information in discussion lines 467-469.

Point 27: Line 357-358: It is unclear how noise eliminated 'distinct highly expressed' genes as that would be expected for lowly expressed genes.

Response to Point 27: Thank you for the feedback. We annotated clusters in ICON data with guidance of highly expressed genes in the cluster. When noise level is high or expression of cell marker genes is low, it is hard to get distinct highly expressed genes per cluster to use them in cell type annotation. We could not get distinct highly expressed cell type specific genes per cluster to separate B cells from T cells in CosMx dataset. We clarified it in the text lines 485-486.

Point 28: Line 371-372: It is again unclear how scRNA-seq data increased the annotation accuracy.

Response to Point 28: Thank you for the feedback. We rephrased the sentence in the text lines 499-503.

Point 29: Line 537-539: As a suggestion, the more recent Leiden algorithm improves on Louvain both in computational speed and partitioning quality (10.1038/s41598-019-41695-z). Also, as the study deals with feature spaces of varying dimensionality, I suggest fine-tuning the clustering hyperparameters (e.g. k in the k-NN graph generation step as well as the resolution term in the modularity function).

Response to Point 29: Thank you for your suggestions. For this study, we opted to analyze data with the standard Seurat pipeline. While we are aware of the Leiden clustering method and agree that it improves upon the Louvain method, Leiden algorithm requires additional environment adaptations to run with the Seurat pipeline. To determine if there was an impact on the clustering given our use of the Louvain method (reviewer Figure 2a), we ran the CosMx ICON data analysis with the Leiden algorithm and attached the UMAP plots of Leiden clusters below (reviewer Figure 2b). While we did not detect a difference between both algorithms in the detected cell types, we did observe an increase in the number of unknown cell types with the Leiden approach without any improvement in our ability to detect B cells in this specific dataset.

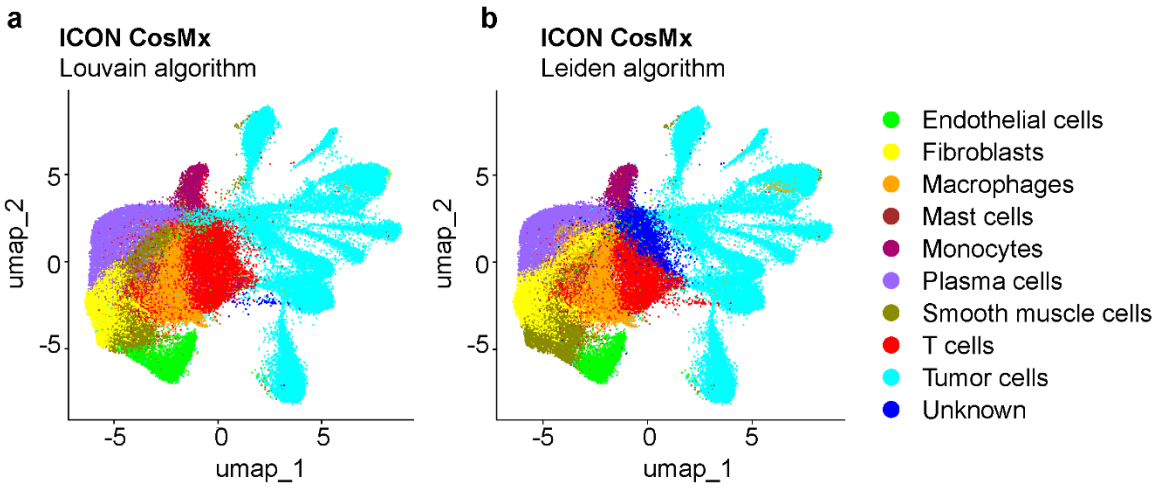

Figure 2: UMAP of CosMx ICON with different clustering algorithms. a, b, Each color represents a cell type.

In addition, increasing resolution in clustering and k in the k-NN graph generation step to get more fine-tuned clusters increased the number of clusters. Some of the clusters were not annotatable due to lack of non-significant differentially expressed genes.

Reviewer #2 (Remarks to the Author):

Point 1: Please normalized data for the read counts and unique gene counts in Figure 2 based on the probe numbers since CoxMx has 1, 000 while MERFISH at 500 and Xenium with 339.

Response to point 1: Thank you for your valuable feedback. We normalized data for the read counts and unique gene counts by panel size and revised Figure 2 accordingly.

Point 2: A summary table with QC matrices such as total tissue area (more or less equal across the 3 platforms?), mean single cell size (Fig3b), gene transcripts/um2 will be helpful for the readers rather than all of them presented in each graph.

Response to Point 2. Thank you for your valuable feedback. We included an additional supplementary table to present a summary table with QC metrics (Supplementary Table 4) in the revised version.

Reviewer #3 (Remarks to the Author):

Point 1: The manuscript has detailed descriptions of sample preparation, data processing, and analytical pipelines. The authors might consider a summary table outlining key differences (e.g., panel sizes, negative control strategies, segmentation algorithms) across platforms could help readers quickly grasp the technical variances and their implications on downstream results.

Response to Point 1: Thank you for your valuable feedback. We included additional Supplementary table to summarize differences (Supplementary Table 1).

Point 2: The comparisons reveal significant differences in transcript counts, cell segmentation accuracy, and noise levels (FDR). In particular, the CosMx platform shows higher transcript counts but also higher background signal, while MERFISH appears to suffer from challenges in cell segmentation. It would strengthen the manuscript to discuss in greater depth how differences in panel design and segmentation strategy might independently contribute to these discrepancies. Are there additional controls or normalization strategies that could help disentangle these effects?

