nature methods

Registered Report

https://doi.org/10.1038/s41592-026-03317-1

Large-scale quantitative assessment of tissue preparation and staining conditions for robust multiplexed imaging of FFPE tissue samples

Received: 6 June 2024

Accepted: 7 October 2026

Published online: 26 November 2026

Check for updates

A list of authors and their affiliations appears at the end of the paper

Multiplexed protein imaging or spatial proteomics enables the simultaneous measurement of over 50 proteins in intact tissues at single-cell resolution, providing insights into tissue organization and cellular interactions. However, variation in tissue preparation and staining protocols remains a major source of technical variability, limiting reproducibility and comparability across studies and platforms. Here we present a large-scale, multisite and multi-platform benchmarking study of tissue preparation conditions for formalin-fixed paraffin-embedded (FFPE) tissues. We systematically evaluated 24 combinations of antigen retrieval and antibody staining conditions using a 27-marker immune panel and ranked performance using quantitative metrics that capture signal intensity, variability and marker failure. Tris–EDTA-based antigen retrieval combined with overnight antibody staining at 4 °C consistently achieved the highest performance, whereas several commonly used alternatives showed reduced robustness. Validation across additional academic and industry laboratories and multiple imaging technologies confirmed the generalizability of these findings across tissues and platforms. We further show that marker signal intensity correlated with greater spatial heterogeneity, improving the detection of diverse cell populations. Together, our results provide practical guidelines and a standardized benchmarking strategy to improve reproducibility, harmonize workflows and facilitate cross-study comparisons in spatial proteomics.

Advancements in multiplexed imaging technologies, also known as spatial proteomics, have enabled the simultaneous detection of more than 50 protein markers at single-cell resolution in a tissue section[1,2]. Such techniques are extremely valuable in the pursuit of a diverse range of biological questions, by retaining the spatial locations of individual cells, unlike dissociative single-cell methods such as mass cytometry[3] and multiparametric flow cytometry[4], and providing a direct readout of protein expression along with their subcellular localization. As such, multiplexed imaging techniques represent valuable tools for quantifying single-cell phenotypes and states, and cell–cell or higher-order interactions within the context of spatially orchestrated processes.

Current multiplexed protein imaging methods can be broadly categorized into two main approaches: mass spectrometry-based and iterative imaging-based methods. Mass spectrometry-based methods, such as imaging mass cytometry (IMC)[5] and multiplexed ion beam imaging (MIBI)[6,7], rely on mass spectrometry readouts to simultaneously localize metal-conjugated antibodies. Iterative imaging methods involve multiple cycles of staining, imaging and signal removal to overcome spectral emission overlap limitations of traditional immunofluorescence. Some examples of iterative imaging methods include codetection by indexing (CODEX) that uses antibodies conjugated to unique DNA oligonucleotide barcodes[8,9], cyclic immunofluorescence[10–12] that

e-mail: sjiang3@bidmc.harvard.edu

involves sequential multichannel imaging followed by fluorophore inactivation and immunostaining with signal amplification by exchange reaction (Immuno-SABER)[13] that achieves highly multiplexed signal amplification by hybridizing single-stranded DNA concatemers to DNA-barcoded antibodies, followed by cycles of hybridization and stripping of fluorophore-bearing imager strands to the concatemers for imaging. Many of these methods are compatible with formalin-fixed paraffin-embedded (FFPE) tissues, making it possible to tap into the tremendous archives of human biopsy and tissue collections worldwide[14]. The development of these techniques has yielded new insights into biological processes in both healthy and diseased states that were previously impossible, contributing to important advancements in biological knowledge and therapeutic strategies[8,9,15–20].

The experimental workflow for spatial proteomics approaches has several parallels with that of traditional immunohistochemistry (IHC) and immunofluorescence, and can be summarized as follows: (1) antigen retrieval, (2) blocking and optional photobleaching, (3) staining with an antibody cocktail followed by an optional post-stain fixation, (4) tissue imaging and (5) data processing[21]. Unlike in IHC where a specific series of conditions are usually optimized for a single antibody clone, in spatial proteomics, all the antibodies need to perform robustly in a single condition. Given the limited quantitative dynamic range in traditional single-target IHC approaches and challenges in high-plex antibody selection, staining and imaging, the current field lacks a systematic and thorough quantitative evaluation of ideal tissue preparation and staining conditions. Across spatial proteomics pioneering laboratories[5,6,12,21–23] and commercial vendor recommendations, there are substantial variations in preferred or recommended tissue preparation conditions, including antigen-retrieval buffers and durations, and antibody staining temperature and durations. These factors together can limit the potential success of the experiments and influence data interpretation due to variations in staining quality. In addition, variations in protocols often contribute to the lack of consistency in the field and hinder robust analysis of large-scale data collection such as in clinical trials[24]. Given the growing interest in the adoption of multiplexed imaging technologies in the study of various archival tissues, we sought here to perform a comprehensive and systematic benchmarking of various tissue preparation and antibody staining conditions to guide new and seasoned users in performing these resource-intensive experiments. Such benchmarking efforts will also help guide the standardization of protocols to maximize the effectiveness and reproducibility of spatial proteomics results within and across laboratories and institutions.

A key step required for the immunostaining of FFPE tissues is antigen retrieval, which involves the reversal of crosslinks formed during tissue formalin fixation, to enable the accessibility of epitopes for antibody detection[25]. Earlier approaches for antigen retrieval involved the use of proteolytic enzymes that directly degrade crosslinks, but this step is detrimental to protein epitopes and tissue morphology and can be hard to control in terms of enzymatic activities due to batch variations[26]. For this reason, heat-induced epitope retrieval (HIER), which involves heating the tissue sections in a solution, is often preferred and has been shown to greatly enhance the detection of antigens in FFPE tissues[26–28]. Key variables of HIER include the pH and salt composition of the retrieval buffer, and the duration and temperature of the retrieval process, which can have a notable impact on both antigen recovery and tissue morphology[26,29,30]. How these variables affect the successful detection of high-plex antibody panels in the context of multiplexed imaging remains unclear, and the systematic evaluation and identification of optimal antigen-retrieval conditions will be critical.

To better assess the impact of the buffer on antigen retrieval, we have selected two HIER buffers for testing: the 'universal retrieval buffer', citraconic anhydride (pH 7.4) buffer[31,32] and the commercial Tris–EDTA (pH 9.0) target retrieval solution. Citraconic anhydride has been shown to be superior to other antigen-retrieval buffers, such as 0.01 M citrate buffer (pH 6.0) and 0.1 M Tris-HCl containing 5% urea (pH 9.0), in conventional IHC of select single-plex targets[31,32], although its utility for high-plex antibody imaging has not been thoroughly evaluated. The commercial Tris–EDTA (pH 9.0) target retrieval solution has been routinely used for both MIBI and CODEX across multiple laboratories, including ours[7,9,21,22,33–36], where the alkaline pH and addition of EDTA have been demonstrated to improve antigen retrieval compared to 0.01 M citrate buffer (pH 6.0) and 0.1 M Tris-HCl (pH 8.0)[26,27].

Apart from the type of HIER buffer used, the optimal duration of antigen retrieval to achieve maximal antigen recovery and minimal tissue damage for each buffer should be empirically determined. Generally, antigen retrieval can be enhanced by extending the duration of treatment to promote reverse-crosslinking, albeit until a point after which continued treatment may not offer substantial improvement or may even be detrimental to antigen recovery[26]. This effect is highly dependent on the type of HIER buffer used, where alkaline pH, EDTA-containing buffers generally require shorter treatment time compared to lower pH, citrate-containing buffers[26].

Another key step that can affect the success of antigen detection in FFPE tissues is antibody staining. Titration of the antibody concentration is highly recommended to obtain the ideal signal-to-noise ratio, and to determine the highest antibody dilution to minimize waste and reduce costs while maintaining a meaningful dynamic range for downstream quantification[37,38]. Nevertheless, at identical antibody concentrations, it is unclear how staining conditions such as temperature and duration of antibody staining affect the quality of staining, particularly of a high-plex antibody cocktail during multiplexed imaging. Many conventional IHC and immunofluorescence protocols recommend incubating with primary antibodies at room temperature for 30 minutes to 60 minutes, with an option of staining at 4 °C overnight[38,39]. A recent study performed for high-parameter flow cytometry demonstrated that incubating cells at 4 °C overnight for antibody staining significantly improves staining quality while reducing the concentration of antibody required, enhances sensitivity for lowly-expressed antigen and reduces inter-experiment variability[40]. We postulate that these differences are likely consistent in the context of multiplexed imaging, which relies on similar antibody-binding thermodynamics.

In this study, we conducted a comprehensive, side-by-side quantitative assessment of the impact of different HIER and antibody staining conditions on the staining performance of 27 immune-centric markers on serial FFPE sections of human lymphoid tissues. To reduce potential confounders due to antibody conjugation efficiencies and lot variations, we partnered with Cell Signaling Technology to develop oligo-tagged, CODEX-compatible antibodies purified using fast protein liquid chromatography in a lot-controlled fashion. Twenty-four different conditions (permutations of two antigen-retrieval buffers, three retrieval durations and four staining conditions; Supplementary Table 1) were tested on serial human tonsil tissue sections on the Akoya PhenoCycler-Fusion at BIDMC-Harvard. In parallel, 12 of these 24 conditions were tested on the RareCyte Orion platform at Stanford and 6 conditions were tested on the Leica BOND RX automated staining device with imaging on the Akoya PhenoCycler-Fusion at Roche during this initial grid search (Supplementary Fig. 1a). We then established a reproducible quantification pipeline to assess the impact of the staining conditions on marker performance, including marker staining intensity, ratio of signal intensity within versus outside the cell mask and cell annotation specificity and sensitivity. Following quantitative evaluation of the data from all sites, the top three and bottom one conditions were identified from the BIDMC-Harvard initial optimization data: the top-ranked protocol was 20 minutes of HIER with a Tris–EDTA (pH 9) buffer and overnight antibody staining at 4 °C, the second-ranked protocol was 40 minutes of HIER with a Tris–EDTA (pH 9) buffer and overnight antibody staining at 4 °C, and the third-ranked protocol was 10 minutes of HIER with a Tris–EDTA (pH 9) buffer and overnight antibody staining at 4 °C. The bottom-ranked

protocol was 20 minutes of HIER with citraconic anhydride (pH 7.4) with 1 hour of room temperature (25 °C) antibody staining. Cross-site validation of the top three and bottom one conditions was then performed across seven additional academic and industry laboratories, including the Agency for Science, Technology and Research (A*STAR), the University of Kentucky, Novartis, Roche and Stanford University. These evaluations were conducted using a range of multiplexed imaging platforms, namely Standard BioTools IMC, Ionpath MIBI, Akoya PhenoCycler-Fusion, RareCyte Orion and Lunaphore COMET (Supplementary Fig. 1a). Across sites, the condition involving 20-min HIER using a Tris–EDTA retrieval buffer coupled with an overnight antibody incubation at 4 °C (where applicable) consistently yielded robust performance across multiple tissue types and platforms. Each site also used its own optimized antibody panels on lymphoid and study tissue serial sections to ensure the generalizability of the findings. This cross-site validation was essential for confirming the reproducibility of the Registered Report and for guiding end-users on the selection of optimal protocols for different instruments. Evaluation of antibody performance and cell type annotation specificity was supported by multiple trained pathologists (D.N., B.H., K.H.) for data produced at BIDMC-Harvard, and subsequently by in-house pathologists at each validation site. We present this Registered Report to guide the development of a standardized and robust protocol for multiplexed tissue imaging methods, along with a reproducible quantification analysis pipeline, for enhanced performance and reduced variability within and across laboratories.

Results

Pilot experiment comparing four HIER and antibody staining conditions

To demonstrate the impact of HIER duration and antibody staining conditions on antibody performance in multiplexed tissue imaging, we first tested 4 out of the 24 proposed conditions on adjacent sections of human FFPE tonsil tissue (Fig. 1a and Supplementary Table 1). The selected conditions included the following: 20-min Tris–EDTA HIER, 37 °C 1-h stain; 20-min Tris–EDTA HIER, 4 °C overnight stain; 10-min Tris–EDTA HIER, 4 °C overnight stain; and 40-min Tris–EDTA HIER, 4 °C overnight stain. All slides and markers were included for analysis after confirming that there were no technical errors and the markers were working for at least one condition (Supplementary Fig. 1b,c). Cell segmentation and extraction of single-cell features were performed on two aligned regions of interest (ROIs) per slide (Fig. 1a). Quantitative assessment of antibody performance and visualization of results were then performed (Fig. 1a).

Visual assessment of the 23 functional and phenotyping markers (Fig. 1b and Supplementary Table 2) showed detectable staining of all markers for the 20-min HIER 4 °C overnight slide (Supplementary Fig. 1d). In contrast, some markers failed to perform in the other conditions; DC-SIGN was not detectable in the 20-min HIER, 37 °C 1-h stain and the 10-min HIER, 4 °C overnight stain slides, while cytokeratin was not detectable in the 10-min HIER, 4 °C overnight stain slide (Fig. 1c and Supplementary Fig. 1e). Side-by-side comparison of the mean Hoechst-normalized signals across all four conditions showed that, in general, 20-min HIER, 4 °C overnight stain yielded the highest staining intensities for 14 out of 23 markers (Fig. 1c and Supplementary Fig. 1f,g). While statistically significant differences were observed for all conditions and markers using the Kruskal–Wallis test with Dunn's post hoc test, large effect sizes (Cohen's *d* within 95% confidence interval >0.8) were notably observed for 20-min HIER, 4 °C overnight stain against the other three conditions for CD138 and H3K27ac (Supplementary Table 3). This suggests that the 20-min HIER, 4 °C overnight stain resulted in the highest staining intensity across all four conditions and the effect was strongest for CD138 and H3K27ac. In contrast, granzyme B, FOXP3, cytokeratin, CD45RA, CD4 and CD31 showed the highest intensities for the 20-min HIER, 37 °C 1-h staining condition (Cohen's *d* within 95% confidence interval >0.8 for granzyme B and CD4), and HLA-DRA, H3K27me3 and CD11c showed the highest intensities for the 10-min HIER, 4 °C overnight staining condition (Fig. 1c, Supplementary Fig. 1f,g and Supplementary Table 3). We also performed orthogonal validation of our results using an alternative cell segmentation method, cellXpress2[41]. Both Mesmer and cellXpress2 yielded similar marker distribution patterns and provided the same rankings of the conditions for all the markers (Supplementary Fig. 1g,h), showing that the staining conditions, rather than the specific image processing or cell segmentation methods used, contribute to the major differences in staining results.

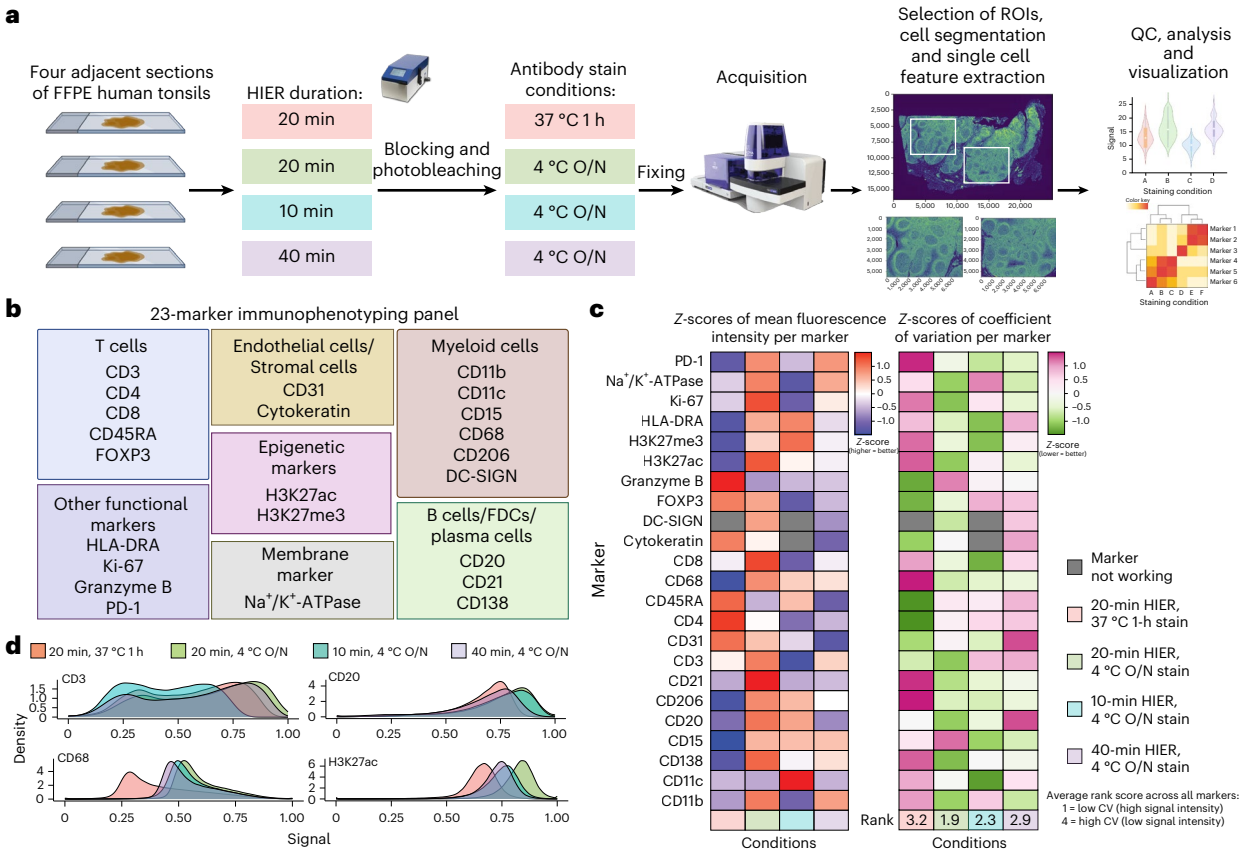
The coefficients of variation (CVs) of each marker across all conditions were also calculated and compared, as an indicator of protein marker signal intensity and dispersion. To quantify the overall average influence of the condition on all markers, each condition was ranked from lowest to highest CV per marker, where 1 denotes lowest CV for that marker and thus highest staining intensity while 4 denotes highest CV and lowest signal intensity. The final average score was calculated. The 20-min HIER, 4 °C stain generally yielded the lowest CV across the markers on average (with a final average rank score of 1.9), while the 20-min HIER, 37 °C stain resulted in the highest CV across markers on average (with a final average rank score of 3.2) (Fig. 1c and Supplementary Table 4).

Markers were then selected for visual comparison, namely CD3 (T cells), CD20 (B cells), CD68 (myeloid cells) and H3K27ac (acetylated histones). Density distributions of the markers and visual comparison of the images revealed that the signal intensities for the markers were the highest for the 20-min HIER, 4 °C overnight staining condition, while the signal intensity for CD3 was the dimmest for the 10-min HIER, 4 °C overnight staining condition and those of CD20, CD68 and H3K27ac were the dimmest for the 20-min HIER, 37 °C 1-h staining condition, which are aligned with the CV rank scores (Fig. 1c–e and Supplementary Fig. 1i).

