Dec 15, 2025 Version 3

Greenwald, Nederlof et al extended methods V.3

DOI
https://dx.doi.org/10.17504/protocols.io.e6nvw44k7lmk/v3

Noah Greenwald[1]

[1]Stanford University School of Medicine

Noah Greenwald

DOI: https://dx.doi.org/10.17504/protocols.io.e6nvw44k7lmk/v3

Protocol Citation: Noah Greenwald 2025. Greenwald, Nederlof et al extended methods. **protocols.io** https://dx.doi.org/10.17504/protocols.io.e6nvw44k7lmk/v3Version created by **Noah Greenwald**

Protocol status: Working
We use this protocol and it's working

Created: December 15, 2025

Last Modified: December 15, 2025

Protocol Integer ID: 234998

Keywords: extended methods detail, methods this protocol, extended method, protocol, nederlof et al

Abstract

This protocol provides extended methods details about our work

MIBI Data Processing

1 **Image compensation with Rosetta**

To systematically correct the sources of contamination and background present in the MIBI data, we used an approach analogous to flow-cytometry channel compensation called Rosetta. To do this, we linked sources of potential contamination or spillover, such as spillover from the gold channel, organics, and adducts (source channels) to the channels where that signal showed up (target channels). We used previously **published** as well as updated values from the MIBIScope manufacturer (IonPath, Menlo Park, USA) to derive coefficients for isotopic impurities in the metal conjugates, as well as instrument-specific coefficients for elemental background in each channel (e.g. adducts and oxides). To account for organic signal, we extracted a range of the spectrum (AMU 117-125) without any true signal to serve as a template for non-specific signal, colloquially referred to as “noodle” due to its characteristic appearance. We also used the gold channel as a proxy for slide background, since the slides are gold-coated. We manually identified the appropriate coefficients for the gold and noodle channel by iterative refinement. This approach was applied to the entire antibody panel included in this study (**Figure 1**).

Following identification of the appropriate coefficients for each source of contamination, we constructed a matrix to hold the coefficients. Each row in the matrix represents a source channel, and each column represents a target channel. The value of each entry ij in the matrix represents the sum of all sources of contamination from the mass in row i for the target channel in column j. This quantifies the proportion of signal in each source channel that appears in the target channel. To correct the target image, we subtracted the source channel multiplied by its coefficient from the target channel (**Figure 2a**). We used gaussian smoothing to convert the integer counts into decimals to enable fractional compensation. Representative images taken before and after noise removal with Rosetta can be seen in **Figure 2b**.

2 **Intensity normalization using median pulse height**

Over the course of a run, the MIBI instrument will gradually lose sensitivity due to aging of the ion detector. Calculating this decrease in sensitivity by looking directly at the image data is challenging, because it is difficult to tease out whether a given change is due to biological or technical reasons. When an ion hits the detector in the MIBI, it produces an electrical pulse. Pulses over a threshold height are recorded as ion hits, whereas pulses under this threshold are discarded. This produces the count-based data that the user interacts with. Over the course of a run, the height of these pulses decreases, such that ions with the same intensity will record shorter and shorter pulses, with more and more of them falling under the threshold. When the sensitivity of the instrument is adjusted, the voltage to the detector is increased such that the height of these pulses is higher. However, looking at the binarized count data (the number of

pulses over the threshold), it is challenging to determine if the decrease in counts is due to a decrease in the height of the pulses (technical, decrease in instrument sensitivity), or a decrease in the number of pulses (biological, less signal in the sample).

To circumvent this issue, we used the median pulse height (MPH) to derive an estimate of the purely technical decrease in sensitivity (**Figure 3a**). Because the height of the pulses is determined exclusively by the voltage supplied to the detector, it is independent of the amount or intensity of the protein staining in a given image. Therefore, by calculating the median of the heights for each channel, we can get a robust, independent estimate of instrument sensitivity. Importantly, we can calculate this quantity directly from the image data being acquired for the study, obviating the need to repeatedly measure control samples over the course of the run.

To use the MPH values to correct for instrument drift, we quantified the relationship between MPH and sensitivity. To do this, we constructed a tuning curve. We used a synthetic polymer, poly-methyl methacrylate (PMMA), sample with fixed ratios of metal isotopes to ensure that any change in signal over was due purely to changes in the instrument setup, not sample-specific differences. We then systematically increased the detector voltage, and hence the MPH, and calculated the change in signal, performing this for three replicates at each detector setting. We normalized the resulting signal by the maximum observed, which allowed us to construct a graph relating MPH to the percentage of maximum signal, which we refer to as sensitivity. We then fit a polynomial to this curve, which we can use to convert MPH values to sensitivity. We use the same sensitivity curve for all images.

After generating the sensitivity curve, we calculate the MPH for each channel in each image in a given run. Because the estimate of MPH can be noisy, we generate a per-mass curve that we fit over the course of a run (**Figure 3b**). We use the fitted value to generate a value for the MPH of each mass in each image. We then plug that MPH value into the sensitivity curve to generate the per-mass, per-image sensitivity (**Figure 3c**). Finally, we use this sensitivity estimate to normalize each image, dividing by the sensitivity to bring the values up to 100% sensitivity. Looking at the pre-normalized images over the course of a run, we can see a decrease in signal, especially in the second half of the run (**Figure 3d**). However, following normalization, this decrease in signal is no longer apparent (**Figure 3e**).

3 **MIBI data QC**

To identify potential batch effects in the data, we measured variation in signal intensity across slides, as well as the spatial variation across cores within each slide. To identify potential batch effects across slides, we used the control samples present on each as a reference. We then computed the distribution of the mean of non-zero pixels in each channel in each control sample. Overall, we saw relatively minor shifts in intensity across slides (**Figure 3f**), and the one channel that showed the strongest differences was

Calprotectin, which had very few positive cells in the control samples, and thus noisy estimates of positive signal. To identify if different slide locations displayed different sensitivity, we calculated the same metrics as above across each of the cores from the TMA samples. We did not observe any spatial bias in signal intensity across the slides in the cohort (**Figure 3g**).

Cell clustering

4 **Pixie clustering**

We performed pixel clustering using the lineage markers contained in the MIBI panel following segmentation with Mesmer (**Figure 4**). We overclustered the data into 225 distinct pixel clusters, which were then combined into 33 pixel meta clusters. We labeled these pixel meta clusters based on marker co-expression patterns, such as CD3+/CD4+/CD45+, ECAD+/CK+, etc, as described in the original publication. We modified the pipeline slightly from the originally described version to improve the clustering on our data. In particular, we made the following changes:

- To ensure that channels with significantly higher intensity values did not swamp the per-pixel signal, we added a 99.9th percentile normalization step prior to pixel clustering.
- To reduce the impacts on downstream cell clustering from noisy channels, we added the option to exclude non-nuclear signal from nuclear markers (which we used on the FOXP3 channel), as well as the option to remove nuclear signal from non-nuclear markers (which we used on the CD11c channel).
- We removed the pixels with the 5th percentile lowest total marker expression, as these were mostly noise and interfered with subsequent cell clustering

Each of these steps is now included in the Pixie GitHub repository as an option so that users can reproduce this workflow, as well as in the code accompanying our paper.