Response to Point 2: Thank you for your comment, in the new version of the manuscript we have further discussed the differences among platforms. Briefly, high transcript counts per cell observed with CosMx could be related to bigger panel size and noise read counts due to non-specific probe bindings. We expanded the discussion section to cover this issue in lines 423-426. Cell segmentation accuracy is dependent on quality of morphology marker staining as we discussed in discussions lines 465-479. Cell segmentation algorithms of Xenium-UM was based on DAPI staining. Cell segmentation algorithms of CosMx, MERFISH and Xenium-MM depended on DAPI and morphology marker staining without including transcript localizations or H&E staining in the cell segmentation algorithms. Cell segmentation accuracy of ST assays may be increased using additional data in the cell segmentation algorithms beside morphology marker staining. Although additional controls or normalization strategies can improve the accuracy of ST data analysis, in their current form, they don't have effects on cell segmentation strategies. The primary goal of our study is to provide an objective comparison under manufacturer-determined ideal conditions as users of these platforms; nevertheless, additional studies using computational algorithms can further help to improve the accuracy of these ST platforms.

Point 3: Although the authors provide extensive correlation analyses with bulk RNA-seq and DSP data, some of the correlation coefficients vary notably between TMAs (e.g., ICON vs. MESO sets). Further discussion on how sample age and tissue processing variability might influence these reproducibility metrics would be helpful. Moreover, details on replicates and the statistical power of these comparisons should be expanded.

Response to Point 3: Thank you for your comment. ICON TMAs were constructed from lung adenocarcinoma years from 2016 and 2018 and MESO TMAs were constructed from mesothelioma years from 2016 to 2018. We inserted the sample preparation differences between bulk RNA-seq and DSP WTA assays in discussion lines 448-464. Correlation coefficients of bulk RNA-seq and ST assays varied between ST assays using same TMAs. In addition, there was variability between correlation coefficients of CosMx and MERFISH with bulk RNA-seq probably due to tissue age differences as we displayed in Figure 4. However, we don't see variation in correlation coefficients between ICON2 and MESO2 of Xenium-UM and Xenium-MM assays. Unfortunately, we don't have any replicates of old and new TMAs in all ST assays; thus, we cannot compute any statistical test to expand our findings.

Point 4: the manuscript would benefit from more context regarding the scoring system (e.g., justification for the 1–5 scale) and how inter-observer variability was managed. Additionally, clarifying the calculation of F1-scores (which combine precision and recall based on pathologists' evaluations) will assist readers in understanding the robustness of the cell segmentation and annotation assessments.

Response to Point 4: Thank you for your comment. We included Supplementary Table 6 to display the evaluations of three independent pathologists. We observed a moderate level of inter-observer variability in the scoring system of pathologists' evaluation in most categories. Thus, we used average scores of cell segmentation and cell type predictions to smooth discrepancies between inter-observer variabilities and performed F1-score calculation with averaged scores. We expanded the explanation of the visual inspection of cell phenotyping and segmentation in methods lines 733-743 and the calculation of F1-scores in methods lines 761-781.

Point 5: A discussion of cost, ease of use, throughput, for tumor samples would be valuable. Furthermore, suggestions for future improvements or potential integration strategies to combine strengths of multiple platforms could be explored if possible.

Response to Point 5: Thank you for your suggestions. In the updated version of the manuscript, we included a comment about the slide preparation difficulties in discussion line 427-432 to provide information on the additional training requirements of each ST platforms to mount tissue on slide scan area. Unfortunately, we did not have these ST platforms in house, thus, discussion of cost and throughput for tumor samples will not be accurate. We included a future directions paragraph to discussions line 552-559.

Point 6: Several figures contain dense information. Consider revising figure legends to be more concise and including schematic summaries or flowcharts to depict experimental design and analysis pipelines.

Response to Point 6: Thank you for your comment. We revised the figure legends and included analysis pipeline to Figure 1.

Point 7: There are minor typographical errors and occasional formatting inconsistencies (e.g., spacing around units like “µm” and “µm²”)

Response to Point 7: Thank you for your feedback. We addressed these inconsistencies in the revised version.

Point 8: Although this is not peer-reviewed, the authors might want to discuss this pre-print in the discussion https://www.biorxiv.org/content/10.1101/2024.12.23.630033v1

Response to Point 8: Thank you for your suggestions. This pre-print was published in bioRxiv after we submitted our manuscript for publication. We included this in the introduction lines 68-71.

Response to Reviewers

We thank the editor and reviewers for their second reviews of our manuscript. We appreciate for thoughtful feedback by the reviewers on the revised manuscript. Their comments have significantly strengthened the paper. In this revision, we answered the comment of Reviewer 2 and formatted the manuscript, figures and legends following the Nature Communications' author checklist.

Reviewer #2 (Remarks to the Author):

Point 1: On page 7, line 150, the authors state: "We normalized transcripts and gene counts per cell with panel size." However, on page 19, line 424-426, they write: "Observation of the highest transcript counts and highest number of uniquely expressed genes per cell with CosMx may be due to the higher number of genes in the panel." Could the statement on page 19 be interpreted as a conclusion drawn from the normalized transcript and gene counts per cell, as described on page 7?

Response to Point 1: Thank you for pointing this out. We replaced the sentence as "Observation of the highest transcript counts and highest number of uniquely expressed genes per cell with CosMx may be due to transcript detection chemistry and probe design strategies." In line 369-370 in revised text.