We next determined the effects of staining conditions on the signal-to-noise ratios of specific markers, since higher background might account for the intensity difference observed in the comparison of the means in Fig. 1c (for example, CD4, Supplementary Fig. 1j, upper panel). Two line scans across 2–5 cells in the same ROI were taken and the signal intensities for CD4, CD3, CD8 and granzyme B were compared. As revealed in the representative ROIs and line graphs, the 20-min HIER, 37 °C 1-h staining condition had high background and

Fig. 1 | Pilot experiment comparing four HIER and antibody staining conditions. a, Workflow for the pilot benchmarking experiment. Four different HIER durations and antibody staining conditions were tested on adjacent FFPE human tonsil sections, and their effects on staining quantified and analyzed. **b**, Protein markers in the 23-marker immunophenotyping panel used in this experiment. PD-1, Programmed cell death protein 1; FDC, follicular dendritic cell. **c**, Heatmaps showing the relative mean fluorescence intensities (left, blue to red) and CVs (right, pink to green) of all 23 phenotypic and functional markers across the four different HIER and staining protocols with *Z*-score transformation. Numbers reflect the average rank score achieved. Lower score indicates higher staining intensity. Grayed-out entries indicate failed markers. See Supplementary Fig. 1d,e for representative images. **d**, Overlay of density plots for the four different HIER and staining conditions of the four example markers shown in **e**. **e**, Representative images of tonsils treated with the four different HIER and staining protocols showing markers for T cells (CD3, green), B cells (CD20, magenta), myeloid cells (CD68, blue) and histone acetylation (H3K27ac, yellow). Corresponding segmentation shown in the inset. Scale bars, 100 µm (main), 20 µm (inset). All images are displayed using identical scales, crop areas and minimum–maximum intensity settings across slides. O/N, overnight; QC, quality control. Panels created in BioRender: **a**, Parmelee, L. https://biorender.com/z0y3ijm (2026); **b**, Parmelee, L. https://biorender.com/hv0m7ph (2026).

https://doi.org/10.1038/s41592-026-03317-1



poor signal-to-noise ratio for CD4 and granzyme B. Across all ROIs and markers, the 10-min HIER, 4 °C overnight staining condition showed the lowest marker signal. Overall, the best signal-to-noise ratios were observed for the 20-min HIER, 4 °C overnight as well as 40 min HIER, 4 °C overnight conditions (for example, CD4, Supplementary Fig. 1j, lower panel).

We also performed spatial clustering and sampling analysis on field of view (FOV) 1 of the 20-min HIER, 4 °C overnight stain slide to demonstrate that canonical cell types such as B cells, T cells, macrophages and endothelial cells can be identified and their spatial localization can be validated (Supplementary Fig. 1k,l). The tau value, defined as the number of FOVs needed to recover a significant proportion of cell phenotypes present in the tissue, was also determined (Supplementary Fig. 1m). In the complete optimization study involving the 24 conditions below, the tau value and the slope of the relationship between tau and FOV widths (α parameter) were compared across different experimental conditions.

Testing the stability of antibodies postlyophilization

To assess the stability of the antibody cocktail following lyophilization, we compared imaging data of tonsil sections stained under three conditions: (1) fresh antibody cocktail applied manually ('fresh manual'), (2) lyophilized antibody cocktail applied manually ('lyophilized manual') and (3) lyophilized antibody cocktail applied using the Opentrons OT-2 Liquid Handling Robot with automated washes ('lyophilized automated') (Supplementary Fig. 2a). These comparisons demonstrated that lyophilization of the antibody cocktail is feasible and largely preserves staining performance. Although some markers exhibited reduced signal intensity in the lyophilized manual condition (Supplementary Fig. 2b–d), the signals remained sufficiently robust for visual inspection and reliable cell type identification, supporting the suitability of lyophilization for long-term antibody storage.

Unexpectedly, all markers displayed lower signal intensities in the lyophilized automated condition compared to the lyophilized manual condition (Supplementary Fig. 2b–d), suggesting that the automated workflow may introduce additional signal loss, likely due to the reduced available staining volume as a result of the capillary action from the setup we had available. As a consequence, all subsequent tissue staining at BIDMC-Harvard was carried out using the 'lyophilized manual' protocol. For both the initial optimization and validation experiments, the antibody cocktail was prepared once and lyophilized in a single batch to maximize comparability and reproducibility across slides and experiments.

Initial optimization of HIER and antibody staining conditions across three sites

For the initial optimization, a total of 24 different conditions, which consist of permutations of 2 antigen-retrieval buffers, 3 retrieval durations and 4 staining conditions, were tested on serial human tonsil tissue sections using the Akoya PhenoCycler-Fusion platform at BIDMC-Harvard. In parallel, 12 of these 24 conditions were tested on human lymph node sections using the RareCyte Orion platform at Stanford and 6 conditions were tested on human tonsil sections using the Leica BOND RX automated staining device with imaging on the Akoya PhenoCycler-Fusion system at Roche (Fig. 2a). Notably, for the testing at Roche, the Leica BOND automated staining platform required the use of system-compatible buffers (BOND Epitope Retrieval Solution 1, which is a citrate-based, pH 6 buffer and BOND Epitope Retrieval Solution 2, which is a EDTA-based, pH 9 buffer). As such, only three of the conditions involving the EDTA-based, pH 9 buffer were comparable to the other two sites (Fig. 2a).

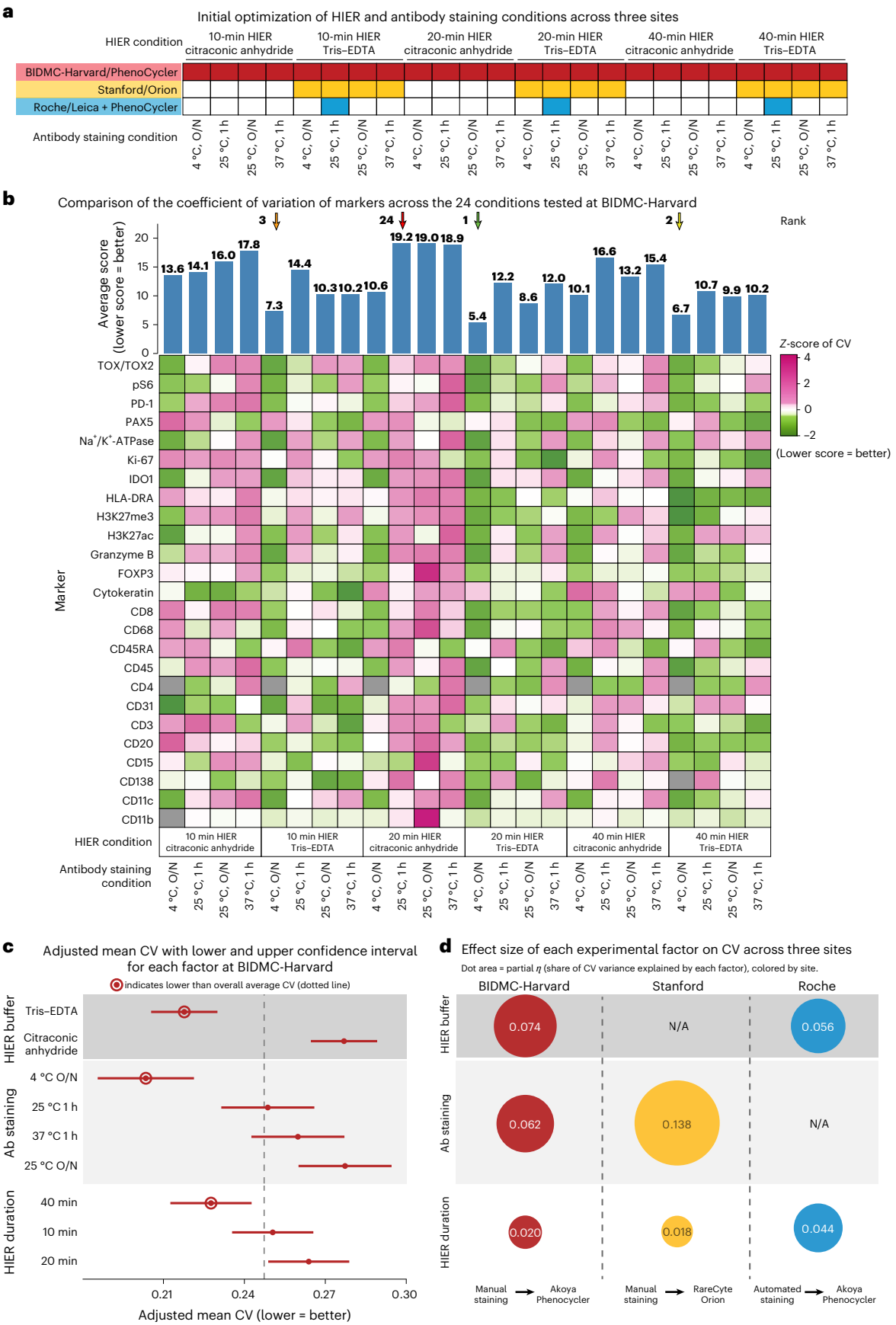
To evaluate the relative performance of the 24 conditions tested at BIDMC-Harvard, the panel of 27 antibodies (Supplementary Table 5) was first visually assessed by three board certified pathologists (D.N., B.H., K.H.) (Supplementary Table 6) and markers that failed to stain specifically for all conditions were excluded from downstream analysis (DC-SIGN and CD56). The mean fluorescent intensities and CV of each of the remaining 25 markers were then determined and compared across the 24 conditions (Fig. 2b and Supplementary Fig. 3). To assess the relative contribution of each experimental factor to staining intensity variation, a factorial linear model was fitted to the CV values across the 24 conditions and 25 markers. Estimated marginal means (EMMs) of CV were then calculated for each factor while averaging over all remaining factors (Fig. 2c). Across all HIER durations and buffer types, antibody staining at 4 °C overnight yielded the lowest model-adjusted CV, indicating the highest overall staining intensities among the staining protocols tested (Fig. 2c). Similarly, HIER using Tris–EDTA was associated with a lower model-adjusted CV than HIER using citraconic anhydride (Fig. 2c). Among the HIER durations tested, 40-min HIER produced the lowest model-adjusted CV compared with 10-min and 20-min HIER (Fig. 2c).

To quantify the relative importance of each factor, effect sizes were estimated using partial η^2 (Fig. 2d). Among the experimental variables, HIER buffer type had the largest effect on CV (partial $\eta^2 = 0.074$), followed by antibody staining protocol (partial $\eta^2 = 0.062$), whereas HIER duration contributed comparatively little to the observed variation (partial $\eta^2 = 0.020$). These findings suggest that optimization of HIER buffer composition and antibody staining conditions has a greater impact on staining reproducibility than varying HIER duration within the range tested. Similar trends were observed in independent datasets generated at Stanford and Roche, where HIER duration consistently exhibited smaller effect sizes than either antibody staining condition or HIER buffer type (Fig. 2d).

The 24 conditions tested at BIDMC-Harvard were then ranked based on their overall average CV values across the 25 markers (Fig. 3a and Supplementary Table 7). Based on this ranking, the top

Fig. 2 | Initial optimization results across HIER and antibody staining conditions. a, Overview of HIER and staining conditions tested at BIDMC-Harvard (red), Stanford (yellow) and Roche (blue). **b**, Heatmap of *Z*-scored CVs for all 25 phenotypic and functional markers across the 24 different HIER and staining protocols evaluated with the Akoya PhenoCycler-Fusion at BIDMC-Harvard. Numbers and bar plots indicate the average rank score achieved by each condition. Arrows indicate the top-ranked (green, 1), second-ranked (yellow, 2), third-ranked (orange, 3) and bottom-ranked (red, 24) conditions. Lower score indicates higher staining intensity. Grayed-out entries indicate failed markers. **c**, Bar plots showing the EMM CV for each HIER duration, antibody staining condition and HIER buffer, computed from a factorial linear model of CV as a function of HIER buffer, HIER duration and antibody staining condition fitted to the BIDMC-Harvard data ($n = 592$ per-condition-and-marker CV). Points indicate model-adjusted mean CVs (EMMs); horizontal bars indicate 95% confidence intervals. The EMMs (95% confidence interval) are: for HIER buffer, Tris–EDTA 0.218 (0.205–0.230), citraconic anhydride 0.277 (0.265–0.289); for HIER duration, 10-min 0.251 (0.235–0.266), 20-min 0.264 (0.249–0.279), 40-min 0.228 (0.213–0.243) and for antibody staining, 4 °C overnight 0.203 (0.186–0.221), 25 °C 1-h 0.249 (0.231–0.266), 37 °C 1-h 0.260 (0.243–0.277), 25 °C overnight 0.277 (0.260–0.295). The dotted line corresponds to the overall mean CV. Lower CV indicates higher staining intensity. **d**, Dot plot showing the experimental factor effect size (HIER buffer, antibody staining condition and HIER duration), measured as partial eta-squared (η_p^2) from Type II ANOVA of additive linear models fitted separately to the initial optimization data from BIDMC-Harvard ($n = 592$ per-condition-and-marker CV), Stanford (Tris–EDTA only; $n = 143$ per-condition-and-marker CV) and Roche (25 °C 1-h staining only; $n = 60$ per-condition-and-marker CV). Dot size represents the effect size. η_p^2 (95% confidence interval) for all sites are: BIDMC-Harvard HIER buffer 0.074 (0.043–1.00), antibody staining 0.062 (0.031–1.00), HIER duration 0.020 (0.004–1.00); Stanford antibody staining 0.138 (0.046–1.00), HIER duration 0.018 (0.000–1.00); Roche HIER buffer 0.056 (0.000–1.00), HIER duration 0.044 (0.000–1.00). Confidence intervals for η_p^2 are one-sided (upper bound fixed at 1).

https://doi.org/10.1038/s41592-026-03317-1



three conditions were identified to be 20-min Tris–EDTA HIER with 4 °C overnight stain, 40-min Tris–EDTA HIER with 4 °C overnight stain and 10-min Tris–EDTA HIER with 4 °C overnight stain, while the bottom condition with the highest overall CV was 20-min citraconic anhydride HIER with 25 °C 1-h stain (Fig. 3a). Example images with selected markers also highlight the impact of different HIER and antibody staining conditions on staining intensities visually (Fig. 3a and Supplementary Fig. 4). For the top three conditions that all involved HIER with Tris–EDTA and 4 °C overnight antibody stain but with variable HIER durations, the staining intensities were visually very similar and the effect sizes of the differences tended to be modest ($d < 0.8$) (Fig. 3a, Supplementary Fig. 4 and Supplementary Table 8).

While the 20-min Tris–EDTA HIER, 4 °C overnight stain condition yielded the overall lowest CV for most markers, we also observed some exceptions, such as PAX5, cytokeratin and CD4, where shorter incubations at higher temperatures yielded better performance (Fig. 2b and Supplementary Fig. 3). We further validated this by performing an experiment on a slide retrieved with Tris–EDTA for 20 min, in which three markers (CD8, CD68 and FOXP3) were stained first at 4 °C overnight, followed by staining of CD4 at 25 °C for 1 hour (Extended Data Fig. 1a). This demonstrated that the optimal staining conditions can vary across antibodies and can be optimized through careful testing and by combining different conditions.

Orthogonal validation using cellXpress2 yielded a similar trend in the relative rank scores across the 24 staining conditions compared to that following Mesmer cell segmentation (Fig. 2b, Extended Data Fig. 1b and Supplementary Tables 9 and 10). This consistency indicates that the observed differences in staining outcomes primarily arise from the staining conditions themselves, rather than the choice of segmentation method.

No significant differences in marker intensities were detected between the initial acquisition (performed immediately after the experiment) and a subsequent acquisition conducted on the same slide after 2.5 months of storage in storage buffer (1× DPBS + 5 mM EDTA + 0.5% bovine serum albumin (BSA) + 0.02% NaN_3 + 0.5 M NaCl) at 4 °C (Supplementary Fig. 5). This finding suggests that the bound and crosslinked antibodies remain stable over time, and that the order of slide acquisition does not appear to account for the observed differences in staining performance.

We then compared the ratio of signal intensity inside versus outside the cell segmentation mask for each marker across the 24 conditions (Extended Data Fig. 2). A higher ratio would suggest higher signal intensity of the marker within the cell mask and/or lower signal intensity correspondingly outside the cell mask. The calculated ratios did not match the ranks of the conditions derived from the relative CVs of the markers staining intensities (Fig. 2b, Extended Data Fig. 2 and Supplementary Fig. 6a). A close look at the original images suggested that this metric is sensitive to diffuse signal and imperfect cell segmentation, which could account for the mismatch (Supplementary Fig. 6). The signal intensity ratios across the 24 slides were also inconsistent between data following Mesmer cell segmentation versus that from cellXpress2, further highlighting that this metric is sensitive to the cell segmentation method used. As such, CV, the values of which were stable across different cell segmentation methods, was the primary metric used to rank the 24 initial optimization conditions (Fig. 2b, Extended Data Fig. 1b and Supplementary Tables 9 and 10).