Following generation of the pixel meta clusters, we calculated the proportion of each pixel meta cluster present in each segmented cell. We then used these proportions as more robust estimates of marker expression in each cell, and fed this into the second step of Pixie, cell clustering. We first over clustered the data into 225 cell clusters, which were then combined into 33 cell meta clusters. We did not make any modifications to the cell clustering portion of the pipeline.

5 **Cell meta cluster inspection and cleanup**

To confirm accurate clustering, images were manually inspected using Mantis Viewer. Representative images were selected, and cell assignments were manually verified by toggling on the appropriate image channels, which were then overlaid with the segmentation boundary and cell meta cluster assignments. Systematic errors in clustering were then addressed by redoing the meta clustering for the pixel clusters, cell clusters, or both, as appropriate. Following final assignment of cell meta clusters with Pixie, an additional round of cleanup was performed using manual thresholds to break up

ambiguous or mixed cell meta clusters, and to merge duplicate clusters. This thresholding was done using cutoffs based on total expression of individual markers, with all thresholds recorded and documented in the accompanying analysis code for reproducibility.

Our clustering process produced 33 total cell meta clusters, which we refer to as the 'detailed' level cell meta clusters because they are the most granular. However, it is challenging to visualize 33 distinct colors on a plot, and many of these cell types are quite similar to one another. As a result, to ease interpretation, improve clarity of the graphs, and reduce the number of duplicative features, these 33 detailed cell meta clusters were merged into 21 intermediate cell meta clusters. This intermediate level of clustering was used as input for the majority of the SpaceCat features. However, even this intermediate level was too detailed for some of our analyses. To capture the broad, overarching trends and changes in cell populations, we collapsed the intermediate clusters into 8 broad cell meta clusters. This three-level hierarchy was used to enable analyses at different levels of granularity.

6 **Clustering tradeoffs**

The limited number of antibody channels available necessarily requires certain decisions to be made about which markers to include and which to exclude. As a result, it is not possible to identify all cell populations at the maximum level of granularity. We highlight below how this impacted our clustering scheme:

Due to limitations in our panel, we were not able to unambiguously identify dendritic cells. As a result, we classified HLA-DR+ and/or CD11c+ cells as antigen presenting cells.

Our classification of fibroblasts is inspired by Costa et al; however, because we did not have the full complement of markers included in our panel, we subclassified fibroblasts into CAF-S1 (FAP+) and CAF-Other (FAP-/SMA-). CAF-S1 fibroblasts are typically described as immunosuppressive, aligning with the traditional pro-tumorigenic view of CAFs. However, Kieffer et al. revealed that human breast CAF-S1 cells consist of distinct subpopulations, including myofibroblastic CAFs (associated with immunosuppression and resistance to immunotherapy) and inflammatory CAFs (**correlated with CD8+ T cell infiltration and inflammatory signaling**). Thus, the CAF-S1 population we identify here may include diverse subpopulations with distinct impacts on the TME.

7 **Cluster validation**

Appropriately assigning cells that have low expression of the markers included in an antibody panel is a significant challenge. Similar to recent work[13], we identified a population of cancer cells that expressed very low levels of traditional epithelial markers. To determine if our clustering had under-counted these ambiguous cancer cells, we examined the alternate clusters that shared similar phenotypes. This included two fibroblast populations that were defined purely by ECM markers (which could conceivably be lowly expressing cancer cells surrounded by ECM), cells expressing only

mesenchymal markers (which could be EMT-like cancer cells), and cells that did not express any of the markers on our panel.

For each of the five populations, we manually inspected the images with the highest cell density to determine if their morphology resembled cancer cells. We observed some cells that appeared to be incorrectly classified. Although the frequency of these misclassified cells appeared low, the misclassified cells seemed to be clustered near one another, presumably because the neighboring cancer cells shared the same phenotype. To get a conservative estimate of the potential misclassification rate, we looked at the frequency of cellular neighbors around each of these cell types. We flagged any cell that had more than 70% of its neighbors as cancer cells (after excluding cells of the same cell type), as these could represent pockets of misclassified cells that are either surrounded by correctly classified cancer cells or other incorrectly classified cells of the same type (**Figure 8c**). While this analysis did identify potentially misclassified cells, the number of cells meeting this criteria was exceedingly low (**Figure 8d-e**), representing 3% of each cell type on average, and 0.2% of cancer cells.

SpaceCat Pipeline

8 **Defining tumor compartments**

We defined four compartments in each image: the cancer core, cancer border, stroma border, and stroma core. We used a smoothed version of the ECAD channel and the segmentations of the cancer cells to define a cancer mask. We binarized this mask, filled in small holes, and kept the regions that were above the size cutoff. We eroded the cancer mask by 50 pixels to define the cancer core, with the eroded regions being defined as the cancer border. We then expanded the cancer mask by 50 pixels, with the expanded region defined as the stroma border. All remaining area was defined as the stroma core.

We performed two steps to clean up these four compartments. We generated a separate mask to represent slide background (i.e. areas with no tissue), which was excluded from the compartment masks. We also generated a separate immune aggregate mask for large clumps of T and/or B cells. Looking across the FOVs in our cohort, we observed that 19% had at least one lymphoid aggregate (**Figure 8f)**. Of FOVs with at least one aggregate, they comprised 17% of the tissue in an image on average (range 0.7%-98%), with an average of 2.6 aggregates per FOV (**Figure 8g**). To ensure that removing the aggregates from the compartment calculations was not biasing our results, we repeated our outcomes analysis with the aggregates included. The features defined based on the aggregates did not effectively separate responders from non-responders (**Figure 8h**), and we did not observe a shift in overall predictive accuracy with the aggregates included (**Figure 8i**).

With these compartments defined, we assigned each cell to its respective compartment, assigning cells at the border between two compartments based on maximum overlap with the compartment mask. We calculated two sets of features that depended exclusively on the compartment masks: The area of each compartment in each image, as well as all pairwise ratios of compartment areas in each image.

Tumor compartments allow us to compare features across images that have different tissue architectures. For example, consider image A, which is primarily composed of a dense tumor nest, and image B, which captures just the edge of a tumor nest and is primarily composed of the surrounding stroma. Any image-level features calculated from these two images will primarily reflect the dramatically different representation of these cell types. For example, we would expect fibroblast density to be much higher in image B, as a consequence of the reduced presence of cancer cells. While this information is important for understanding the broad differences in representation between the images, we wanted a way to look at cellular features that wasn't so confounded by the region being imaged. Tumor compartments represent a way to extract this information. By looking at fibroblast density within the stroma compartment, rather than across the whole image, we can control for differences in the total number of cancer cells in an image. This lets us catalogue and compare cell type abundances, ratios, or any of the other features in SpaceCat independent of the compositional differences in the images.

9 **Cell abundance features**

To quantify cell type abundance, we calculated three categories of features. The first category was densities of individual cells. For each cell type at the broad and intermediate level of clustering, we calculated the density by dividing the number of cells by the area of the region. Densities were computed across the entire tissue area in the image, as well as in each compartment.