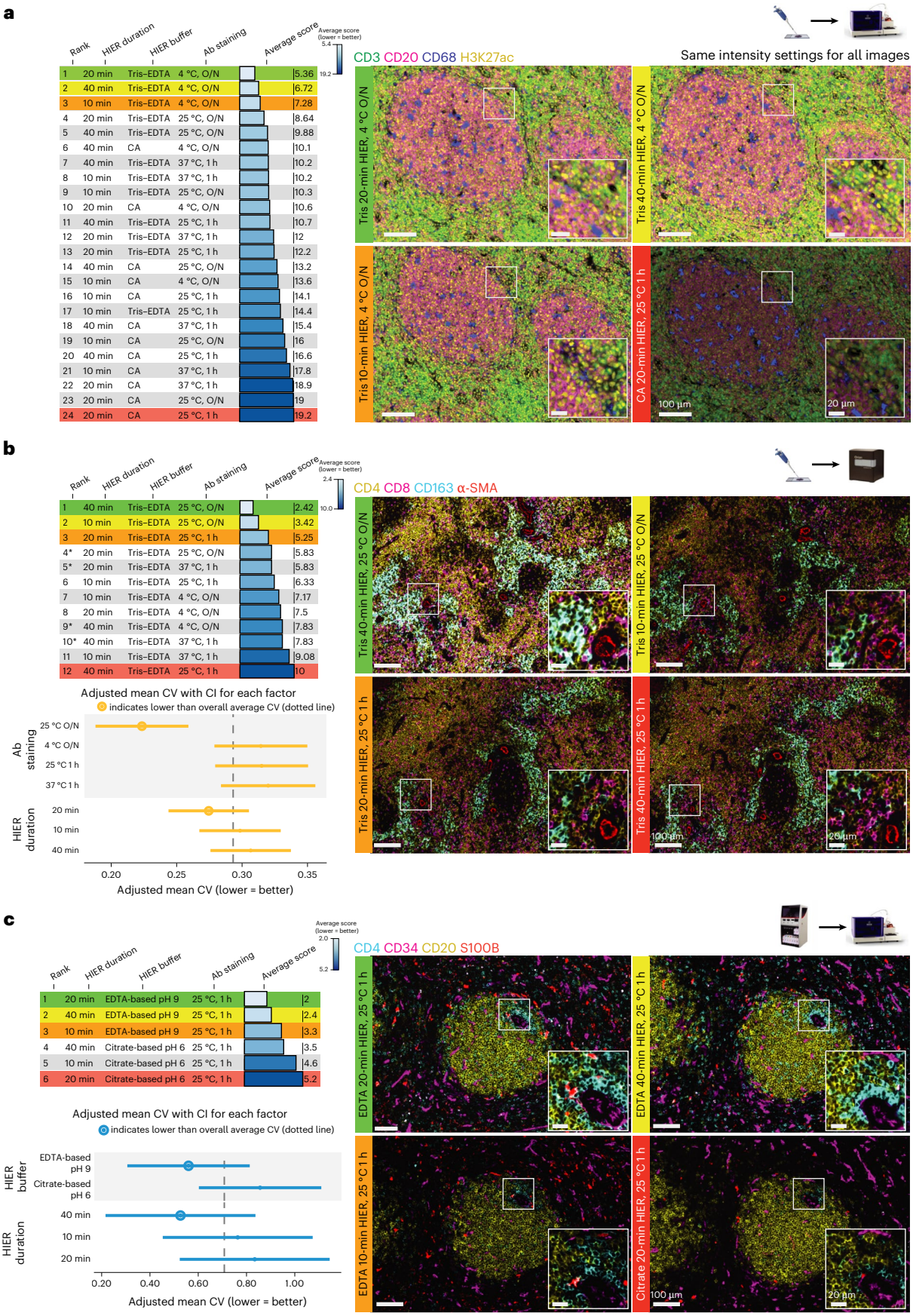
The optimization performed at Stanford and Roche provided valuable insights into optimal protocols for different platforms and the use of automated staining respectively. Among the 12 conditions with Tris–EDTA as the HIER buffer tested on the RareCyte Orion at Stanford, the top performing condition with overall lowest CV of marker staining intensity was the 40-min HIER, 25 °C overnight stain while the bottom condition was 40-min HIER, 25 °C 1-h stain (Fig. 3b and Supplementary Fig. 7a). The factorial linear model fitting showed that 25 °C overnight stain yielded the lowest adjusted mean CV among all factors tested.

Conversely, for the six conditions involving 25 °C 1-h antibody stain tested using the Leica BOND RX automated staining device with imaging on the Akoya PhenoCycler-Fusion system at Roche, the top three conditions were the 20-min HIER, 40-min HIER and 10-min HIER with the EDTA-based, pH 9 buffer, while the bottom condition was 20-min HIER with the citrate-based, pH 6 buffer (Fig. 3c and Supplementary Fig. 7b). Consistent with the results from BIDMC-Harvard, the data from Roche showed that EDTA-based, high pH buffers are superior to their acidic counterparts in yielding lower mean CV values and therefore higher antibody staining intensities (Figs. 2c and 3c).

The CV values across the conditions tested at Stanford and Roche, following both Mesmer and cellXpress2 segmentation, resulted in similar results, indicating that the cell segmentation method used did not significantly affect the relative ranks of the conditions (Supplementary Fig. 8 and Supplementary Tables 9 and 10). The average signal intensity ratios comparing within and outside the cell segmentation mask for the conditions tested at Stanford corresponded to their relative ranks previously determined by the average CV of marker staining intensities, suggesting that conditions with higher staining intensities within the cell mask tend to have lower staining outside the cell mask (Supplementary Fig. 9a). Akin to the BIDMC-Harvard dataset, the signal intensity ratios for the Roche dataset did not match the ranks

Fig. 3 | Ranking of all HIER and antibody staining conditions tested across three sites. a, Twenty-four conditions ranked from top to bottom based on their average CV scores for experiments performed on the Akoya Phenocycler-Fusion with manual staining at BIDMC-Harvard (left). Representative images of the top three (green, yellow, orange) and bottom-ranked (red) conditions. Markers reflect T cells (CD3, green), B cells (CD20, magenta), myeloid cells (CD68, blue) and active chromatin (H3K27ac, yellow) (right). Ab, antibody. **b**, Twelve conditions ranked by their average CV scores for experiments performed on the RareCyte Orion with manual staining at Stanford (top left). Bar plots showing the EMM CV for each HIER duration and antibody staining condition, computed from a factorial linear model of CV as a function of HIER duration and antibody staining condition fitted to the Stanford data (Tris–EDTA only; $n = 143$ per-condition-and-marker CV). Points indicate model-adjusted mean CVs (EMMs); horizontal bars indicate 95% confidence intervals (CIs). The EMMs (95% CI) are: for HIER duration, 10-min 0.298 (0.267–0.330), 20-min 0.275 (0.244–0.305), 40-min 0.307 (0.276–0.337); and for antibody staining, 4 °C overnight 0.314 (0.279–0.350), 25 °C 1-h 0.315 (0.279–0.350), 25 °C overnight 0.223 (0.188–0.259), 37 °C 1-h 0.320 (0.284–0.356). Lower CV indicates higher staining intensity (bottom left). Representative images of the top three (green, yellow, orange) and the bottom-ranked (red) ranked conditions on the RareCyte Orion. Markers show T cells (CD4, yellow; CD8, magenta), macrophages (CD163, cyan) and smooth muscle cells (α-SMA, red) (right). **c**, Six conditions ranked by their average CV scores for experiments performed on the Akoya PhenoCycler-Fusion with automated staining on the Leica BOND RX at Roche (top left). Bar plots showing the EMM CV for each HIER duration and HIER buffer condition, computed from a factorial linear model fitted to the Roche data. Points indicate model-adjusted mean CVs; horizontal bars indicate 95% confidence intervals. Bar plots showing the EMM CV for each HIER duration and HIER buffer condition, computed from a factorial linear model of CV as a function of HIER buffer and HIER duration fitted to the Roche data (25 °C 1-h staining only; $n = 60$ per-condition-and-marker CV). Points indicate model-adjusted mean CVs (EMMs); horizontal bars indicate 95% CIs. The EMMs (95% CI) are: for HIER buffer, EDTA-based pH 9 0.560 (0.306–0.813), citrate-based pH 6 0.856 (0.603–1.110) and for HIER duration, 10-min 0.764 (0.453–1.074), 20-min 0.834 (0.523–1.145), 40-min 0.527 (0.216–0.837). Lower CV indicates higher staining intensity (bottom left). Lower CV indicates higher staining intensity (bottom left). Representative images of the top three (green, yellow, orange) and the bottom-ranked (red) conditions on the PhenoCycler-Fusion. Markers show T cells (CD4, cyan), endothelial cells (CD34, magenta), B cells (CD20, yellow) and dendritic cells (S100B, red) (right). Scale bars, 100 µm (main) and 20 µm (inset). All images on identical scale, crop areas and minimum–maximum intensity settings across slides. Tris, Tris-EDTA; CA, citraconic anhydride.

https://doi.org/10.1038/s41592-026-03317-1



of the conditions derived from the relative CVs of the markers staining intensities (Supplementary Fig. 9b).

Altogether, the initial optimization across the three sites demonstrated the impact of HIER buffer and antibody staining conditions on antibody staining intensities and provided insights into optimal protocols for different platforms and the use of automated stainers.

Single-round annotation of BIDMC-Harvard initial optimization dataset

We next evaluated how staining intensities affect cell type annotation efficiency and robustness. Each slide was subjected to a single round of clustering using the same parameters (annotation markers and *k* parameter) for a meaningful comparison, and the fractions of annotated, mixed and other cell types were compared across the 24 staining conditions performed at BIDMC-Harvard (Fig. 4a and Supplementary Fig. 10). Each cluster from the single round of clustering was visually validated before they were annotated (Supplementary Fig. 11a). Overall, the proportions of annotated cell types corresponded to the expected tissue composition, with B cells being the largest cluster, followed by T cells (Fig. 4a). However, the relative fractions of annotated cell types, mixed and others clusters varied across the different staining conditions and did not seem to match the relative rank of the staining conditions based on the CV values (Fig. 4a and Supplementary Table 11). Marker enrichment analyses further confirmed the robustness of the annotations, where canonical markers (such as CD20 and CD3) were predominantly enriched in their corresponding annotated clusters (B cells and T cells, respectively), whereas sporadic enrichment in nonspecific clusters was occasionally observed, likely due to marker spillover in the tightly packed lymphoid tissues[42] (Fig. 4b).

We next compared the phenotype maps and images of the top-ranked slide (20-min Tris–EDTA HIER, 4 °C overnight stain) and the bottom-ranked slide (20-min citraconic anhydride HIER, 25 °C 1-h stain) based on CV values (Fig. 4c). Despite differences in signal intensity, main cell types, such as B and T cells, were still reliably identified in the bottom-ranked slide following a single clustering run (Fig. 4c). Notably, a distinct T_{reg} cluster ($CD3^+$ $FOXP3^+$) was detected in the citraconic anhydride condition but not in the Tris–EDTA conditions, indicating that moderate differences in staining intensity do not necessarily compromise the identification of major immune cell types (Fig. 4c and Supplementary Fig. 11b).

As an orthogonal validation, we applied Balagan analysis[43–45] to quantitatively compare spatial sampling efficiency across the 24 staining conditions. The Balagan analysis is a statistical framework that determines the minimal number of FOVs required to identify all major cell types in a tissue. Louvain clustering consistently recovered key cell clusters including B cells, epithelial cells, neutrophils, macrophages, $CD4^+$ T cells and $CD8^+$ T cells (Extended Data Fig. 3a,b). The median τ value, which represents the number of regions that must be imaged to recover most cell clusters, from 100 iterations across 10 FOV size ranges (50–500 µm) for each slide was calculated and ranked against the other conditions from high (low τ) to low sampling efficiency (Extended Data Fig. 3c). The median τ rank did not correlate with the relative slide ranking by CV (Extended Data Fig. 3c). Consistent with the PhenoGraph annotation results (Fig. 4), this indicates that spatial sampling efficiency was largely independent of overall staining intensity.

The α parameter derived from the Balagan analysis quantifies the rate at which sampling efficiency improves with increasing FOV size, as a proxy for spatial heterogeneity. A higher α indicates greater spatial organization of distinct cell types, that is, the same number of phenotypes can be recovered either from a few large FOVs or many small ones. The α values were significantly correlated with both the relative CV scores (Fig. 4d and Supplementary Fig. 12a) and the mean fluorescence intensities of the ten phenotypic markers used for clustering (Supplementary Fig. 12b), but not with the τ parameter (Supplementary Fig. 12c) and fractions of mixed and others clusters from manual annotation (Supplementary Fig. 12d). Our findings suggest that staining conditions yielding lower CV values (higher signal intensities) enhance the detection of spatially distinct cell populations, resulting in higher estimated spatial heterogeneity (α), even though overall sampling efficiency (τ) remains unaffected.

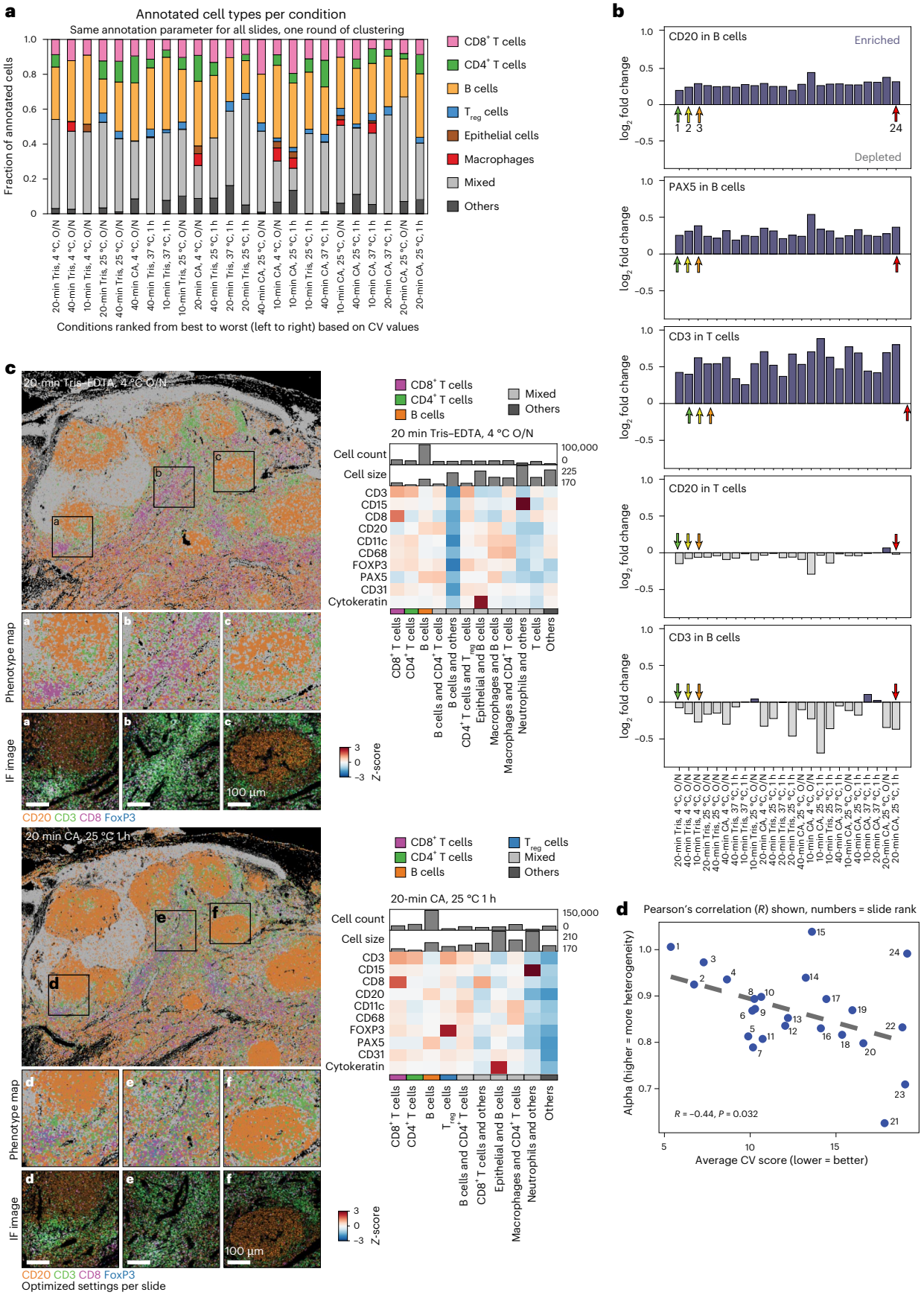
Validation of top three and bottom one conditions

We next expanded our findings to a multi-center level to determine the generalizability and robustness of our initial observations across eight academic and pharmaceutical sites, five imaging platforms, with two different tissues and individualized panels at each site (Fig. 5a and Supplementary Table 12). Since the BIDMC-Harvard dataset involved benchmarking of the full 24 conditions with minimal variations to the standard protocol, the top three and bottom one conditions identified by the relative CV values at BIDMC-Harvard were used for validation at the eight sites (Fig. 5a). Each site used their own FFPE tissues, internal immune-centric antibody panels and internal instruments. All images generated were further assessed by pathologists and nonworking markers were excluded from analysis (Supplementary Table 13).

Validation was performed with the Akoya PhenoCycler-Fusion system at three sites: BIDMC-Harvard, Roche and Novartis (Fig. 5b). For the same tissue type (tonsil), the top three conditions consistently yielded lower CV values and brighter staining intensities visually compared to the bottom condition, with minor differences observed in the relative rank of the top three conditions (Fig. 5b–d and Supplementary Table 14). This trend remained consistent for the other tissue types (diffuse large B-cell lymphoma tissue at BIDMC-Harvard, intestine tissue at Roche and lung cancer tissue at Novartis) (Fig. 5b–d and Supplementary Table 14). The effect sizes of the differences in staining intensities between the top three conditions were also modest (Cohen's *d* within 95% confidence interval <0.8), likely reflecting tissue source and instrumental variations, and suggesting that varying HIER duration within the range of 10–40 minutes is likely to have marginal effects on staining intensities (Supplementary Table 15). Together, the cross-site validation demonstrated that HIER with Tris–EDTA and a 4 °C overnight antibody stain reproducibly yielded higher staining intensities compared to HIER with citraconic anhydride and 25 °C 1-h staining across different markers, tissues and sites on the Akoya PhenoCycler-Fusion platform.

Fig. 4 | Assessment of the impact of protein marker staining intensity on downstream analysis. a, Stacked plots depicting fractions of annotated, mixed and other cell types across the 24 HIER and staining conditions tested at BIDMC-Harvard. All slides went through one round of clustering and annotation with identical parameters. T_{reg} cells, regulatory T cells. **b**, Enrichment of marker expression in annotated cell types. Indigo or gray bar plots show enrichment or lack of enrichment of marker expression in the specified cluster compared to all other annotated clusters (excluding 'Mixed' and 'Others'), respectively. Slides are ordered from left to right based on CV ranking. Arrows indicate the top 3 (1, 2, 3) and bottom 1 (24) conditions. The first three plots show enrichment (CD20/PAX5 in B cells, CD3 in T cells); the last two plots show the lack of enrichment of marker expression (CD20 in T cells, CD3 in B cells). **c**, Phenotype maps of the top and bottom conditions according to their CV rank showing the distribution of the annotated cell types in the tonsil, together with zoomed-in images of the phenotype map of annotated cell types and the corresponding immunofluorescence (IF) image. Heatmaps display average protein intensity per cell cluster, with colors representing *Z*-score scaling across rows (blue, low, red, high). **d**, Correlation between the average CV score (calculated in Fig. 2b) and the alpha parameter, which measures tissue heterogeneity, determined using the Balagan analysis, across the 24 staining conditions ($n = 24$ conditions). Statistical significance was assessed using a two-sided Pearson's correlation test: Pearson's $r = -0.44$ (95% confidence interval −0.72 to −0.04), $t_{22} = -2.30$, $P = 0.032$; the correlation coefficient (r) serves as the effect size. Numbering indicates slide rank (lower CV score means higher intensity).

https://doi.org/10.1038/s41592-026-03317-1



Results from the RareCyte platform (lymph nodes), IMC platform (tonsil and oral squamous cell carcinoma tissues) and MIBI platform (lymph nodes, liver and colon tissues) confirmed 20-min citraconic anhydride HIER, 25 °C 1-h stain as the bottom-ranked condition across most markers (Extended Data Fig. 4). An exception was the endometrial cancer tissue data assessed on the RareCyte at Stanford, where the average scores of conditions 3 and 24 were equivalent (Extended Data Fig. 4a), in which 9 out of 13 markers performed better for condition 24, while 4 markers failed (Extended Data Fig. 4a). Conversely, the markers were observed to be dimmest when stained with condition 3 compared to the other three conditions (Extended Data Fig. 4a). In general, the top two conditions performed well across these platforms and tissue types and the bottom condition remained consistent, highlighting the generalizability of our optimized HIER and staining protocols (Extended Data Fig. 4).