The second category was ratios between cell types. We computed all pairwise ratios between cell types at the broad level of clustering, as well as biologically motivated ratios for cells at the intermediate level of clustering (CD8 T/CD4 T, CD4 T/Treg, CD8 T/Treg, CD68 Mac / CD163 Mac). We set a minimum density threshold of 5×10^{-7} cells per square pixel based on manually inspecting the density histograms and looking at the corresponding images. If one of the density values in the ratio was below this threshold, it would be rounded up to the minimum density to avoid issues with division by zero. If both cell densities were below the threshold, the ratio for that image was not calculated. Ratios were computed across the entire image, as well as in each compartment

The third category was cell type proportions. We calculated this feature for cells at the broad level of clustering that were composed of at least two distinct intermediate cell types. For each intermediate cell type in a given broad cell type, we calculated the proportion of the number of broad cells that the intermediate cell type

represented. Proportions were calculated across the entire image, as well as in each compartment.

10 **Functional marker frequencies**

For each cell type at the intermediate clustering level, we calculated the frequencies of positivity for the functional markers on the panel. We manually inspected images to determine an appropriate marker-specific threshold for positivity and applied the same threshold for a given marker across all cell types. For a given cell type, we computed marker frequencies for each marker that was positive in at least 5% of cells of that type, to avoid markers that were not expressed in a given cell type. For each cell type/marker combination, we calculated the proportion of cells above the marker-specific threshold. We set a minimum cell count threshold of at least 5 cells in an image; we did not calculate marker frequencies in images with fewer than the minimum number of cells. Marker frequencies were calculated across the entire image.

11 **Cell morphology features**

We previously defined a range of morphology metrics to summarize differences in both cell and nuclear shape and size[22]. These metrics were computed across all of the cells in the dataset as part of our segmentation pipeline[61]. Manual inspection revealed that these metrics could reliably pick up differences in cancer cell morphology, but that morphological shifts in the other cell types were too subtle to be effectively captured by these automated metrics. We therefore included cell size for all cell types, and a non-redundant subset of the morphology metrics only for the cancer cells. We computed the average value for each of these metrics for each cell type across the entire image.

12 **Cell diversity features**

We calculated a range of metrics to capture the diversity of the cells in the image, which fell into two broad categories. The first category was based purely on cell abundances, not physical proximity. We first identified the clustering granularity used to perform the diversity calculation. We did this both at the broad cluster level granularity, as well as at the intermediate clustering granularity within immune, cancer, and structural cell populations. For each of these clustering resolutions, we extracted the proportion of cells of each cell type that was present. We then computed the Shannon index on these proportions. We computed these diversity scores across the entire image, as well as within each compartment.

The second category of diversity feature was those based on physical proximity. For each cell in each image, we computed the number of cells of each cell type within a 50-pixel radius. We then calculated the same Shannon diversity as above, but based on the proportions within the 50-pixel radius. We then calculated the average of this diversity value across the cells at the intermediate clustering resolution. This value was computed across the entire image, as well as within each compartment.

13 **Cell-cell distances**

For each cell at the broad clustering resolution, we calculated the centroid distance to the **nearest cell of every other type**. We then calculated the average across all of the cells in the image, as well as in each compartment in the image. We removed linear distances that were highly correlated with the density of the target cell type to only retain features that represented spatial structure, rather than increased abundance of the target cell.

14 **Mixing scores**

We previously described a classification of TNBC based on the degree of separation between cancer and non-cancer cell populations[19]. Patients with immune and cancer cells that associated closely with one another were termed “mixed”, while patients without close associations between these two populations were termed “compartmentalized.” The underlying metric used to define this classification was a mixing score, which is based on the fraction of cellular neighbors belonging to each population. We used the same approach to generate pairwise mixing scores between cell populations at the broad cluster resolution, with some modifications to how it was previously described. For each cell in the selected populations pairs, we computed the number of surrounding cells in a 50-pixel radius. We then calculated the number of heterotypic interactions as the number of cells of the opposite type within the radius, and the number of homotypic interactions as the number of cells of the same type within the radius. We used the ratio of heterotypic to homotypic interactions as the mixing score, which was averaged across all cells in the image.

15 **Cell neighborhoods**

We used k-means clustering to define cell neighborhoods in each image as previously described[19,58,59]. We computed the relative proportions of cells at the intermediate clustering resolution in a 50-pixel radius surrounding each cell. We used these proportions as inputs to kmeans clustering, selecting 12 total clusters based on the heatmap of cell frequency loadings, as well as manual inspection of the underlying images and neighborhood assignments. We then calculated the proportion of cells belonging to each of the identified cell neighborhoods across the entire image, as well as within each compartment

16 **Fiber features**

We used our previously described fiber segmentation pipeline[19,58,59] to segment out individual fiber objects in the images. For each fiber object, we defined key summary statistics such as length, area, elongation, eccentricity, and alignment with neighboring fibers. Fiber length was defined as the longest axis of each segmented object. As shown in **Figure 5i**, our fiber segmentation pipeline identified images with short vs long fiber length. We also highlight another key metric of fiber composition, the alignment score. Calculated as the average angle between a fiber and its five closest neighbors, this allows us to identify regions of high alignment, and contrast it with fibers that are more randomly arranged (**Figure 5j**). We averaged each of these

metrics across the entire image, as well as within 512×512 tiles, and added these as features.

17 **Extracellular matrix (ECM) image clusters**

To capture changes in acellular features, we generated image-level clusters based on the expression level of Collagen, Fibronectin, and FAP. We divided the images into tiles of 256×256 pixels, and generated a binary mask based on the three markers to define the ECM area. We calculated the total per marker expression, normalized by the mask area, for each tile. We discarded tiles that had less than 10% ECM area, and clustered the remaining tiles using k-means clustering into Cold Collagen (those with predominantly Collagen expression) and Hot Collagen (those with coexpression of Collagen and Fibronectin and/or FAP). We calculated the proportion of each image that was composed of Cold Collagen, Hot Collagen, or non-ECM tiles.

18 **ECM pixel clusters**

To capture pixel-level co-expression of ECM markers, we performed pixel clustering on Collagen, Fibronectin, FAP, SMA, and Vimentin using Pixie[23]. We identified a total of 15 distinct ECM pixel clusters (**Figure 5g**), which we used to compute a number of distinct features. We calculated the density and proportion of each cluster in each image as features. For each pixel cluster, we identified groups of contiguous pixels from the same cluster, which were binarized into masks. We then calculated morphological features of these pixel cluster masks to capture their shape. We calculated the per-cluster average of these morphological features for each cluster in each image, which we included as additional features.

To quantify the relationships between the distinct ECM clusters, we generated ECM neighborhoods. We looked at a 100-pixel window around each pixel, counted the number of each ECM cluster within that window, and clustered the pixels using these counts. Clustering was done using k-means clustering with the number of clusters set to five, based on visual inspection. We calculated the density and proportion of each neighborhood in each image as features.

19 **Aggregating computed features**

After computing each of the above image features, we aggregated them into a single data structure for downstream analysis. Each feature was given an informative name, as well as metadata relating to which image compartment it was calculated in, the cell types and/or markers used to calculate the feature, and the broad feature category it belonged to. For features that were in compartments, we assessed the correlation between the compartment-specific value and the image-wide value. Compartment features that had a greater than 0.8 correlation with the corresponding image-wide value were excluded. We then Z-scored each feature to enable easy comparison across feature types. Finally, for samples with multiple FOVs per timpeoint, we computed the average across all the distinct FOVs to use for downstream analysis.