Due to the inability to modify the antibody staining condition on the Lunaphore COMET platform, we were only able to assess different HIER conditions at A*STAR and University of Kentucky (Extended Data Fig. 5). While the relative performance of the four selected conditions was variable across the different tissue types and sites, 20-min HIER with Tris–EDTA generally performed consistently well across the different tissue types at both sites (Extended Data Fig. 5). The 20-min HIER with citraconic anhydride yielded the lowest CV and hence highest staining intensity for 8 out of the 15-marker panel for the colorectal cancer tissue, which suggests that optimal HIER conditions can be tissue-dependent and marker-dependent (Extended Data Fig. 5a). We further evaluated the CV of each marker for all data from the multi-center validation study using a different cell segmentation method, cellXpress2 and observed very similar trends with those from the Mesmer segmentation, further highlighting the robustness of our observations (Supplementary Fig. 13 and Supplementary Tables 16 and 17). Together, our results reflect the consistent high performance of 20-min HIER with a Tris–EDTA buffer, coupled with a 4 °C overnight antibody stain, where applicable, across multiple tissue types and multiplexed imaging platforms.

Assessment of FFPE tissue section storage conditions on staining intensities

In addition to tissue fixation and processing conditions of FFPE tissue blocks, the storage conditions also affect the antigenicity loss of FFPE tissue sections[46]. We thus sought to assess the impact of tissue section storage conditions for spatial proteomics. Adjacent tonsil sections stored at (1) room temperature, bench top for 6 months, (2) room temperature, vacuum chamber for 6 months, (3) −20 °C, vacuum chamber for 6 months and (4) freshly sectioned 2 days before the experiment were stained in parallel with the same protocol (20-min HIER with Tris–EDTA, 4 °C overnight stain) and the same antibody master mix, before imaging on the PhenoCycler-Fusion (Supplementary Fig. 14a). The performance ranking based on the average CV of the markers demonstrated that both the slides stored at −20 °C vacuum chamber for 6 months (3) and freshly sectioned (4) yielded the highest marker intensities, followed by room temperature, vacuum chamber for 6 months (2) and room temperature, bench top for 6 months (1) (Supplementary Fig. 14b). A visual comparison of the four slides and their raw normalized intensities suggest marginal differences in marker performance (Supplementary Fig. 14c).

Discussion

The field of spatial proteomics has seen rapid technology development and widespread applications in recent years, contributing to its recognition as the 'Method of the Year' in 2024[47]. We sought here to perform an extensive benchmarking and quantitative assessment of antigen retrieval and antibody staining conditions of human FFPE tissue sections to help guide the standardization of protocols to improve the quality and reproducibility of spatial proteomics results within and across laboratories and institutions.

Our results demonstrate that antibody staining intensities are strongly influenced by antigen-retrieval buffer and antibody staining conditions, and to a lesser extent by antigen retrieval duration. By calculating the average CVs of marker signal distributions together with a penalty score for markers that failed under specific conditions, we ranked the tested protocols according to their imaging signal performance. The top-ranked protocol was 20 minutes of HIER with Tris–EDTA (pH 9) buffer, coupled with overnight antibody staining at 4 °C. The second-ranked and third-ranked protocols were 40 minutes and 10 minutes of HIER, respectively, both with Tris–EDTA (pH 9) buffer and overnight antibody staining at 4 °C. The bottom-ranked protocol was 20 minutes of HIER with citraconic anhydride (pH 7.4) buffer and 1 hour of antibody staining at room temperature (25 °C). These results were validated across 8 academic and industry laboratories, 11 tissue types and 5 multiplexed imaging platforms. Across these validation datasets, HIER with Tris–EDTA (pH 9) and overnight antibody staining at 4 °C yielded robust staining, supporting these conditions as a generalizable starting point for spatial proteomics. For cyclic imaging platforms, where antibody staining conditions are fixed, other adjustable parameters, such as antigen-retrieval buffer and duration, can be optimized. Validation results from two sites using the COMET platform also identified 20 minutes of HIER with Tris–EDTA (pH 9) as the optimal protocol across different tissue types and antibodies. The optimization of 12 conditions at Stanford on the RareCyte Orion and 6 conditions at Roche on the Akoya PhenoCycler-Fusion with automated staining on the Leica BOND RX further revealed additional platform-specific insights for users.

Akin to IHC, antigen retrieval is a key determinant to the success of a spatial proteomics experiment. The development of buffer-based HIER methods under high-heat was a key breakthrough in the field of IHC on FFPE tissues for antibody detection in controlled and reproducible settings[26,28,48,49]. The initial optimization performed on the

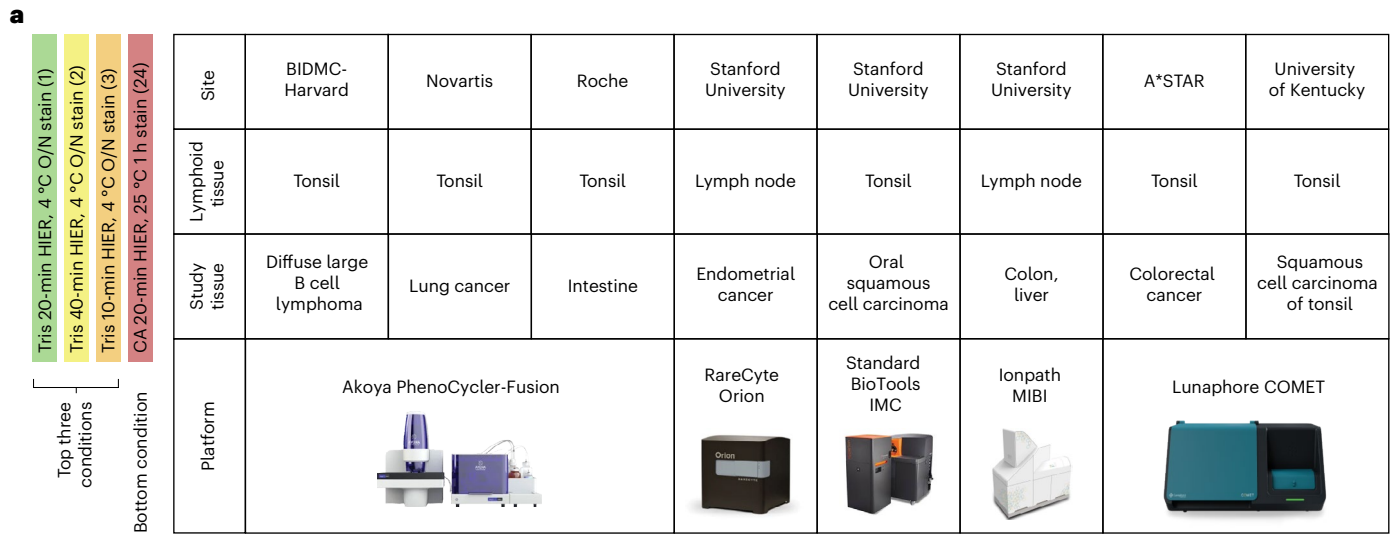
Fig. 5 | Cross validation of spatial proteomics tissue processing and staining conditions. a, Summary of tissue types and platforms involved in the cross-site validation of the top 3 and bottom 1 conditions: (1) 20-min HIER with Tris–EDTA, 4 °C O/N stain; (2) 40-min HIER with Tris–EDTA, 4 °C O/N stain; (3) 10-min HIER with Tris–EDTA, 4 °C O/N stain and (24) 20-min HIER with citraconic anhydride, 25 °C 1-h stain. **b**–**d**, Heatmaps showing the *Z*-scored CVs and average rank scores of conditions tested on the Akoya PhenoCycler-Fusion. **b**, Results for all 24 markers across the 4 tested conditions on lymphoid (tonsil, top left) and 22 markers on study (diffuse large B-cell lymphoma (DLBCL), top right) tissue evaluated at BIDMC-Harvard. Representative images of tonsil (bottom left) and DLBCL (bottom right) treated with the top-ranked and bottom-ranked conditions are displayed with markers for T cells (CD3, green), B cells (CD20, magenta), chromatin (H3K27ac, yellow) and myeloid cells (CD68, blue). **c**, Results for all 12 markers across the 4 tested conditions on lymphoid (tonsil, top left) and study (intestine, top right) tissue evaluated at Roche. Representative images of tonsil and intestine treated with the top-ranked and bottom-ranked conditions are displayed with markers for T cells (CD4, green; CD8, magenta), dendritic cells (S100B, yellow) and B cells (CD20, blue) for tonsil tissue (bottom left), and mesenchymal cells (vimentin, green), endothelial cells (CD34, magenta), epithelial cells (cytokeratin, yellow) and T cells (CD8, blue) for intestine (bottom right). **d**, Results for all nine (tonsil) or five (lung cancer) markers across the four tested conditions on lymphoid (tonsil, top left) and study (lung cancer, top right) tissue evaluated at Novartis. Representative images of tonsil (bottom left) and lung cancer (bottom right) treated with the top-ranked and bottom-ranked conditions are displayed with markers for T cells (CD3, green), myeloid cells (CD15, magenta), epithelial cells (E-cadherin, yellow) and dendritic cells (CD11c, blue). Lower score indicates higher staining intensity. All images from the same tissue type and site are on identical scales, crop areas and minimum–maximum intensity settings across slides. Scale bar, 100 µm.

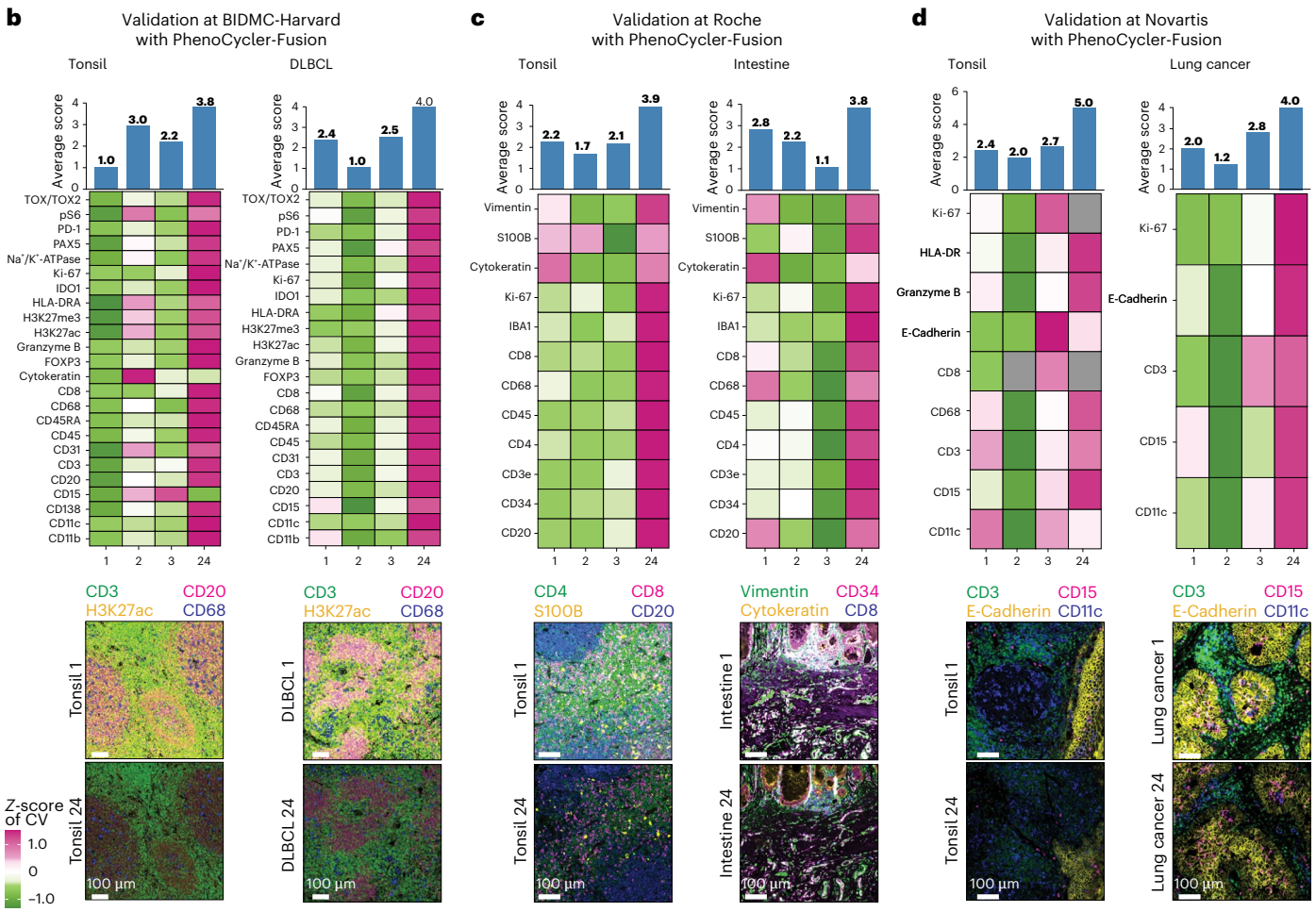
https://doi.org/10.1038/s41592-026-03317-1

PhenoCycler-Fusion platform at both BIDMC-Harvard and Roche consistently demonstrated that HIER using a Tris–EDTA (pH 9) buffer generally yielded higher staining intensities compared to citraconic anhydride (pH 7.4) and citrate buffer (pH 6). Two key factors might account for the superior performance of Tris–EDTA (pH 9): (1) its high pH and (2) the presence of EDTA as a calcium-chelating agent. A previous study that assessed seven buffer solutions across pH 1–10 concluded that for the same buffer composition, a higher pH (pH 8–10) generally yielded higher staining intensities compared to an acidic pH (pH 4–6)[30].

Antigen retrieval with a Tris-HCl buffer at pH 9 also yielded the most resolvable proteins via mass spectrometry, compared to retrieval with Tris-HCl at pH 4 and pH 6[50]. Together with our results, these indicate that HIER at an alkaline pH is more efficient in reversing protein formaldehyde adducts and crosslinks than a lower pH. Another important contributing factor is the presence of EDTA in the buffer, which can improve protein epitope exposure by removing tissue-bound calcium ions[27,51,52].

Relative to the type of HIER buffer, the duration of HIER has a lower effect on staining intensities; tissue sections from the same block that





underwent 10-min, 20-min and 40-min HIER with the same buffer tended to have smaller differences in staining intensities compared to different HIER buffers. It has been previously reported that prolonged HIER might result in the destruction of tissue morphology, likely through 'over' reverse-crosslinking, and thus careful consideration is warranted when planning experiments involving precious samples[26].

Another key determinant to the successful staining of markers in multiplexed imaging is the antibody staining condition. Our results demonstrate that overnight antibody staining at 4 °C resulted in higher signal intensities and lower CVs for markers, while 1-h staining at either 25 °C or 37 °C generally performed poorly. This is consistent with previous observations in flow cytometry, where overnight incubation with antibodies improved the signal-to-noise ratio of markers compared to 1 hour of staining at 4 °C (ref. 40). While increased temperatures translate to higher association rates of antibody to antigen, conversely, higher temperatures will also result in increased dissociation rates, which might explain the inferior performance of slides stained at higher temperatures[53]. Nevertheless, the ideal antibody staining condition can vary with antibody and spatial proteomics platform. For example, for the BIDMC-Harvard data, while CD4 did not stain specifically for overnight antibody staining at 4 °C, specific staining was only observed for 1-h staining at either 25 °C or 37 °C. The initial optimization performed by Stanford on the RareCyte Orion also showed that overnight staining at 25 °C resulted in higher average staining intensities. Apart from antibody staining temperature and duration, antibody concentration is also a key factor to ensuring high signal-to-noise ratio. As such, we recommend titrating antibodies before the actual experiment to determine the optimal concentration and staining protocol for the best signal-to-noise ratio[38].

Studies benchmarking tissue preparation and staining conditions in IHC traditionally rely on semi-quantitative grading assessments[30,31,54]. In this study, we used a quantitative approach, where the relative performance of different HIER and staining conditions was determined and ranked by their average CV of the signal distribution in segmented cells across the markers; a low CV corresponds to a tighter signal distribution with a higher peak. We recognize that the CV value is limited to measuring signal intensities and variance, and does not necessarily translate to better signal-to-noise ratios. For example, a slide where an antibody stains nonspecifically to all cells will have a lower CV and be ranked better than a slide where an antibody stains specifically. To partially mitigate this challenge, slides with markers evaluated by pathologists as not staining specifically were given a penalty score and excluded from the CV ranking process. We also compared the signal intensity ratios inside and outside the cell mask, where signals outside the cell mask were assumed to represent a nonspecific background.

The signal intensities or CVs of the markers did not significantly influence spatial sampling or annotation efficiency from our limited results, suggesting that unsupervised clustering-based cell type annotation is more likely to be influenced by factors such as tissue type, antibody specificity and numbers, factors that were fixed in our study. We did observe a negative correlation between the CVs of the markers and spatial heterogeneity, as defined by the alpha parameter[44]. This suggests that lower CVs or higher staining intensities are more sensitive for elucidating rare cell phenotypes, likely due to higher dynamic range of signals.

We also compared the effects of different slide storage conditions on staining intensities, and observed that slides that were freshly sectioned or stored at −20 °C in vacuum chamber for 6 months yielded the highest overall staining intensities, followed by the slide stored at room temperature in vacuum chamber for 6 months, with the slide stored at room temperature on bench top for 6 months having the overall lowest staining intensity. This is consistent with previous studies that reported enhanced antigenic loss for FFPE slides stored at room temperature and/or in the presence of moisture[46,55,56]. We note that the effects of storage conditions of the FFPE tissues on staining intensities were minimal in our study, potentially due to the use of optimized HIER and antibody staining conditions, along with robust antibody clones and marker selections. It is important to note that others have reported loss of antigenicity of specific markers such as p53, MIB1 and PD-L1 during long-term storage of FFPE tissue sections[46,56–58]. We therefore recommend for slides to be freshly sectioned or stored in vacuum chamber (at −20 °C if possible) to ensure maximal antigen preservation[55,59].

We further demonstrate in this study that pre-aliquoting and lyophilizing antibodies can be performed to maximize the reproducibility of results for large cohort studies, where samples are stained in batches and possibly across different users. In general, lyophilized antibodies can be resuspended at similar or slightly higher concentrations as the optimized fresh antibody staining concentrations without noticeable signal loss.

This study involved an initial large-scale quantitative assessment of 24 conditions consisting of different HIER buffers and antibody staining conditions and durations, followed by validations across eight academic and pharmaceutical institutions with five different imaging platforms, tissue types and antibody panels. We foresee the results, data and best practices presented here will benefit users of various multiplexed imaging platforms in maximizing the success of their spatial proteomics experiments and promote the intra-site and inter-site reproducibility of results.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-026-03317-1.