20 **Assessing pipeline robustness**

To determine the impact of the key design decisions in our pipeline, we systematically varied different thresholds and cutoffs to understand their impact on the output. In general, we found that small changes to any of the parameters resulted in negligible changes, whereas larger (log_2 ratio of at least 1) produced a noticeable impact on the results. We first looked at the minimum cells per image used for calculating per-cell statistics. In line with the different abundance of different cell types, we observed varying baseline levels of exclusion across distinct cell types (**Figure 5b**). Altering this threshold (which is set at a five cell minimum) affected the number of FOVs excluded, with changes roughly proportional to the change in threshold. We next looked at the functional marker positivity thresholds. We observed relatively minor shifts in total positive cells for small relative shifts in the threshold, with the most significant changes for very low values of the thresholding, which resulted in many cells being called positive (**Figure 5c**). We then looked at the compartment masks, and analyzed how altering the radius of the border region impacted border size. We found a consistent relationship between border radius and total border area (**Figure 5d**), with the most significant deviations only occurring for large (log_2 ratio of at least 1) changes in the radius.

We performed a similar analysis for the neighborhood-based metrics to evaluate how the radius used to define interactions impacted our results by systematically varying the pixel radius. The neighborhood diversity metric showed a gradual, progressive increase in diversity as the radius was expanded, and a corresponding decrease as it was contracted (**Figure 5e**). Although the numerical value changed as the radius was adjusted, the relative position of the scores stayed quite consistent for moderate changes, and only deviated substantially once the radius became too small to effectively capture the neighboring cells. The mixing score demonstrated a similar trend, with progressive increases in mixing as the radius was expanded to include more cells (**Figure 5f**). These gradual changes did not impact the relative position of the mixing scores to one another, although there were global shifts.

Assessing univariate outcomes associations

21 **Importance score robustness**

We performed a permutation test to assess the robustness of the identified features and ensure that the importance score was accurately capturing informative features. We first randomly shuffled the outcome labels for each patient for 100 trials. We used these shuffled labels to regenerate the univariate outcome association metrics. We then computed the top 100 most predictive features in the shuffled data using the same importance score criteria as above, and compared them with the top 100 features in the real data. Looking at the distribution of p-values in the real data compared to one of the shuffled replicates, we see that the distribution of the real data is shifted to the right with almost no overlap, indicating that the real data has

much more significant p-values (**Figure 6a**). We then compared the average p-value in the real data with the average p-value in each of the 100 shuffled iterations, finding that the real data is completely non-overlapping with the observed distribution of shuffled averages (**Figure 6b**).

We performed the same analysis looking at the distribution of effect sizes, measured using the difference in medians between the two populations, as this is the second metric that is used to define the importance score. Looking at the distribution of medians from the real data and an individual permutation, we see a rightward shift in the medians, though not as extreme as was observed for the p-values (**Figure 6c**). Looking at the averages across all the replicates, we see that the observed average is almost completely non-overlapping with the randomized averages (**Figure 6d**), indicating that we are identifying features with much larger effect sizes and much smaller p-values than expected.

To determine if there were certain features which were more likely than others to come up as false positives, we looked at the specific features selected as part of the top 100 in each of the above replicates (**Figure 6e**). Overall, we observed relatively little overlap between the same features across distinct replicates, suggesting that there were not underlying factors influencing the features which we identified. We also looked at the correlation of the features selected as part of the actual top 100 compared to the rest of the dataset (**Figure 6f**). These features were more correlated with one another than with the non-selected features, which makes sense given the shared patterns in the types of response-associated features we identified.

22 **Evaluating enrichment within specific compartments**

We observed a strong enrichment of outcome associated features in specific compartments. To determine how the specific features included in the SpaceCat pipeline influenced this finding, we first looked at the distribution of features across compartments independent of outcome. We found that the cancer border was over-represented in features compared to the number expected to be present if all compartments were equally represented (**Figure 8a**). This over-representation occurs because fewer cancer border features are removed during our QC process compared to features in other compartments. The two reasons that a feature would be removed are 1) correlation with the corresponding image-wide value and 2) not containing the minimum number of cells to calculate a given feature.

This enrichment of the border compartment was true when looking at the panel as a whole, but might have been the result of the specific antibodies selected for our panel. Thus, we repeated the analysis, this time progressively removing immune features from the dataset to simulate a panel with a less of an immune focus. We saw similar patterns even after removing 50% of the immune features (**Figure 8a**), indicating that the enrichment is not due to an inherent property of the panel. Thus, the over-representation

of the cancer border region indicates that it is a dynamic area of the tumor with features that are not correlated with the image-wide value, and tends to be well-populated by the cell types we measured.

We then performed the same subsampling analysis, this time looking at the top 100 outcome associated features. The trend we observed in the paper, with outcome-associated features over-represented in the cancer border, held true even as we reduced the number of immune related features (**Figure 8b**). This shows that the enrichment of cancer border features in the top predictors is not due simply to a greater number of potential immune features, but is indicative of the importance of this region in understanding response.

23 **Potential confounders**

We observed a wide range of cell counts across the images in our cohort (**Figure 8j**). We wanted to understand what, if anything, was different about the sparse, low-cellularity images we observed. The low cellularity images were relatively evenly distributed, without obvious bias towards specific patients or timepoints (**Figure 8k-l**). We examined the features present in these low cellularity images, and compared them with the features across the entire cohort. We found that diversity-based features differed most strongly between low and high cellularity images (**Figure 8m**), which makes sense given that high diversity values require the presence and enumeration of many distinct cell types, which will be less likely in an image with many fewer cells.

Assessing multivariate outcomes associations

24 **Lasso model analysis**

To gain insights into the model's decision-making process, we examined the weights assigned by the Lasso model to each feature. The Lasso model's sparse solution results in many feature weights being set to zero, effectively performing feature selection. By analyzing the non-zero feature weights, ranked by their absolute magnitude, we can identify the most important features contributing to the model’s predictions. The magnitude of the non-zero weights provides a measure of the relative importance of each selected feature, with larger absolute values indicating a stronger influence on the model's predictions of patient response.

To identify the features most important for the model’s prediction, we defined a set of top features across the 10 replicates. In order for a feature to be considered a top feature, it needed to be selected by the model in at least three separate replicates, and it needed to have a weight of at least 30% as large as the largest feature. Using this criteria, we sought to understand how consistently top features were identified across distinct replicates. We found that more than half of the top features were identified in at least 80% of trials, indicating good consistency across the distinct splits.

One of the reasons to use a Lasso model is that the output is sparse, meaning only a small subset of the features are included. We compared the univariate scores of the top features selected by the model. For the on-nivo MIBI model, the top features had very high importance scores. The on-nivo RNA model had one top feature with high importance score (the cytolytic activity score), and one feature with a relatively low importance score (anti-tumor cytokines), indicating that the model prioritized different information from the univariate analysis. In agreement with the relatively small degree of sharing across timepoints for features identified in the univariate analysis, top features from the models likewise were not often shared across timepoints.

Our analysis showed the on-nivo timepoint as the most predictive. To further evaluate the predictive capability and generalizability of this model, we divided the dataset using stratified splitting into training and test sets, with a 70% (43 patients) and 30% (19 patients) ratio. We standardized the training set and applied the derived mean and standard deviation to the test set to prevent data leakage. The optimal level of sparsity in the Lasso model was determined through stratified CV within the training set. The model's performance was then evaluated on the unseen test set, with results reported in terms of both the AUROC and Area Under the Precision-Recall Curve (AUPRC). We observed similar performance with this setup (AUROC=0.875) compared with the cross-validation framework we used for the preceding analyses, indicating that the performance metrics we used were unlikely to be inflated by data leakage.

25 **SpaceCat validation**

To validate SpaceCat's performance, we reprocessed the data from Wang et al. to facilitate a direct comparison. We generated compartment masks using the same pipeline as described above for the TONIC cohort, and used the author's original cell type annotations. We then ran SpaceCat on their data to generate features, and fed these features into the same Lasso-based multivariate modeling framework to predict patient outcome. We used the timepoint definitions from the original publication, where the authors analyzed baseline only, on-treatment only, or baseline + on treatment. They also looked at either the chemotherapy arm alone, or combined the chemotherapy and immunotherapy arms.

We found that SpaceCat-based features performed as well as the original author's features for predicting outcome, and we observed highly concordant results using our features or the original features from the study. SpaceCat-based features tended to do slightly better at predicting response for some timepoints (such as chemotherapy at baseline and chemotherapy at baseline + on-treatment), whereas at others the performance was equivalent (such as on-treatment for the chemotherapy + immunotherapy arms).

We also performed the reciprocal analysis, where we implemented the feature engineering pipeline from Wang et al. and applied it to the TONIC samples. We then re-ran our multivariate modeling using these features instead. We likewise saw similar results, with minor differences in accuracy across timepoints but no substantial shifts. Although the features from Wang et al. captured sufficient information to predict response, they were not as information-rich as the SpaceCat features, as they did not contain information about specific compartments or cell ratios. This was true even when we re-ran our clustering to more closely resemble the specific cell labels from Wang et al (**Figure 7**)

Finally, we investigated whether combining the two feature engineering approaches would yield better results. Looking at both the TONIC and NeoTRIP datasets, we did not see an increase in accuracy when using both sets of features as inputs to the model.

Figures

26

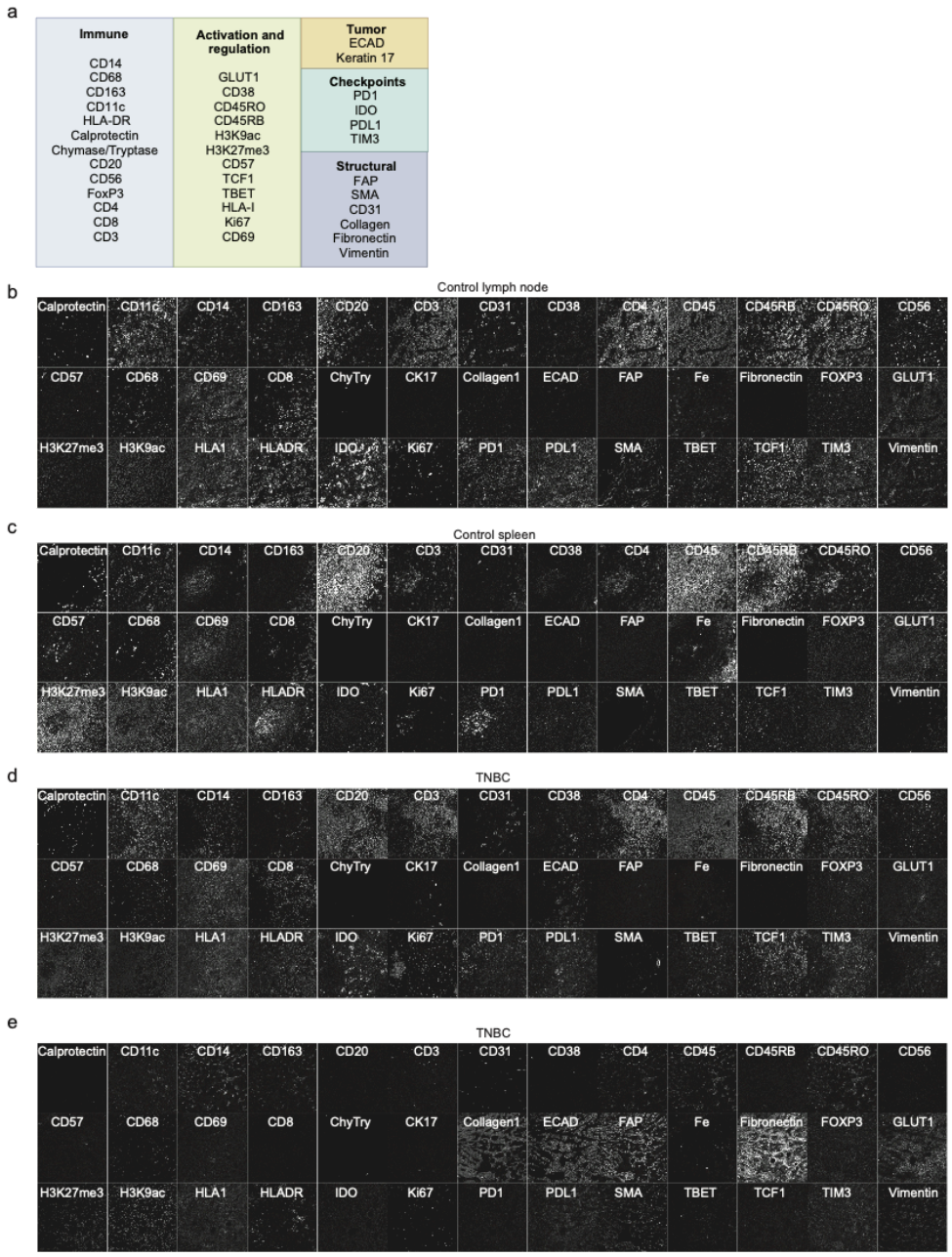

Figure 1. Antibody panel validation

a. MIBI panel, including immune markers, tumor markers, checkpoint markers, structural markers and markers related to activation and regulation.