References

1. de Souza, N., Zhao, S. & Bodenmiller, B. Multiplex protein imaging in tumour biology. *Nat. Rev. Cancer* **24**, 171–191 https://doi.org/10.1038/s41568-023-00657-4 (2024).
2. Mund, A., Brunner, A. D. & Mann, M. Unbiased spatial proteomics with single-cell resolution in tissues. *Mol. Cell* **82**, 2335–2349 https://doi.org/10.1016/j.molcel.2022.05.022 (2022).
3. Bendall, S. C. et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332**, 687–696 https://doi.org/10.1126/science.1198704 (2011).
4. Porwit, A. & Béné, M. C. Multiparameter flow cytometry applications in the diagnosis of mixed phenotype acute leukemia. *Cytom. Part B: Clin. Cytom.* **96**, 183–194 https://doi.org/10.1002/cyto.b.21783 (2019).
5. Giesen, C. et al. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat. Methods* **11**, 417–422 https://doi.org/10.1038/nmeth.2869 (2014).
6. Angelo, M. et al. Multiplexed ion beam imaging of human breast tumors. *Nat. Med.* **20**, 436–442 https://doi.org/10.1038/nm.3488 (2014).
7. Yeo, Y. Y. et al. A hitchhiker's guide to high-dimensional tissue imaging with multiplexed ion beam imaging. *Methods Cell Biol.* **186**, 213–231 https://doi.org/10.1016/bs.mcb.2024.02.018 (2024).
8. Goltsev, Y. et al. Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. *Cell* **174**, 968–981 https://doi.org/10.1016/j.cell.2018.07.010 (2018).
9. Schürch, C. M. et al. Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. *Cell* **183**, 838 https://doi.org/10.1016/j.cell.2020.10.021 (2020).
10. Schubert, W. et al. Analyzing proteome topology and function by automated multidimensional fluorescence microscopy. *Nat. Biotechnol.* **24**, 1270–1278 https://doi.org/10.1038/nbt1250 (2006).

11. Rivest, F. et al. Fully automated sequential immunofluorescence (seqIF) for hyperplex spatial proteomics. *Sci. Rep.* **13**, 16994 https://doi.org/10.1038/s41598-023-43435-w (2023).
12. Lin, J. R., Fallahi-Sichani, M., Chen, J. Y. & Sorger, P. K. Cyclic immunofluorescence (CycIF), a highly multiplexed method for single-cell imaging. *Curr. Protoc. Chem. Biol.* **8**, 251–264 https://doi.org/10.1002/cpch.14 (2016).
13. Saka, S. K. et al. Immuno-SABER enables highly multiplexed and amplified protein imaging in tissues. *Nat. Biotechnol.* **37**, 1080–1090 https://doi.org/10.1038/s41587-019-0207-y (2019).
14. Blow, N. Tissue issues. *Nature* **448**, 959–960 https://doi.org/10.1038/448959a (2007).
15. Phillips, D. et al. Immune cell topography predicts response to PD-1 blockade in cutaneous t cell lymphoma. *Nat. Commun.* **12**, 6726 https://doi.org/10.1038/s41467-021-26974-6 (2021).
16. Hickey, J. W. et al. Organization of the human intestine at single-cell resolution. *Nature* **619**, 572–584 https://doi.org/10.1038/s41586-023-05915-x (2023).
17. Wang, X. Q. et al. Spatial predictors of immunotherapy response in triple-negative breast cancer. *Nature* **621**, 868–876 https://doi.org/10.1038/s41586-023-06498-3 (2023).
18. Greenbaum, S. et al. A spatially resolved timeline of the human maternal-fetal interface. *Nature* **619**, 595–605 https://doi.org/10.1038/s41586-023-06298-9 (2023).
19. Lin, J. R. et al. Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes. *eLife* **7**, e31657 https://doi.org/10.7554/elife.31657 (2018).
20. Kinkhabwala, A. et al. MACSima imaging cyclic staining (MICS) technology reveals combinatorial target pairs for CAR T cell treatment of solid tumors. *Sci. Rep.* **12**, 1911 https://doi.org/10.1038/s41598-022-05841-4 (2022).
21. Black, S. et al. CODEX multiplexed tissue imaging with DNA-conjugated antibodies. *Nat. Protoc.* **16**, 3802–3835 https://doi.org/10.1038/s41596-021-00556-8 (2021).
22. Jiang, S. et al. Combined protein and nucleic acid imaging reveals virus-dependent B cell and macrophage immunosuppression of tissue microenvironments. *Immunity* **55**, 1118–1134 https://doi.org/10.1016/j.immuni.2022.03.020 (2022).
23. Liu, C. C. et al. Reproducible, high-dimensional imaging in archival human tissue by multiplexed ion beam imaging by time-of-flight (MIBI-TOF). *Lab. Investig.* **102**, 762–770 https://doi.org/10.1038/s41374-022-00778-8 (2022).
24. Danenberg, E. et al. Breast tumor microenvironment structures are associated with genomic features and clinical outcome. *Nat. Genet.* **54**, 660–669 https://doi.org/10.1038/s41588-022-01041-y (2022).
25. Bogen, S. A., Vani, K. & Sompuram, S. R. Molecular mechanisms of antigen retrieval: antigen retrieval reverses steric interference caused by formalin-induced cross-links. *Biotech. Histochem.* **84**, 207–215 https://doi.org/10.3109/10520290903039078 (2009).
26. Krenacs, L., Krenacs, T., Stelkovics, E. & Raffeld, M. Heat-induced antigen retrieval for immunohistochemical reactions in routinely processed paraffin sections. *Methods Mol. Biol.* **588**, 103–119 https://doi.org/10.1007/978-1-59745-324-0_14 (2009).
27. Pileri, S. A. et al. Antigen retrieval techniques in immuno-histochemistry: comparison of different methods. *J. Pathol.* **183**, 116–123 (1997).
28. Shi, S. R., Key, M. E. & Kalra, K. L. Antigen retrieval in formalin-fixed, paraffin-embedded tissues: an enhancement method for immunohistochemical staining based on microwave oven heating of tissue sections. *J. Histochem. Cytochem.* **39**, 741–748 https://doi.org/10.1177/39.6.1709656 (1991).
29. Evers, P. & Uylings, H. B. Microwave-stimulated antigen retrieval is pH and temperature dependent. *J. Histochem. Cytochem.* **42**, 1555–1563 https://doi.org/10.1177/42.12.7983356 (1994).
30. Shi, S. R., Imam, S. A., Young, L., Cote, R. J. & Taylor, C. R. Antigen retrieval immunohistochemistry under the influence of pH using monoclonal antibodies. *J. Histochem. Cytochem.* **43**, 193–201 https://doi.org/10.1177/43.2.7822775 (1995).
31. Namimatsu, S., Ghazizadeh, M. & Sugisaki, Y. Reversing the effects of formalin fixation with citraconic anhydride and heat: a universal antigen retrieval method. *J. Histochem. Cytochem.* **53**, 3–11 https://doi.org/10.1177/002215540505300102 (2004).
32. Leong, A. S. Y. & Haffajee, Z. Citraconic anhydride. *Pathology* **42**, 77–81 https://doi.org/10.3109/00313020903434439 (2010).
33. Keren, L. et al. A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging. *Cell* **174**, 1373–1387 https://doi.org/10.1016/j.cell.2018.08.039 (2018).
34. Keren, L. et al. MIBI-TOF: a multiplexed imaging platform relates cellular phenotypes and tissue structure. *Sci. Adv.* **5**, eaax5851 https://doi.org/10.1126/sciadv.aax5851 (2019).
35. Shaban, M. et al. MAPS: pathologist-level cell type annotation from tissue images through machine learning. *Nat. Commun.* **15**, 28 https://doi.org/10.1038/s41467-023-44188-w (2024).
36. Yeo, Y. Y. et al. Epstein–Barr virus orchestrates spatial reorganization and immunomodulation in the classic Hodgkin lymphoma tumor microenvironment. *Cell Rep. Med.* **7**, 102722 https://doi.org/10.1016/j.xcrm.2026.102722 (2026).
37. Taylor, C. R. Immunohistochemistry in surgical pathology: principles and practice. *Methods Mol. Biol.* **1180**, 81–109 https://doi.org/10.1007/978-1-4939-1050-2_5 (2014).
38. Magaki, S., Hojat, S. A., Wei, B., So, A. & Yong, W. H. An introduction to the performance of immunohistochemistry. *Methods Mol. Biol.* **1897**, 289–298 https://doi.org/10.1007/978-1-4939-8935-5_25 (2018).
39. Im, K., Mareninov, S., Diaz, M. F. P. & Yong, W. H. An introduction to performing immunofluorescence staining. *Methods Mol. Biol.* **1897**, 299–311 https://doi.org/10.1007/978-1-4939-8935-5_26 (2018).
40. Whyte, C. E., Tumes, D. J., Liston, A. & Burton, O. T. Do more with less: improving high parameter cytometry through overnight staining. *Curr. Protoc.* **2**, e589 https://doi.org/10.1002/cpz1.589 (2022).
41. Laksameethanasan, D., Tan, R. Z., Toh, G. W. L. & Loo, L. H. cellXpress: a fast and user-friendly software platform for profiling cellular phenotypes. *BMC Bioinforma.* **14**, S4 https://doi.org/10.1186/1471-2105-14-s16-s4 (2013).
42. Bai, Y. et al. Adjacent cell marker lateral spillover compensation and reinforcement for multiplexed images. *Front. Immunol.* **12**, 652631 https://doi.org/10.3389/fimmu.2021.652631 (2021).
43. Bost, P. et al. Statistical modeling and analysis of cell counts from multiplexed imaging data. *Cell Syst.* **16**, 101296 https://doi.org/10.1016/j.cels.2025.101296 (2025).
44. Bost, P., Schulz, D., Engler, S., Wasserfall, C. & Bodenmiller, B. Optimizing multiplexed imaging experimental design through tissue spatial segregation estimation. *Nat. Methods* **20**, 418–423 https://doi.org/10.1038/s41592-022-01692-z (2023).
45. PierreBSC. Balagan: a collection of miscellaneous code. *GitHub* https://github.com/PierreBSC/Balagan (2025).
46. Xie, R. et al. Factors influencing the degradation of archival formalin-fixed paraffin-embedded tissue sections. *J. Histochem. Cytochem.* **59**, 356–365 https://doi.org/10.1369/0022155411398488 (2011).

47. Method of the year 2024: spatial proteomics. *Nat. Methods* **21**, 2195–2196 https://doi.org/10.1038/s41592-024-02565-3 (2024).
48. Cattoretti, G. et al. Antigen unmasking on formalin-fixed, paraffin-embedded tissue sections. *J. Pathol.* **171**, 83–98 https://doi.org/10.1002/path.1711710205 (1993).
49. Norton, A. J., Jordan, S. & Yeomans, P. Brief, high-temperature heat denaturation (pressure cooking): a simple and effective method of antigen retrieval for routinely processed tissues. *J. Pathol.* **173**, 371–379 https://doi.org/10.1002/path.1711730413 (1994).
50. Rahimi, F., Shepherd, C. E., Halliday, G. M., Geczy, C. L. & Raftery, M. J. Antigen-epitope retrieval to facilitate proteomic analysis of formalin-fixed archival brain tissue. *Anal. Chem.* **78**, 7216–7221 https://doi.org/10.1021/ac060294s (2006).
51. Morgan, J. M., Navabi, H. & Jasani, B. Role of calcium chelation in high-temperature antigen retrieval at different pH values. *J. Pathol.* **182** 233–237 (1997).
52. Kim, Y. S., Ju Lee, S., Kim, I., Kim, G. M. & Paik, S. R. The use of alkaline EDTA solution improves heat-induced epitope retrieval for immunohistochemical localization of MIB-1 antigen. *Acta Histochem. Cytochem.* **32** 287–294 https://doi.org/10.1267/ahc.32.287 (1999).
53. Johnstone, R. W., Andrew, S. M., Hogarth, M. P., Pietersz, G. A. & McKenzie, I. F. C. The effect of temperature on the binding kinetics and equilibrium constants of monoclonal antibodies to cell surface antigens. *Mol. Immunol.* **27**, 327–333 https://doi.org/10.1016/0161-5890(90)90045-2 (1990).
54. Shi, S. R., Cote, R. J., Young, L., Ashraf Imam, S. & Taylor, C. R. Use of pH 9.5 tris-HCl buffer containing 5% urea for antigen retrieval immunohistochemistry. *Biotech. Histochem.* **71**, 190–196 https://doi.org/10.3109/10520299609117158 (1996).
55. Forse, C. L., Pinnaduwage, D., Bull, S. B., Mulligan, A. M. & Andrulis, I. L. Fresh cut versus stored cut paraffin-embedded tissue: effect on immunohistochemical staining for common breast cancer markers. *Appl. Immunohistochem. Mol. Morphol.* **27**, 231–237 https://doi.org/10.1097/pai.0000000000000579 (2017).
56. Jacobs, T. W., Prioleau, J. E., Stillman, I. E. & Schnitt, S. J. Loss of tumor marker-immunostaining intensity on stored paraffin slides of breast cancer. *J. Natl Cancer Inst.* **88**, 1054–1059 https://doi.org/10.1093/jnci/88.15.1054 (1996).
57. Wester, K. et al. Paraffin section storage and immunohistochemistry. *Appl. Immunohistochem. Mol. Morphol.* **8**, 61–70 https://doi.org/10.1097/00022744-200003000-00010 (2000).
58. Haragan, A. et al. Accelerated instability testing reveals quantitative mass spectrometry overcomes specimen storage limitations associated with PD-L1 immunohistochemistry. *Lab. Investig.* **100**, 874–886 https://doi.org/10.1038/s41374-019-0366-y (2020).
59. Baena-Del Valle, J. A. et al. Rapid loss of RNA detection by in situ hybridization in stored tissue blocks and preservation by cold storage of unstained slides. *Am. J. Clin. Pathol.* **148**, 398–415 https://doi.org/10.1093/ajcp/aqx094 (2017).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2026

Jia Le Lee[1,19], Johanna Schaffenrath[1,2,19], Cankun Wang[3,4,19], Lollija Gladiseva[2,5], Jia-Ying Joey Lee[6], Wenrui Wu[1], Yuzhou Chang[1,3,4], Shaohong Feng[3,4], Dingani Nkosi[7], José A. Galván[2], Jahanbanoo Shahryari[8], Jason Yeung[1], Huaying Qiu[1], Yao Yu Yeo[1,9], Lindsay Parmelee[1], Chi Ngai Chan[1], Precious Cramer[1], Hendrik Michel[1,9], Pierre Stephan[7], Yang Wang[1], Li Yen Chong[10], Stephen Capodilupo[11], Jean Loebelenz[11], Christine Camacho Fullaway[8], Candace C. Liu[8], Joe Yeong[10], Jakob Einhaus[12], Dorien Feyaerts[12], Dana Napier[13], Giorgio Basile[14], Nadja Kopp[14], Wendy Fantl[14], Karen Wang[14], Brice Gaudillière[12], Derek Allison[13,15], Michael Angelo[8], Sean C. Bendall[8], Lit-Hsin Loo[6], Jeremy Fisher[11], Scott Rodig[7], Brooke Howitt[8], Qin Ma[3,4], Kerstin Hahn[2] & Sizun Jiang[1,7,9,16,17,18] ✉

[1]Center for Virology and Vaccine Research, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA. [2]Roche Pharma Research and Early Development, Roche Innovation Center Basel, Basel, Switzerland. [3]Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH, USA. [4]Pelotonia Institute for Immuno-Oncology, The James Comprehensive Cancer Center, The Ohio State University, Columbus, OH, USA. [5]Institute of Animal Pathology, Vetsuisse Faculty, University of Bern, Bern, Switzerland. [6]Bioinformatics Institute, Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [7]Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [8]Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA. [9]Program in Virology, Division of Medical Sciences, Harvard Medical School, Boston, MA, USA. [10]Institute of Molecular and Cell Biology, Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [11]Cell Signaling Technology, Danvers, MA, USA. [12]Department of Anesthesiology, Perioperative and Pain Medicine, Stanford University, Stanford, CA, USA. [13]Markey Cancer Center, University of Kentucky, Lexington, KY, USA. [14]Novartis Biomedical Research, Cambridge, MA, USA. [15]Department of Pathology and Laboratory Medicine, University of Kentucky, Lexington, KY, USA. [16]Department of Pathology, Dana Farber Cancer Institute, Boston, MA, USA. [17]Department of Computational Pathology, Brigham and Women's Hospital, Boston, MA, USA. [18]Broad Institute of Harvard and MIT, Cambridge, MA, USA. [19]These authors contributed equally: Jia Le Lee, Johanna Schaffenrath, Cankun Wang. ✉e-mail: sjiang3@bidmc.harvard.edu

Methods

Antibody cocktail preparation

The commercial grade, lot-controlled antibody cocktail for the 24 slides for the initial optimization and 4 slides for validation at BIDMC-Harvard was prepared in a single batch, aliquoted into 0.2-ml PCR tubes, and stored lyophilized. The antibody panel and staining concentrations are found in Supplementary Table 5. Before lyophilization, all conjugated antibodies were spun at 12,500*g* for 15 min to minimize aggregates, then pipetted into 100 µl of antibody diluent (5% donkey serum, 0.05% NaN_3, ultrapure water (Invitrogen, cat. no. 10977015), 1× Tris-buffered saline (TBS) wash buffer with Tween 20 (Sigma, cat. no. 935B-09)). The mix was transferred to a prewet 50-kDa centrifugal filter column (Amicon, cat. no. UFC505096) and centrifuged at 12,500*g* for 8 min. The filter containing the antibodies was inverted into a new collection tube and centrifuged at 12,500*g* for 1 min. Then 350 µl of FFPE block (0.05 mg ml^{-1} mouse IgG, 0.05 mg ml^{-1} rat IgG, 0.5 mg ml^{-1} salmon sperm DNA, Oligo master mix (35 nM each of 57 different custom oligos (IDT)), CODEX S2 Buffer (250 mM NaCl, 39 mM NaH_2PO_4, 61 mM Na_2HPO_4, 5 mM EDTA, 0.5% BSA, 0.02% NaN_3 in 1× TBS)) was added to the antibody mix and the mixture was diluted with antibody diluent to a final volume of 1,400 µl. It was then passed through a prewet 0.1-µm centrifugal filter column (Merck Millipore, cat. no. UFC30VV00) at 12,500*g* for 2 min in 3 batches. The antibody cocktail was split into 14 aliquots of 100 µl each and lyophilized (below).

Lyophilization and reconstitution of antibody cocktail

The cocktail was lyophilized using a Dry Freezer System (Labconco FreeZone 2.5L −84 °C). Once the instrument was cooled to below −40 °C, lid-open aliquots in 0.2-ml PCR tubes (stabilized with paper tissue) were placed in the chamber and the vacuum set to 0.006 mbar. Lyophilization was complete after 1 h and 20 min, after which tubes were carefully closed and stored at 4 °C until use. On the experiment day, each tube was reconstituted in 100 µl of 0.25× TBS (sufficient for 2 slides at 50 µl each) by spinning down, resuspending and incubating for 10 min on ice with gentle resuspension at the 0-min, 5-min and 10-min marks before staining.