b. Single-plex images of each marker in the panel in a control lymph node sample

c. Same as b) for a control spleen sample

d. Same as b) for first representative TNBC sample

e. Same as b) for second representative TNBC sample

27

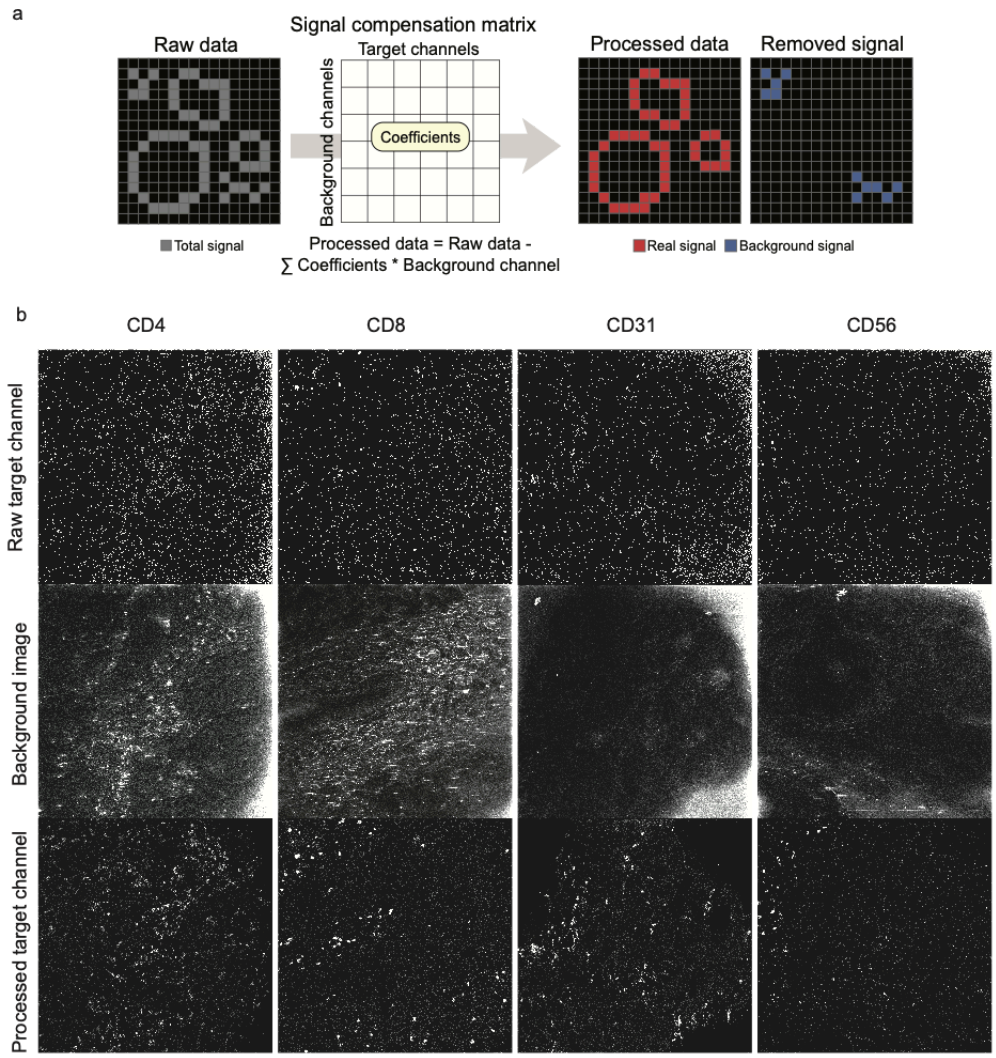

Figure 2. Rosetta image compensation
a. Cartoon illustrating the Rosetta image compensation process, where background signal is subtracted using a compensation matrix to produce cleaned up images
b. Representative examples of different channels before compensation (top), after compensation (bottom), along with the corresponding background channel used for compensation (middle) from the same image

28

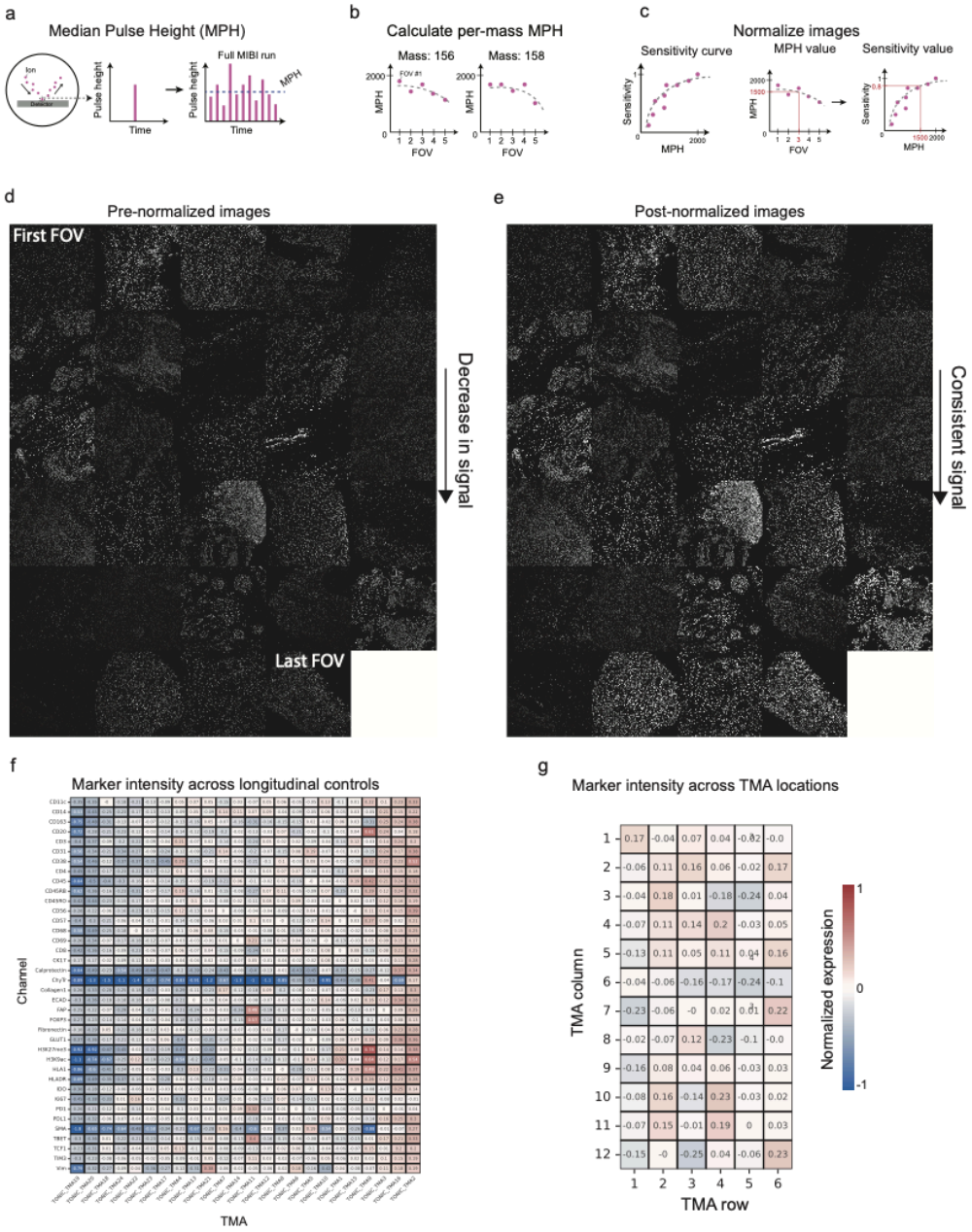

Figure 3. Image normalization and quality control

a. Cartoon illustrating how median pulse height (MPH) is calculated

b. Cartoon showing how a per-mass MPH curve is fit for an imaging run

c. Cartoon showing how the sensitivity curve is combined with the per-mass MPH curve to correct for changes in sensitivity over the course of an imaging run

d. Representative stitched image showing the uncorrected images from a single channel across an entire run, stitched in acquisition order from the start of the run (top left) to the end of the run (bottom right)

e. Representative stitched image showing the corrected intensities from the images in d) following MPH normalization

f. QC plot showing the normalized expression in the control tissues of all the markers in the panel (y-axis) across the different TMAs in the cohort (x-axis).

g. QC plot showing the normalized expression per field-of-view, plotted with the same row/column location as in the actual TMA

29

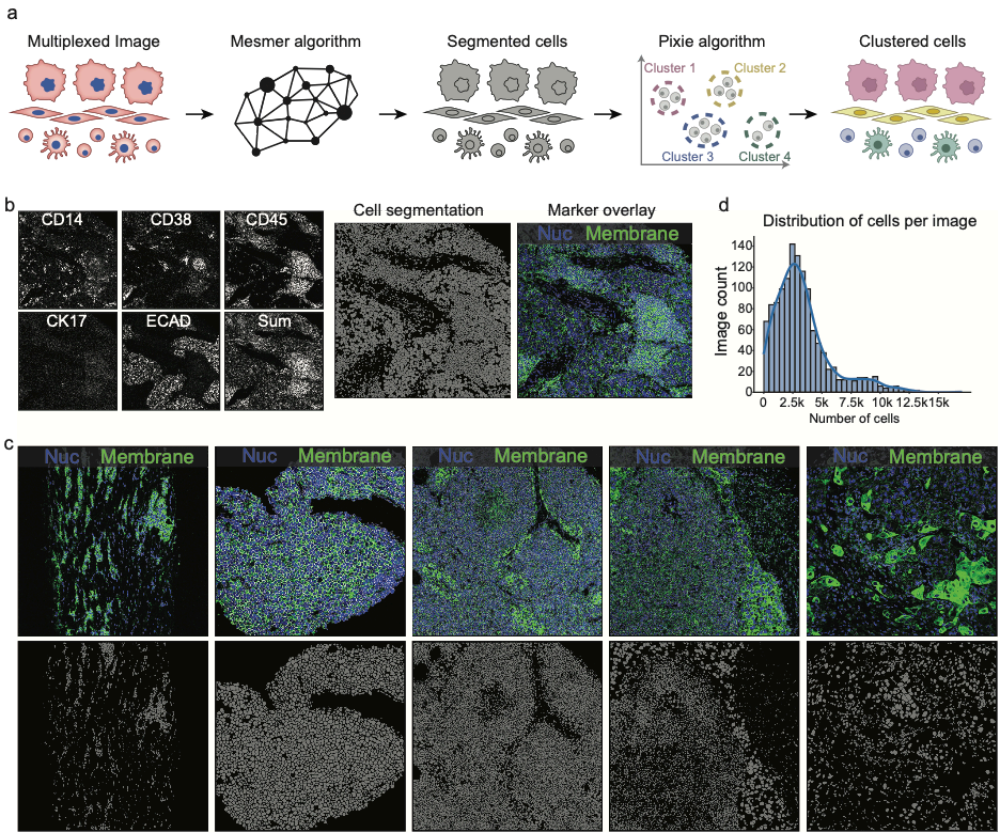

Figure 4. Cell segmentation

a. Cartoon of the segmentation and clustering pipeline

b. Representative image showing the different membrane markers that were combined together into a single membrane channel for segmentation (left), the resulting segmentation mask generated by Mesmer (middle), along with an overlay of segmentation mask and channels used for segmentation (right)

c. Additional representative images of showing the segmentation channels overlaid with the segmentation outlines (top), along with the segmentation mask on its own (bottom)

d. Histogram showing the number of cells per image across all images in the cohort

30

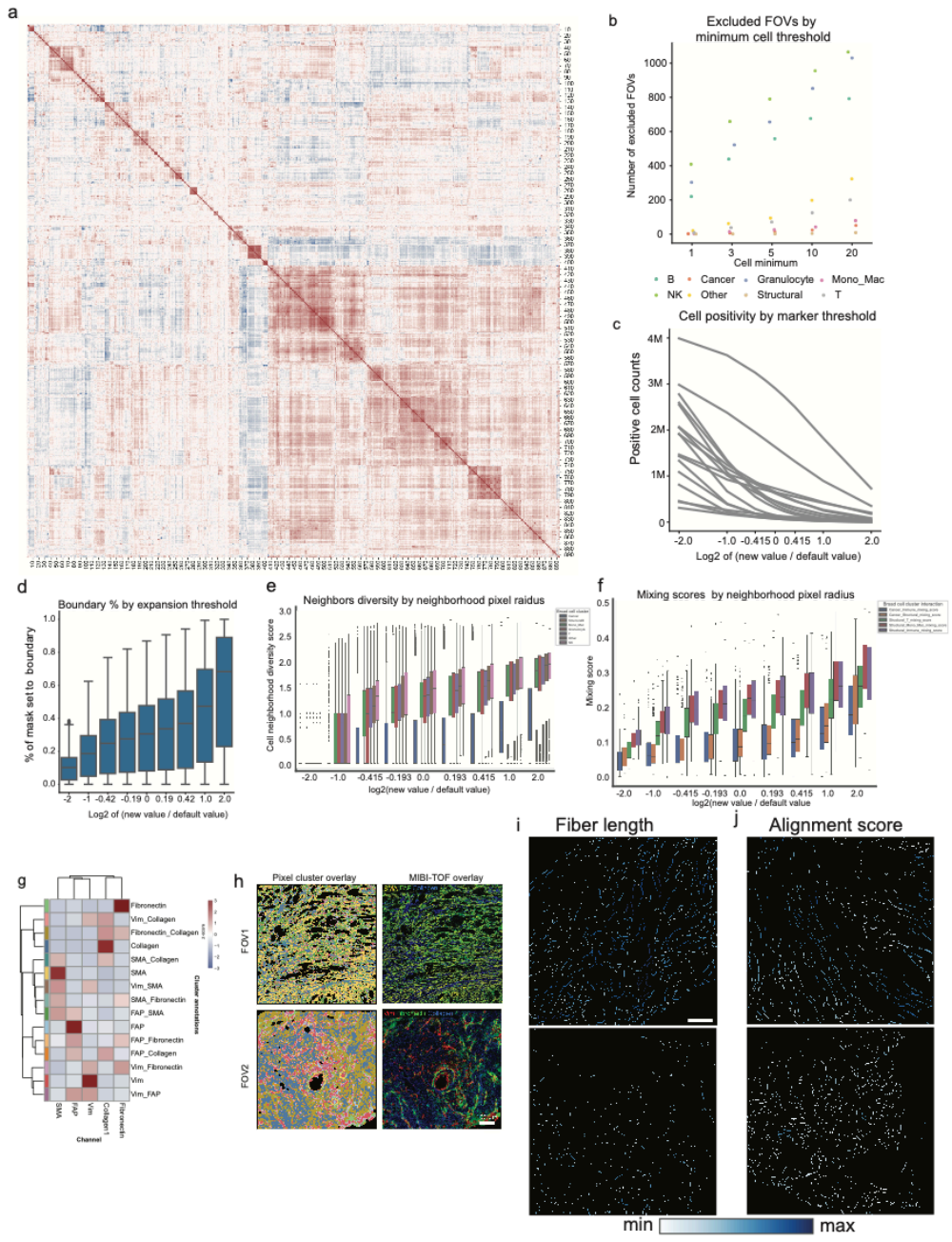

Figure 5. Validation of SpaceCat features

a. Figure2C correlation heatmap with numbered labels

b. Number of images removed from analysis as a function of different minimum cell thresholds, stratified by cell type

c. The number of functional marker positive cells across different thresholds

d. The proportion of the image assigned to the border compartment across different expansion thresholds

e. Cell neighborhood diversity across different pixel radius thresholds

f. Mixing scores across different pixel radius thresholds

g. ECM pixel clustering heatmap

h. ECM pixel cluster overlays

i. Representative images of long vs short fibers. Scale bars: 100um

j. Representative images of aligned vs unaligned fibers

Box plot: Lower bound is 1st quartile, center is median, upper bound is 3rd quartile, whiskers extend to 1.5*IQR beyond bound.