Antigen retrieval and CODEX staining protocol at BIDMC-Harvard

The 24 HIER and antibody staining conditions tested are listed in Supplementary Table 1; staining was performed in three batches of eight slides (slides 1–8, 9–16, 17–24). FFPE human tonsil was sectioned at 5 µm onto glass slides and stored under vacuum (Bel-Art, SP Scienceware). Slides were baked for 1 h at 70 °C (VWR, 10055-006), rinsed twice in neat xylene (5 min each) and deparaffinized on a linear stainer (Leica, ST4020): 3× neat xylene, 2× 100% EtOH, 2× 95% EtOH, 1× 80% EtOH, 1× 70% EtOH, 3× ddH_2O (180 s each, constant dipping), then held in ddH_2O. Meanwhile the pretreatment (PT) module (Thermo Fisher, cat. no. A80400012) with the antigen-retrieval chambers was preheated to 75 °C. Tris–EDTA pH 9.0 (Dako, cat. no. S2367) was diluted 10× in ultrapure water; citraconic anhydride pH 7.4 (Sigma, cat. no. 125318) was prepared as in ref. 32 (40.1 µl in ~50 ml ddH_2O, titrated to pH 7.4 with 1 M NaOH, brought to 100 ml). After deparaffinization, slides were rapidly transferred into their preheated HIER buffer (Supplementary Table 1). After completion of HIER, the PT module was cooled in a controlled manner to 65 °C, then chambers were left on the benchtop for 15 min until cooled to room temperature. Slides were rinsed twice in ddH_2O, ringed with a hydrophobic barrier (Vector Laboratories, cat. no. H-4000), washed twice in 1× TBS-T (2 min each), then blocked with 200 µl of blocking buffer while photobleached under a custom LED array for 1 h.

Blocking buffer comprised 150 µl of BBDG block (5% donkey serum, 0.1% Triton X-100, 0.05% NaN_3 in 1× TBS wash buffer with Tween 20 (Cell Marque, cat. no. 935B-09)), 20 µl of oligo block (57 custom oligos, IDT; 0.5 µmol l^{-1} each final), 10 µl of sheared salmon sperm DNA (ThermoFisher, cat. no. AM9680), 10 µl of mouse IgG (Sigma, cat. no. I5381) and 10 µl of rat IgG (Sigma, cat. no. I4131). The humidity chamber was a water-filled and wet-towel-filled tip box on an ice-block cool box (Corning, cat. no. 432021), flanked by two happy lights (Best Buy, cat. no. 6460231) with an overhead LED grow light (Amazon, cat. no. B07C68N7PC).

After blocking, slides were stained with 50 µl of antibody mix under the test conditions (Supplementary Table 1). After antibody staining, the slides were washed twice in CODEX S2 buffer (2 min each) and fixed twice in fresh 1.6% PFA (16% PFA (EMS, cat. no. 15710) diluted 1:10 in storage buffer: 1× DPBS, 5 mM EDTA, 0.5% BSA, 0.02% NaN_3, 0.5 M NaCl) for 10 min each. After 3 1× TBS rinses, tissues were fixed in ice-cold 100% methanol (5 min on ice; barrier scraped off at the 3-min mark), rinsed twice in 1× TBS, then fixed twice (10 min, room temperature) in CODEX final fixative (20 µl of BS3 (Thermo, cat. no. 21580) in 1,000 µl of 1× TBS). Slides were washed twice and stored in 1× TBS at 4 °C for imaging; slides not imaged on the same day were held in storage buffer (stable ~2 weeks at 4 °C). Within each of the 3 optimization batches, all 1-h-staining slides were fixed and imaged first.

Comparison of fresh, lyophilized and automated antibody cocktail for pilot data

The master mix was split into 3 equal tubes; 2 were lyophilized and 1 kept fresh, all were stored at 4 °C for 24 h before the lyophilized tubes were reconstituted in 0.25× TBS. For the fresh-versus-lyophilized comparison, two tonsil tissues on one slide were stained with fresh or lyophilized mix under otherwise identical manual conditions. An adjacent slide, stained with the second lyophilized cocktail, tested an automated Opentrons OT-2 protocol in which all postretrieval washes and staining, except blocking and methanol fixation, were performed on the robot. Subsequent imaging was performed as above.

CODEX imaging on the Akoya PhenoCycler-Fusion

A black flat bottom 96-well plate (Corning, cat. no. 07-200-762) was used for the reporter plate, with each well corresponding to an imaging cycle and 2 additional wells for the first and last blank cycles, respectively. Each well was filled with 245 µl of plate master mix, containing 1:300 Hoechst 33342 (Thermo, cat. no. H3570) and 2.5 µl each of two fluorescent oligonucleotides on the Cy3 and Cy5 channels (GenScript, stock concentration 100 nM). The wells for the blank cycles were filled with 250 µl of plate master mix without fluorescent oligonucleotides. The plate was then sealed with aluminum film (Thermo, cat. no. AB-0626). All reporter plates for an experiment were prepared together for consistency, and they were stored at 4 °C until use.

Right before imaging, flow cells were assembled onto slides using the Flow Cell Assembly Device (Akoya Biosciences) as per the manufacturer's protocol. Slides were acquired in pairs on the PhenoCycler-Fusion (Akoya Biosciences).

Additional optimization of conditions on the RareCyte Orion and CODEX + Leica BOND RX automated staining

In parallel with the 24 BIDMC-Harvard conditions, 12 conditions were tested on the Orion (RareCyte) and 6 on the PhenoCycler-Fusion (Akoya Biosciences) (Supplementary Table 1).

For the Orion (RareCyte), FFPE tissues were baked at 60 °C for 120 min and deparaffinized and rehydrated (2× xylene 4 min; 2× 100% EtOH 1 min; 95% then 70% EtOH 1 min each; fresh 1× PBS 1 min), then washed twice in 1× PBS. Antigen retrieval was performed in preheated buffer in a decloaking chamber for the stipulated time. After which, slides were cooled for 5 min, then washed in 1× PBS. Autofluorescence was quenched in AF Quench Buffer (4.5% H_2O_2/24 mM NaOH in PBS) under a white-light box for 60 min, then transferred to an ultraviolet light transilluminator for 30 min. After quenching, the slides were washed and blocked with Image-iT FX Signal Enhancer for 15 min. After two surfactant wash buffer sprays, the slides were incubated with

antibodies (Supplementary Table 12) on a humidity tray. Slides were coverslipped, cured overnight and cleaned before scanning.

For the CODEX with Leica BOND RX automated staining, the slides were baked at 60 °C for 30 min, and deparaffinized and rehydrated on the BOND RX (30 s in 150 µl of Dewax (AR9222) at 72 °C, then 2× Dewax; 3× 100% EtOH; 3× with 1× wash buffer (AR9590)). Antigen retrieval was performed using the BOND Epitope Retrieval Solution 1 (citrate-based pH 6; Leica, AR9961) or Solution 2 (EDTA-based pH 9; Leica, AR9640) at 95 °C for 10 min, 20 min or 40 min, followed by a flush and 4 washes with 1× wash buffer. Slides underwent three 2-min washes in hydration buffer (Akoya, P/N 7000017), two ddH_2O washes (2 min) and two staining-buffer washes (2 min). Antibody cocktail (150 µl; Akoya and in-house conjugates, Supplementary Table 12) was applied manually under the stipulated conditions, then washed 3× for 2 min in staining buffer. Poststaining followed Akoya guidelines: 1.6% PFA (10 min, room temperature), 3× with 1× PBS; ice-cold methanol (5 min, 4 °C), 3× with 1× PBS; Final Fixative (20 µl PhenoCycler Fixative Reagent in 1,000 µl 1× PBS, 20 min), 3× with 1× PBS.

Qualitative assessment of tissue morphology and staining efficacy

Final QPTIFF and OME-TIFF files were opened in QuPath 0.4.4 and two representative ROIs, conserved across same-site slides, were selected (inclusion and exclusion criteria in Supplementary Fig. 1b; FOV coordinates in Supplementary Table 18). Each marker was assessed for correct subcellular localization (nuclear, cytoplasmic, membranous), anatomical location and expected frequency, supported by trained pathologists (D.N., B.H., K.H.) at BIDMC-Harvard and in-house pathologists at each validation site. Markers failing across all conditions were excluded (Supplementary Fig. 1c), and consistent intensity settings were applied across slides for figure generation.

Assessment of antibody performance

Three trained pathologists (D.N., B.H., K.H.) independently assessed specificity, referencing IHC from the Human Protein Atlas (https://www.proteinatlas.org/), and exploiting multiplex context (for example CD3/CD4 colocalization, CD4/CD8 mutual exclusion). Each marker in one FOV per section was scored: a score of 2 meant that the cell and tissue pattern of immunoreactivity fully agreed with known patterns of protein expression, a score of 1 meant that the cell and tissue pattern of immunoreactivity overall agreed with known patterns of protein expression, but background staining was observed, and score of 0 meant that the cell and tissue pattern of immunoreactivity did not agree with known patterns of protein expression and were consistent with nonspecific staining patterns. A result of 'NP' denoted assessment of positive staining was not possible and further information was added as a comment.

Scores agreed by two or more pathologists were final, and markers with a mean score <1 were excluded (Supplementary Fig. 1c).

Quantitative assessment of mean fluorescent intensities of markers and CVs following cell segmentation and feature extraction

Cell segmentation and feature extraction were performed on representative ROIs using Mesmer (parameters: image_mpp = 0.50, maxima_threshold = 0.075, interior_threshold = 0.2)[60] and cellXpress version 2.4.0 (https://cellXpress.org, default segmentation parameters except gapless segmentation = true)[41]. At least two ROIs were selected per slide where tissue was consistently represented across all conditions (Supplementary Fig. 1b and Supplementary Table 18). Hoechst signal was used as a nuclear marker and membrane markers assessed to be working were used as membrane signal for cell segmentation (Supplementary Fig. 1b). All segmentation markers used were consistent across all conditions. In Fig. 1, CD3, CD4, CD8, CD11b, CD20, CD68, CD31, CD206, CD138, Na/K-ATPase, CD45RA and cytokeratin were used as membrane signals, while in Figs. 2 and 4, CD45, HLA-DR, cytokeratin and CD11b were used (Supplementary Table 18). The extracted single-cell features were then exported and data from the two ROIs were pooled.

Per-cell signals were processed in R 4.3.2. Cells with Hoechst = 0 or below $Q1 - 1.5$ IQR (Q1, first quartile; IQR = $Q3 - Q1$) were removed; signals were normalized to the median Hoechst of all same-experiment slides, arcsinh-transformed (cofactor 0.001), percentile-normalized (floor 0.001, ceiling 0.999), then scaled to [0, 1]. Per-marker distributions (geom_density) and Z-score heatmaps of mean expression and of CV (CV = σ/μ, where σ refers to the standard deviation and µ refers to the mean) (geom_tile) were plotted with ggplot. For CV scoring, conditions were ranked per marker from lowest CV (score 1) to highest (score = number of slides; for example 24 at BIDMC-Harvard); nonworking markers incurred a penalty (10 for 4–6-slide, 30 for 24-slide, 18 for 12-slide experiments). Mean rank scores were compared across all conditions tested at the three optimization sites and eight validation sites.

Differences across conditions per marker were tested by Kruskal–Wallis with Benjamini–Hochberg correction, followed by Dunn's post hoc (Benjamini–Hochberg-corrected) where significant ($P < 0.05$). Effect sizes used Cohen's d:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s_{\text{pooled}}}, \quad s_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

where $\bar{X}_i$ refers to the mean, s_i refers to the standard deviation and n_i refers to the sample size of group i.

The effect size was deemed as small if $0 \le d \le 0.2$, medium if $0.2 < d < 0.8$ and large if $d \ge 0.8$.

The 95% confidence interval for Cohen's d is determined by the following formula:

$$\left(d - 1.96 \times \sigma(d),\ d + 1.96 \times \sigma(d) \right)$$

and the variance of d is

$$\sigma(d) = \sqrt{\frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}}.$$

The same pipeline was applied to Mesmer and cellXpress2 outputs to confirm differences reflected staining conditions rather than the segmentation method.

Factorial analysis of experimental factors on staining intensity

To quantify each preparation variable's contribution to staining intensity, per-marker CVs from the initial optimization were modeled by linear regression with marker identity as a covariate to account for marker-intrinsic CV. At BIDMC-Harvard, where all 24 conditions were tested, CV was modeled as a function of HIER buffer, HIER duration and antibody staining condition ($n = 592$); at Stanford (Tris–EDTA only) as a function of HIER duration and antibody staining condition ($n = 143$) and at Roche (25 °C 1-h staining only) as a function of HIER buffer and HIER duration ($n = 60$). All factors were categorical with sum-to-zero contrasts. Relative importance was assessed by Type II ANOVA, yielding partial η^2 (residual CV variance uniquely attributable to each factor) with 95% confidence intervals. EMMs of CV per factor level (averaging over other factors and marker identity; 95% confidence intervals from the t-distribution) compared adjusted mean CV (lower shows a higher intensity) across preparation options. These analyses were performed using the car, effectsize and emmeans R packages.

Quantitative comparison of signal-to-noise ratios using line scans

Line scan analyses were performed on the four slides from the pilot experiment to compare the relative signal-to-noise of selected markers.

On an aligned ROI with consistent intensity settings, a line was drawn across two-to-five cells in ImageJ 2.14.0 and copied to all images via the ROI manager; per-pixel intensities from Plot Profile were plotted in R 4.3.2. For the large-scale optimization, the standardized metrics below were used instead.

Quantitative assessment of signal intensity ratio inside versus outside cell mask

Per-marker TIFFs were generated per slide and the FOV segmentation mask applied to compute signal within (signal_in) and outside (signal_out) the mask. Each was normalized to the corresponding 4,6-diamidino-2-phenylindole (DAPI) signal for that slide, and the area-normalized ratio of in-mask to out-of-mask DAPI-normalized signal was computed as:

$$\left(\frac{\text{DAPI}-\text{normalized signal}_{\text{in}}}{\text{area}_{\text{in}}}\right)/\left(\frac{\text{DAPI}-\text{normalized signal}_{\text{out}}}{\text{area}_{\text{out}}}\right)$$

The ratios for each marker were compared on all slides and displayed on a heatmap.

Cell type annotation and comparison of annotation specificity and efficiency

Cell annotation on the aforementioned 2 ROIs per slide was performed on the 24 slides from the initial optimization performed at BIDMC-Harvard. An iterative clustering process using PhenoGraph version 1.5.7 was performed[22,35,36,42,61,62]. The markers used for clustering were CD3, CD15, CD8, CD20, CD11c, CD68, FOXP3, PAX5, CD31 and cytokeratin (phenotypic markers, see Supplementary Table 5 for details). Clustering and annotation were done on each slide independently to minimize the impact on cell phenotyping due to differences in staining intensities. Single-cell marker expressions from each slide were first normalized and filtered as described above and used as input for PhenoGraph clustering. Cluster marker enrichment was visualized as *Z*-score heatmaps for annotation; annotations and masks were exported as GeoJSON and inspected by overlay in QuPath 0.4.4. Mixed-cell-type clusters were labeled 'Mixed' and ambiguous clusters 'Others', and cluster proportions compared across the 24 slides.

Annotation specificity was assessed by bar plots of marker enrichment in specific versus nonspecific clusters ($\log_2 \frac{\text{MFI marker in respective cluster}}{\text{MFI marker in all other clusters}}$); accurate annotation shows lineage-marker enrichment and absence of nonspecific markers. Phenotype maps of aligned ROIs confirmed expected spatial localization.

Uniform manifold approximation and projection (UMAP) plots were generated using Scanpy (Python). Per-slide clustering results were matched back to the corresponding AnnData objects using cell identifiers, and UMAPs were generated both after concatenating all 24 slides and independently for each slide. Marker-expression values were stored before scaling for visualization, after which expression values were scaled, a nearest-neighbor graph was constructed using 30 neighbors and UMAP embeddings were computed with a minimum distance of 0.5. Combined UMAPs were colored by PhenoGraph cluster ID, manual annotation and slide ID to assess slide-level effects, with marker-expression color scales capped at the 99th percentile. Per-slide UMAPs were stitched with marker-expression panels for review.

Sampling analysis to evaluate the effect of HIER and staining conditions on sampling efficiency and tissue heterogeneity

Spatial clustering and sampling analyses were performed on FOV1 of the 20 min HIER, 4 °C overnight staining slide from the pilot experiment and FOV1 of all 24 slides from the initial optimization experiment using the Balagan v1.0.0 R package[45]. Raw data were imported into a SingleCellExperiment (2D, bit_mode 16) and normalized by Pearson-residual count transform (per-channel log-Poisson regression of total intensity on cell size). A *k*-nearest neighbors graph (Euclidean, $k = 15$) with Louvain clustering was built. Sampling analysis (Perform_sampling_analysis) varied sampled FOVs from 1 to 20, counting a cluster as detected at >2 cells, with 20 repeats (N_times = 20) and width/height_FOV_vector = 400.

For the large-scale initial optimization of 24 conditions, an expanded set of sampling parameters was used to more comprehensively evaluate slide performance. Square window sizes used ranged from 50 × 50 µm to 500 × 500 µm in increments of 50 µm (10 sizes in total), and the number of sampled FOVs ranged from 1 to 20. The same ten phenotypic markers used previously for annotation were used for Louvain clustering. Because Balagan's *k*-nearest neighbors-based clustering introduces stochasticity, repeated runs of the same slide produced variation in cluster assignments and tau (τ) values. To obtain stable consensus metrics across slides, the full pipeline was run 100 times for each slide. For each window size, the median tau value (τ) across 100 runs was calculated, representing the minimum number of sampled FOVs required to recover all major cell phenotypes at that scale.

All slides were ranked by their median τ value independently for each of the 10 FOV sizes: the slide with the lowest τ (most sampling-efficient) at that FOV size was assigned a rank of 1, and the slide with the highest τ (least efficient) was assigned a rank of 24. The average rank across all FOV sizes was then used to determine overall slide performance. The alpha (α) parameter, calculated from the relationship between $\log(\tau)$ and log(FOV size), describes how τ scales with FOV size and reflects tissue heterogeneity. Slides with a higher α are considered more heterogeneous and typically require either many small FOVs or a few large FOVs to recover the full set of cell phenotypes.

Data visualization

All plots were generated using ggplot2[63].

Multisite validation of nominated conditions

The top three and bottom one conditions were selected for multisite and multi-platform validation (Supplementary Fig. 1a). Collaborators at A*STAR, University of Kentucky, Novartis, Roche and Stanford performed spatial proteomics on their chosen platforms using these four conditions, each with their own optimized immune-focused panels and tissues (lymphoid and study tissue) (Supplementary Table 12) and the standardized baking–deparaffinization protocol above.

Data were transferred to BIDMC-Harvard for coordinated analysis with the standardized segmentation and feature-extraction pipeline, with orthogonal cellXpress2 segmentation by A*STAR (FOV coordinates and membrane markers in Supplementary Table 18). *Z*-score heatmaps of mean expression and CV per marker across the four conditions from both segmentation methods were compiled to assess reproducibility.

Statistics and reproducibility

No statistical method was used to predetermine sample size, and all experiments in this study were carried out once. Because the study benchmarks tissue preparation and staining conditions rather than biological groups, each condition was represented by a single stained tissue section per site and platform, and reproducibility was instead assessed across the three sites and multiple platforms in the initial optimization and across eight sites in the second-stage validation.

Following extensive quality control, markers of poor performance were excluded from analysis. Antibody specificity was scored independently by three trained pathologists (D.N., B.H., K.H.), with scores agreed by two or more pathologists taken as final; markers that failed to stain specifically (mean score <1) or that failed across all conditions were excluded (Supplementary Fig. 1c). Slides affected by technical errors were excluded according to the criteria in Supplementary Fig. 1b. The investigators were not blinded to condition during acquisition or analysis, as knowledge of the preparation and staining condition was required to run the benchmarking pipeline; conditions were not

randomized for the same reason. Pathologist scoring of antibody performance was performed blinded and independently to mitigate subjective bias.

For each slide, two representative ROIs that were conserved across slides from the same site were selected and pooled, such that each condition comprised cells from two ROIs (Supplementary Table 18). Single-cell data were filtered for outliers, normalized and arcsinh-transformed as described above, and the same pipeline was applied to both Mesmer and cellXpress2 segmentation outputs to confirm that observed differences reflected staining conditions rather than the segmentation method.

Differences in signal intensity across conditions were tested per marker using the Kruskal–Wallis test with Benjamini–Hochberg correction; where significant, pairwise comparisons used Dunn's post hoc test with Benjamini–Hochberg correction. A two-sided $P < 0.05$ was considered statistically significant in all tests. Effect sizes were quantified using Cohen's d with 95% confidence intervals (small, $0 \le d \le 0.2$; medium, $0.2 < d < 0.8$; large, $d \ge 0.8$). Factor contributions to CV were assessed by Type II analysis of variance with partial η^2 and 95% confidence intervals. All analyses were performed in R version 4.3.2 unless otherwise stated.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data from the pilot experiment including the raw images and .csv files from cell segmentation are available via Zenodo at https://doi.org/10.5281/zenodo.11410274 (ref. 64). All raw images for the initial optimization, validation and additional experiments, and annotation input and output files are available via the BioImage Archive at https://doi.org/10.6019/S-BIAD2491. Selected images are available via ImmunoAtlas. All .csv files from cell segmentation are available via Zenodo at https://doi.org/10.5281/zenodo.17843231 (ref. 65).

Code availability

All code used are available via GitHub at https://github.com/SizunJiangLab/Benchmarking_tissue_preparation/.

References

60. Greenwald, N. F. et al. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nat. Biotechnol.* **40**, 555–565 https://doi.org/10.1038/s41587-021-01094-0 (2022).
61. Phillips, D. et al. Highly multiplexed phenotyping of immunoregulatory proteins in the tumor microenvironment by CODEX tissue imaging. *Front. Immunol.* **12**, 687673 https://doi.org/10.3389/fimmu.2021.687673 (2021).
62. Zhu, B. et al. CellLENS enables cross-domain information fusion for enhanced cell population delineation in single-cell spatial omics data. *Nat. Immunol.* **26**, 963–974 https://doi.org/10.1038/s41590-025-02163-1 (2025).
63. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016).
64. Lee, J. L. & Schaffenrath, J. Benchmarking_tissue_preparation_pilotdata. *Zenodo* https://doi.org/10.5281/zenodo.11410274 (2024).
65. Lee, J. L., Schaffenrath, J., Wang, C. & Jiang, S. Benchmarking tissue preparation Stage2. *Zenodo* https://doi.org/10.5281/zenodo.17843231 (2025).

Acknowledgements

This article reflects the views of the authors and should not be construed as representing the views or policies of the institutions that provided funding.

Author contributions

Conceptualization was done by S.J., J.L.L., J.S. and C.W. Experiments were carried out by J.L.L., J.S., J.A.G., J.S., L.Y.C., C.C.F., C.C.L., J.E., D.F., D.N., G.B. and N.K. Analysis was carried out by J.L.L., J.S., C.W., L.G. and J.-Y.J.L. Pathological assessment was done by D.N., B.H. and K.H. Contribution of unique reagents, tools or technical expertise came from Y.C., W.W., S.F., J.Y., H.Q., Y.Y.Y., L.P., C.N.C., P.C., H.M., P.S., Y.W., S.C., J.L., J.Y., W.F., K.W., B.G., D.A., M.A., S.C.B., L.-H.L., J.F., S.R. and Q.M. Writing was done by J.L.L., J.S. and S.J. Supervision and funding were done by S.J.

Funding

J.L.L. is supported by a National Science Scholarship (PhD) from the Agency for Science, Technology and Research, Singapore (grant no. BM/NDR/18/003) and is a Schmidt Science Fellow. J.S. is supported by a Roche Postdoctoral Fellowship. Y.Y.Y. is a recipient of the Albert J. Ryan Fellowship. P.S. is supported by the Lymphoma Research Foundation. D.N. and D.A. are supported by the Biospecimen Procurement & Translational Pathology Shared Resource Facility of the University of Kentucky Markey Cancer Center (grant no. P30CA177558). D.F. is supported by grant no. K99HD115829. Q.M. is supported by the following grant nos.: R01GM152585, P01CA278732 and U54AG075931. S.J. is supported in part by grant nos. NIH DP2AI171139, P01AI177687, R01AI149672, R01GM152585, U24CA224331, R01NS139479, P50CA272390, a Gilead's Research Scholars Program in Hematologic Malignancies, the Bill & Melinda Gates Foundation INV-002704, the Broad Next Generation Award, the Dye Family Foundation and the Bridge Project, a partnership between the Koch Institute for Integrative Cancer Research at MIT and the Dana-Farber/Harvard Cancer Center.

Competing interests

S.J. is a co-founder of Elucidate Bio Inc. and Saterra AI Inc., serves on their Board of Directors and Scientific Advisory Board, and has received research support from Roche and Novartis. S.J. has received speaking honorariums or consulting fees from Cell Signaling Technology, Quanterix and AstraZeneca. S.R. has received research funding from Bristol Myers Squibb, KITE/Gilead and Surface Pharmaceuticals, and is a member of the Scientific Advisory Board for Immunitas Therapeutics. S.C., J.L. and J.F. are employed by Cell Signaling Technology Inc. G.B., N.K., W.F. and K.W. are employed by Novartis Inc. J.Y., J.-Y.J.L. and L.-H.L. are shareholders of, and J.-YJ.L. and L.-H.L. are directors of, ImmunoQs Pte Ltd, which provides commercial tissue image analysis services based on the cellXpress software. J.S., L.G., J.A.G. and K.H. are employed by F. Hoffmann-La Roche Ltd. P.S. reports research funding (to institution) from AbbVie, BeiGene and Pfizer and a mobility grant from Institut Servier. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41592-026-03317-1.

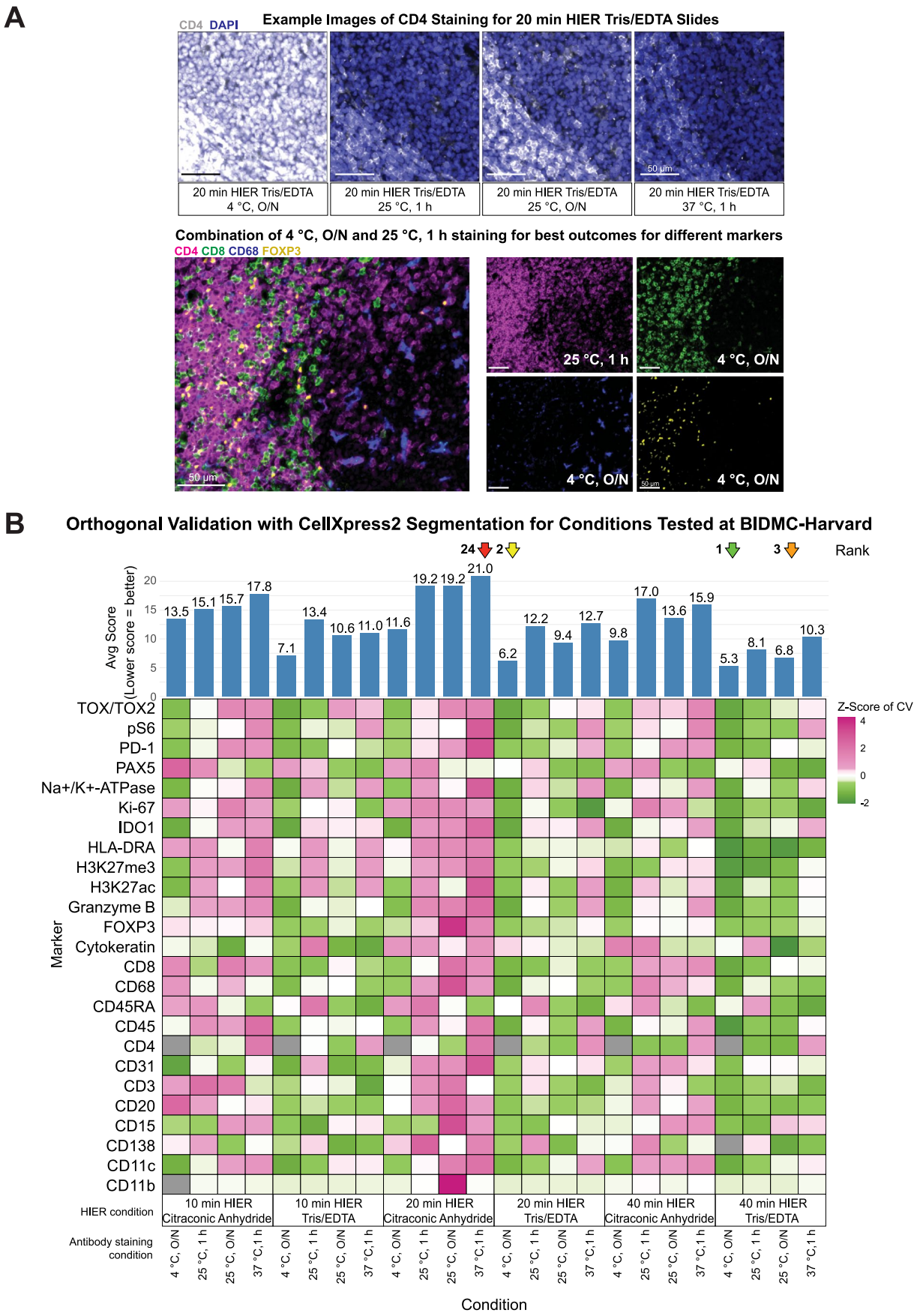
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-026-03317-1.

Correspondence and requests for materials should be addressed to Sizun Jiang.

Peer review information *Nature Methods* thanks Anja Hauser, Arutha Kulasinghe and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Rita Strack and Madhura Mukhopadhyay, in collaboration with the *Nature Methods* team.

Reprints and permissions information is available at www.nature.com/reprints.

https://doi.org/10.1038/s41592-026-03317-1

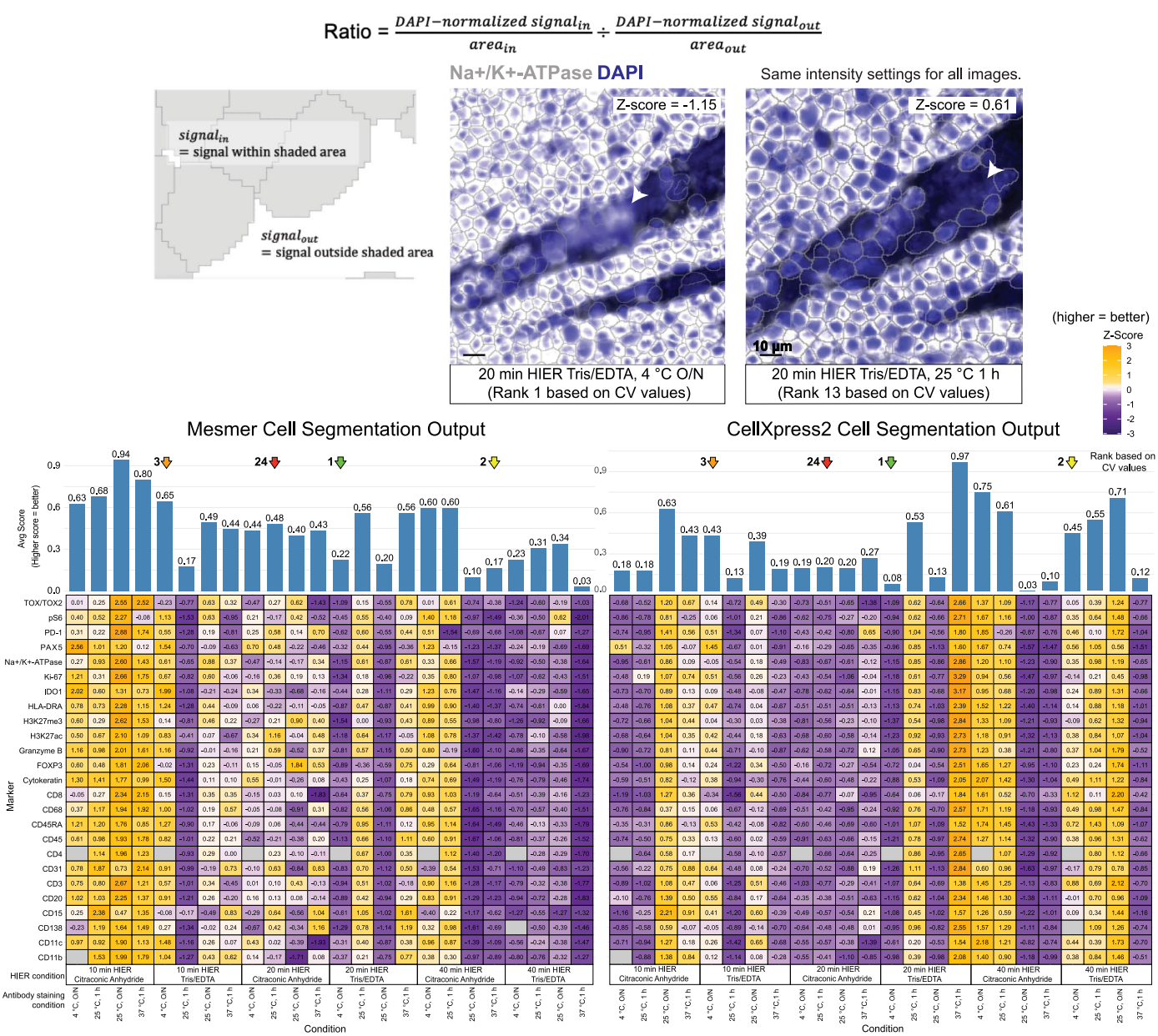


Extended Data Fig. 1 | See next page for caption.

https://doi.org/10.1038/s41592-026-03317-1

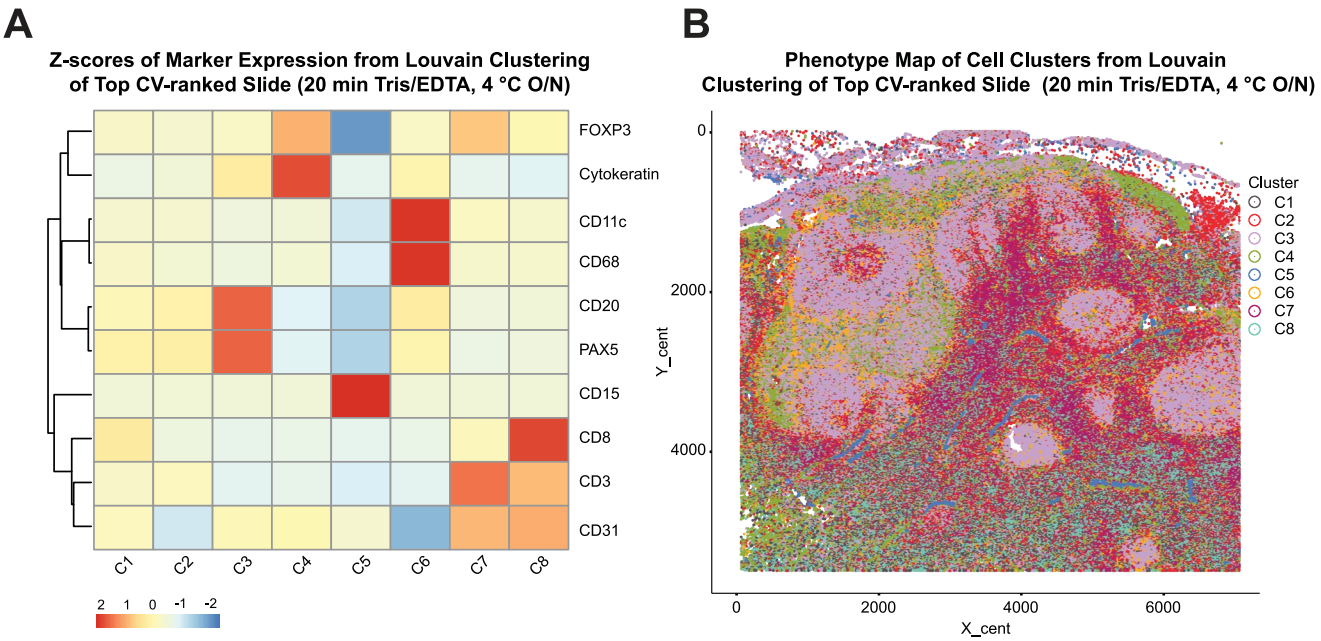
Extended Data Fig. 1 | Effects of different antibody staining conditions on CD4 staining and orthogonal validation with an alternative cell segmentation method, cellXpress2. (A) Upper panel: CD4 (white) and nuclei (blue) for tissues treated with 20 minutes Tris/EDTA HIER and different antibody staining conditions. All images on identical scales, crop areas and min/max intensity settings across slides. Scale bar = 50 μm. Lower panel: 4 °C overnight staining of antibody cocktail (CD8, green; CD68, blue; FOXP3, yellow) with subsequent 25 °C, 1h staining of CD4 (magenta) to improve CD4 marker staining performance (lower panel). Merged image (left): Scalebar = 50 μm; single markers (right): Scalebar = 50 μm. **(B)** Heatmap of Z-scored coefficients of variation (CVs) for all 25 phenotypic and functional markers across the 24 different HIER and staining protocols evaluated with the Akoya PhenoCycler-Fusion at BIDMC-Harvard following cell segmentation with cellXpress2. Numbers and bar plots indicate the average rank score achieved by each condition. Arrows indicate the top-ranked (green, 1), second-ranked (yellow, 2), third-ranked (orange, 3) and bottom-ranked (red, 24) conditions. Lower scores indicate higher staining intensity. Grayed-out entries indicate failed markers.

https://doi.org/10.1038/s41592-026-03317-1



Extended Data Fig. 2 | Comparison of the signal intensity ratios inside versus outside the cell mask across 24 conditions tested at BIDMC-Harvard. Upper panel: Schematic of signal intensity ratios evaluation by dividing DAPI-normalized signal within the cell with DAPI-normalized signal outside of the cell boundaries. Example images of Na+/K+-ATPase (white) and nuclei (DAPI, blue) for 20 min HIER with Tris/EDTA with 4 °C overnight staining and 25 °C 1 hour staining and their corresponding Z-scores (upper panel). Scale bar = 10 μm. Lower panel: Heatmap showing the signal intensity ratios of all 25 markers across the 24 conditions assessed at BIDMC-Harvard following Mesmer cell segmentation (left) versus cellXpress2 cell segmentation (right). Yellow = high signal intensity ratio; Purple = low signal intensity ratio. Avg Score = Average signal intensity ratio Z-score.

https://doi.org/10.1038/s41592-026-03317-1



C

Comparison of Spatial Sampling Efficiency across the 24 Staining Conditions using Balagan Analysis

Rank based on CV values (left) and Median Rank based on τ value across multiple FOV widths (right).

Slide (Ordered by Median Tau Rank)	CV Rank	FOV Width (μm) 50	100	150	200	250	300	350	400	450	500
10 min CA, 37 °C,1 h	21	1	1	1	1	1	1	1	3	6	10
10 min CA, 25 °C, 1 h	16	2	2	2	2	2	3	3	1	2	3
20 min Tris, 4 °C, O/N	1	12	4	4	3	3	2	2	2	1	1
40 min CA, 25 °C, O/N	14	7	5	3	4	4	4	4	4	3	2
10 min CA, 4 °C O/N	15	23	15	5	6	5	5	5	5	4	4
20 min CA, 25 °C 1 h	24	21	10	7	5	6	6	6	6	7	6
20 min CA, 4 °C O/N	10	5	18	18	10	7	7	8	7	5	5
40 min CA, 37 °C 1 h	18	4	3	6	7	9	8	9	9	9	8
10 min Tris, 4 °C O/N	3	22	12	9	9	8	9	7	8	8	7
40 min Tris, 37 °C 1 h	7	6	7	8	8	10	14	12	11	15	16
40 min Tris, 4 °C O/N	2	16	21	16	11	11	12	13	10	11	11
10 min Tris, 37 °C 1 h	8	18	11	11	12	12	11	15	14	13	14
10 min Tris, 25 °C 1 h	17	9	14	15	13	13	13	11	12	14	12
10 min CA, 25°C O/N	19	11	8	12	14	15	15	14	15	12	13
20 min Tris, 4 °C O/N	4	15	17	17	15	14	10	10	13	10	9
40 min Tris, 25 °C 1 h	11	13	9	14	16	16	18	16	18	18	18
10 min Tris, 25 °C O/N	9	8	20	19	19	19	16	17	16	16	15
20 min Tris, 25 °C 1 h	13	19	13	13	17	17	17	20	17	17	17
20 min CA, 25 °C 1 h	23	3	6	10	18	18	19	18	19	19	19
20 min Tris, 37 °C 1 h	12	17	19	20	21	20	20	19	20	20	21
40 min Tris, 25 °C O/N	5	14	16	22	20	22	21	21	21	23	20
40 min CA, 25 °C 1 h	20	24	22	21	22	21	22	23	24	24	24
40 min CA, 4 °C O/N	6	20	23	23	23	23	23	22	23	22	22
20 min CA, 37 °C 1 h	22	10	24	24	24	24	24	24	22	21	23

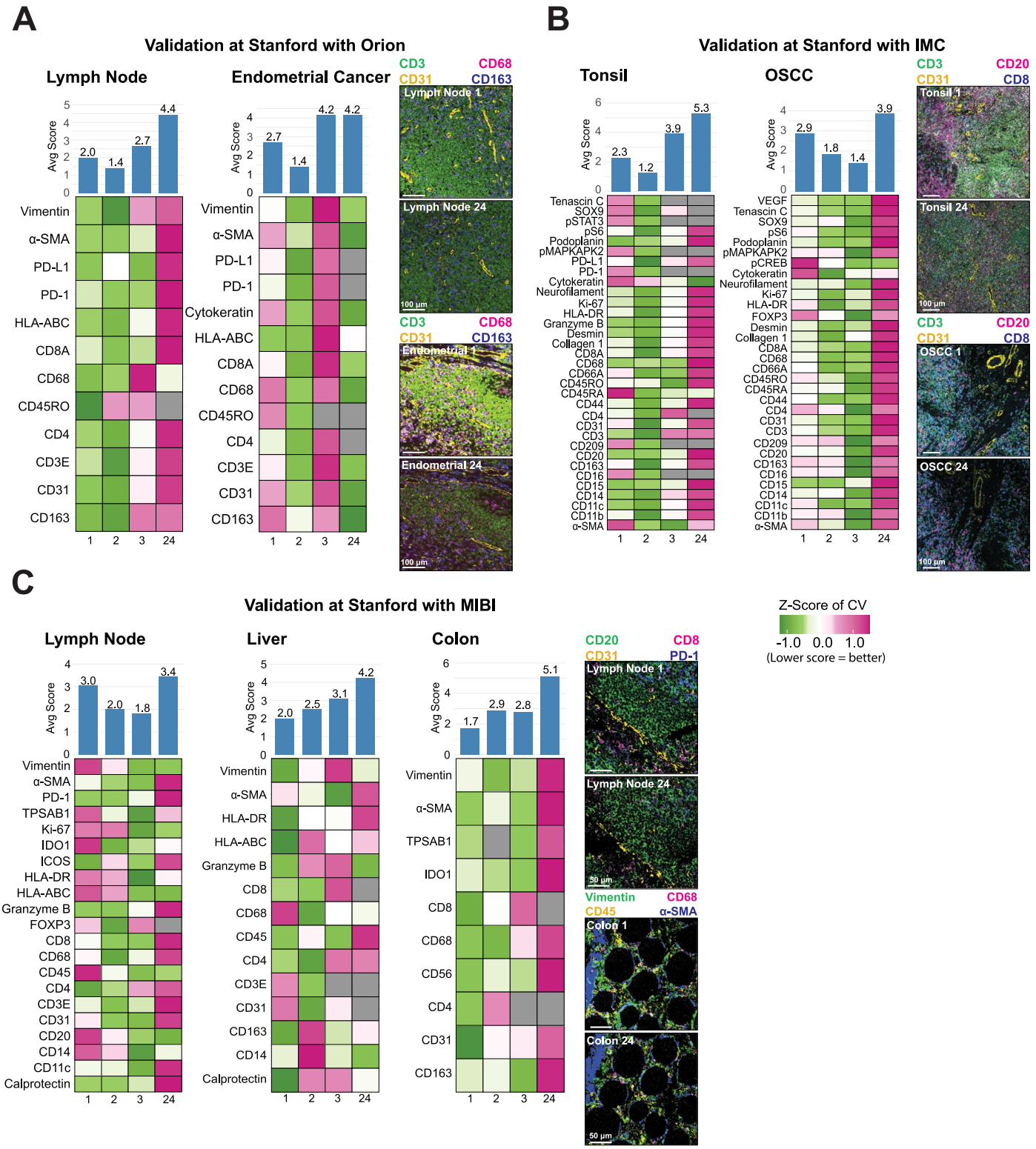
high sampling efficiency

Extended Data Fig. 3 | See next page for caption.

https://doi.org/10.1038/s41592-026-03317-1

Extended Data Fig. 3 | Louvain clustering and Balagan analysis of images from the 24 conditions tested on the Akoya PhenoCycler-Fusion at BIDMC-Harvard. **(A)** Heatmap showing the average protein marker expression for each cell cluster identified by Louvain clustering of the top ranked slide (20 min Tris/EDTA HIER, 4 °C O/N antibody staining), with colors representing Z-score scaling across rows (blue indicating relatively low expression, red indicating relatively high expression). C1-C8 = cluster numbers. **(B)** Phenotype map showing the spatial distribution of the clusters identified from Louvain clustering in A. **(C)** Comparison of spatial sampling efficiency (τ value) of all 24 staining conditions across multiple FOV widths assessed on the PhenoCycler-Fusion at BIDMC-Harvard. Conditions ordered by median tau rank from high (top, yellow) to low (bottom, blue) sampling efficiency across different FOV widths, with the ranks derived from the average CV scores shown on the left. CV = coefficient of variation.

https://doi.org/10.1038/s41592-026-03317-1

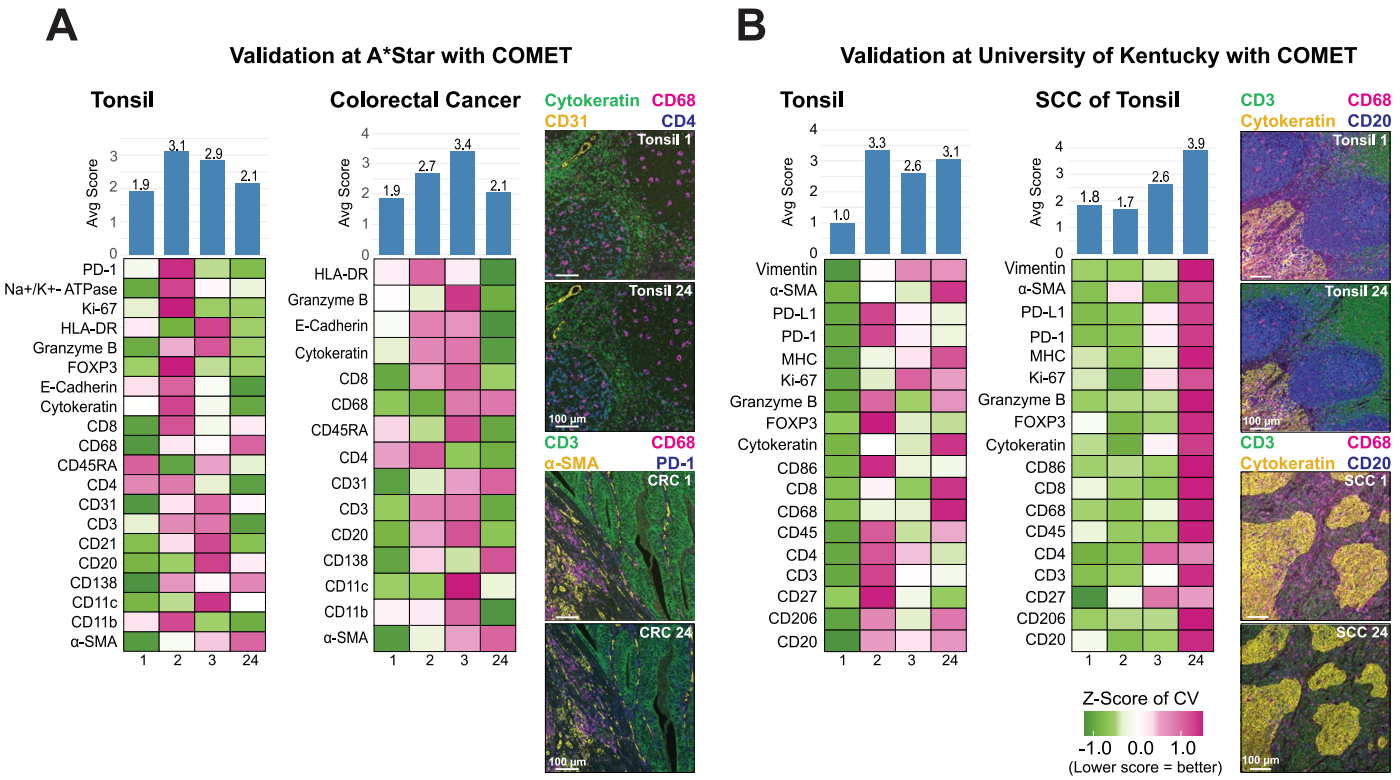


Extended Data Fig. 4 | See next page for caption.

https://doi.org/10.1038/s41592-026-03317-1

Extended Data Fig. 4 | Inter-site validation of the top three and bottom one conditions across different tissues and panels on the Orion, IMC, and MIBI. **(A-C)** Heatmap showing the Z-scored CVs and average rank scores of: **(A)** All 12 (lymph node) or 13 (endometrial cancer) markers across the 4 validation conditions (specified in Fig. 3A) on lymphoid (lymph node, left) and secondary (endometrial cancer, middle) tissue performed on the RareCyte Orion at Stanford. Images of lymph node and endometrial cancer tissues stained with the top and bottom conditions with markers showing T cells (CD3, green), myeloid cells (CD68, magenta), endothelial cells (CD31, yellow), and macrophages (CD163, blue) (right). (**B**) All 33 (tonsil) or 32 (OSCC) markers across the 4 validation conditions (specified in Fig. 3A) on lymphoid (tonsil, left) and secondary (OSCC, middle) tissue performed on the IMC at Stanford. Images of tonsil and OSCC tissues stained with the top and bottom conditions with markers showing T cells (CD3, green; CD8, blue), B cells (CD20, magenta), and endothelial cells (CD31, yellow) (right). **(C)** All 21 (lymph node), 14 (liver) or 10 (colon) markers across the 4 validation conditions (specified in Fig. 3A) on lymphoid (lymph node, left) and secondary (liver, middle left; colon, middle right) tissues performed on the MIBI at Stanford. Images of lymph node and colon tissues stained with the top and bottom conditions with markers showing B cells (CD20, green), T cells (CD8, magenta), endothelial cells (CD31, yellow), and activated lymphocytes (PD-1, blue) for lymph node, and mesenchymal cells (vimentin, green), myeloid cells (CD68, magenta), lymphocytes (CD45, yellow), and smooth muscle cells (α-SMA, blue) for colon. Low average rank score = high staining intensity. Grayed out = marker not working. All images from the same tissue type and site on the same scale and intensity settings. Scalebar = 100 μm (Orion, IMC); 50 μm (MIBI). CV = coefficient of variation. IMC = imaging mass cytometry. MIBI = multiplexed ion beam imaging. OSCC = oral squamous cell carcinoma.

https://doi.org/10.1038/s41592-026-03317-1



Extended Data Fig. 5 | Inter-site validation of the top three and bottom one conditions across different tissues and panels on the Lunaphore COMET. **(A-B)** Heatmap showing the Z-scored CVs and average rank scores of: **(A)** All 20 (tonsil) or 15 (colorectal cancer) markers across the 4 validation conditions (specified in Fig. 3A) on lymphoid (tonsil, left) and secondary (colorectal cancer, middle) tissues performed on the Lunaphore COMET at A*STAR. Images of tonsil and colorectal cancer tissues stained with the top and bottom conditions with markers showing epithelial cells (cytokeratin, green), myeloid cells (CD68, magenta), endothelial cells (CD31, yellow), and T cells (CD4, blue) for tonsil and T cells (CD3, green), myeloid cells (CD68, magenta), smooth muscle cells (α-SMA, yellow), and activated lymphocytes (PD-1, blue) for colorectal cancer (right). **(B)** All 18 markers across the 4 validation conditions (specified in Fig. 3A) on lymphoid (tonsil, left) and secondary (SCC, middle) tissues performed on the Lunaphore COMET at University of Kentucky. Images of tonsil and SCC tissues stained with the top and bottom conditions with markers showing T cells (CD3, green), myeloid cells (CD68, magenta), epithelial cells (cytokeratin, yellow), and B cells (CD20, blue) (right). Low average rank score = high staining intensity. All images from the same tissue type and site on identical scales, crop areas, and min/max intensity settings. Scale bar = 100 μm. CV = coefficient of variation. SCC = squamous cell carcinoma.

nature portfolio

Corresponding author(s): Sizun Jiang

Last updated by author(s): Jan 7, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Fusion data was collected with the manufacturer supplied software for the PhenoCycler/PhenoImager Fusion.
Data analysis	Cell segmentation and single-cell feature extraction were performed using Mesmer (DeepCell Python package v0.12.10) and cellXpress v2.4.0. Analysis code was executed in Python v3.10.16 and R v4.3.2. Image visualization, region-of-interest selection, and overlay of cell annotations were carried out in QuPath v0.4.4, and line-scan signal-to-noise quantification in ImageJ v2.14.0. Spatial sampling and tissue-heterogeneity analyses used the Balagan v1.0.0 R package. Cell clustering was performed with PhenoGraph (v1.5.7), and Uniform Manifold Approximation and Projection (UMAP) embeddings were generated with Scanpy v1.11.0 and AnnData v0.11.3 in Python. Factorial modelling of the coefficient of variation, including Type II ANOVA, partial eta-squared, and estimated marginal means, used the car (v3.1.3), effectsize (v1.0.1), and emmeans (v2.0.2) R packages. All plots were generated using ggplot2 (v4.0.1). Additional Python dependencies for the segmentation pipeline included TensorFlow v2.8.4, NumPy v1.26.4, pandas v2.2.2, scikit-image v0.24.0, and imagecodecs v2024.6.1. All analysis code is available at https://github.com/SizunJiangLab/Benchmarking_tissue_preparation.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The data from the pilot experiment including the raw images and .csv files from cell segmentation is available on Zenodo at 10.5281/zenodo.11410274. All raw images for the Initial Optimization, Validation, and Additional Experiments, and annotation input and output files are available on BioImage Archive at 10.6019/S-BIAD2491. Selected images are available on ImmunoAtlas. All .csv files from cell segmentation are available on Zenodo at 10.5281/zenodo.17843231.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	The sex and gender of donors for the FFPE tissues used in this study has not been provided as no clinical data was used in the analysis.
Reporting on race, ethnicity, or other socially relevant groupings	The race and ethnicity of donors for the FFPE tissues used in this study has not been provided as no clinical data was used in the analysis.
Population characteristics	The population characteristics of donors for the FFPE tissues used in this study has not been provided as no clinical data was used in the analysis.
Recruitment	No patients were recruited for this study.
Ethics oversight	No samples were collected or provided explicitly for this study and are covered by institutional ethics oversight for their respective studies.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Adjacent sections of the same tissue were used to benchmark the different staining or storage conditions. For the pilot data, 4 consecutive tissue sections were used. For the initial optimization performed at BIDMC-Harvard, Stanford, and Roche, 24, 12, and 6 consecutive tissue sections were used respectively. For the validation experiments of the top three and bottom one conditions, each site used 4 consecutive tissue sections each of two tissue types. These sample sizes were selected to cover all permutations of different antigen retrieval buffer, antigen retrieval durations, and antibody staining conditions.
Data exclusions	Markers that were assessed as not working were excluded from analysis (more details in Supplementary Figure 1C).
Replication	Following initial optimization performed at BIDMC-Harvard, validation of the top three and bottom one conditions were performed across 8 sites to validate reproducibility of the results.
Randomization	Sample randomization was not applicable as adjacent sections of identical samples were used for all experiments. This is because this manuscript aims to compare the effect of different processing of identical samples.
Blinding	Blinding was not relevant to the study as investigators will need to be cognizant of the assigned staining condition for each slide in order to carry out the experiment. However, pathologists were blinded from the staining conditions when performing pathological assessment of marker performance to ensure fair judgment.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used: All antibodies including clone, company name, catalog number, and concentrations/dilutions used are listed in Supplementary Tables 2,5, and 12.

Validation: We have partnered with Cell Signaling Technologies to access oligo-tagged CODEX-compatible antibodies purified using fast protein liquid chromatography that are lot-controlled and each antibody was validated internally on human tonsil sections. Other antibodies are bought commercially, conjugated in-house, and validated internally. Where relevant, staining results are reviewed by certified pathologists.

Plants

Seed stocks: NA

Novel plant genotypes: NA

Authentication: NA