31

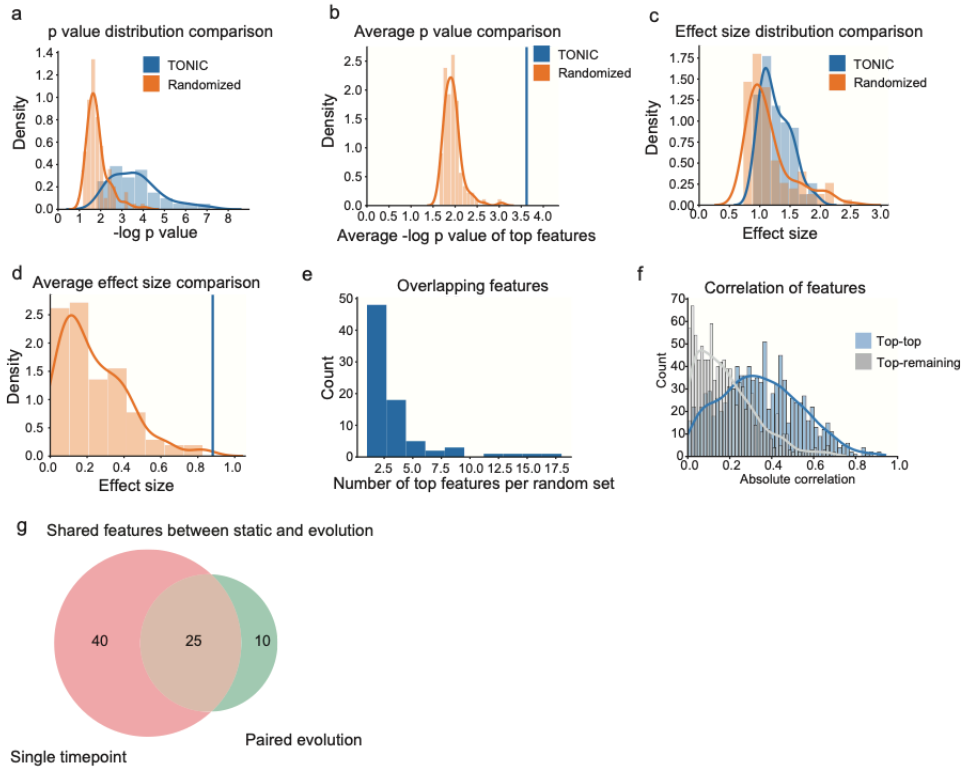

Figure 6. Univariate outcomes pipeline robustness

a. Distribution of p-values from the top features 100, as well as distribution of p-values from the top 100 features recalculated after the labels have been randomized

b. Average p-value in the top features, as well as distribution of average p-values across 100 different randomizations

c. Distribution of effect sizes from the top 100 features, as well as distribution of effect sizes from top 100 features recalculated after the labels have been randomized

d. Average effect size in the top features, as well as distribution of average effect sizes across 100 different randomizations

e. Number of times the same feature come up in different randomizations

f. Correlation between top features, compared with correlation between top features and other features

g. Overlap in features identified when examining each timepoint independently, compared to looking at the change in each feature across timepoints

32

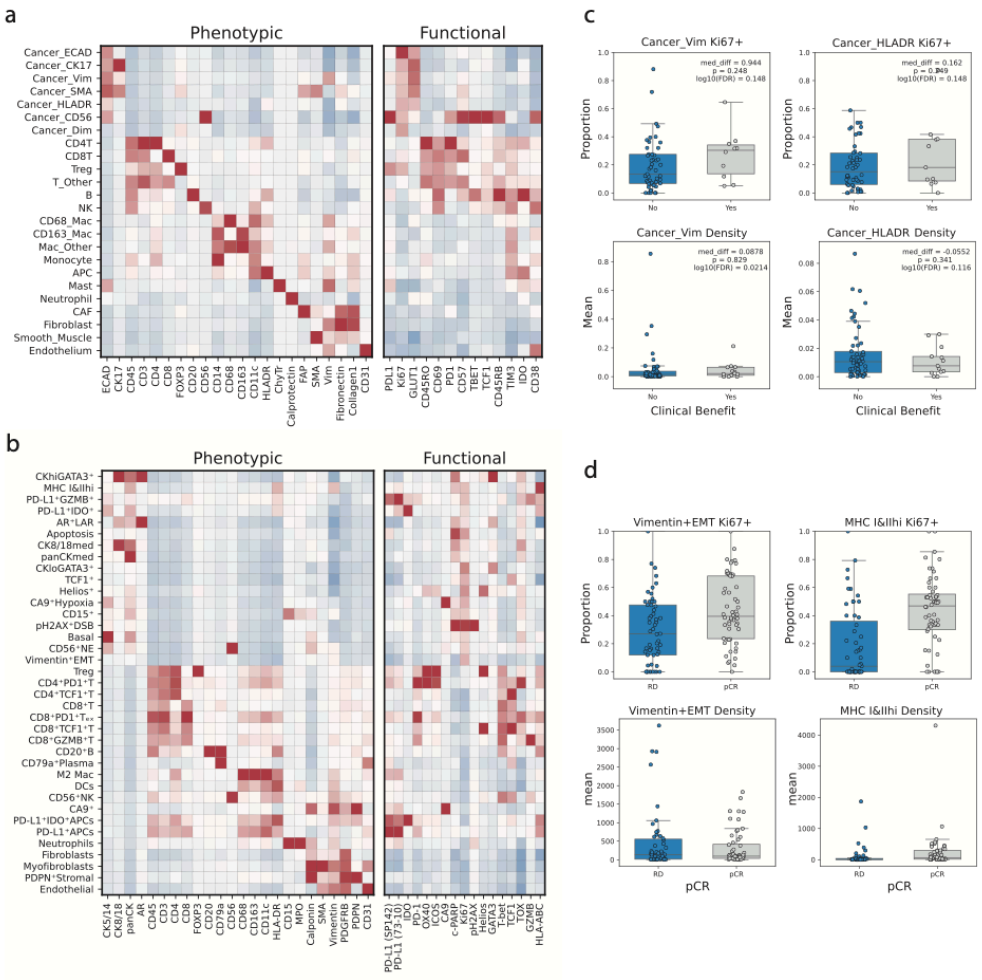

Figure 7. Matching cell types with Wang et al.

a. Heatmap with all cell clusters from TONIC cohort

b. Heatmap with all cell clusters from Wang et al

c. Outcome associations in TONIC cohort of cancer clusters identified as predictive in Wang et al, p-value generated from an unpaired two-sided t-test

d. Outcome associations in Wang et al cohort of cancer clusters identified as predictive in Wang et al

Box plot: Lower bound is 1st quartile, center is median, upper bound is 3rd quartile, whiskers extend to 1.5*IQR beyond bound.

33

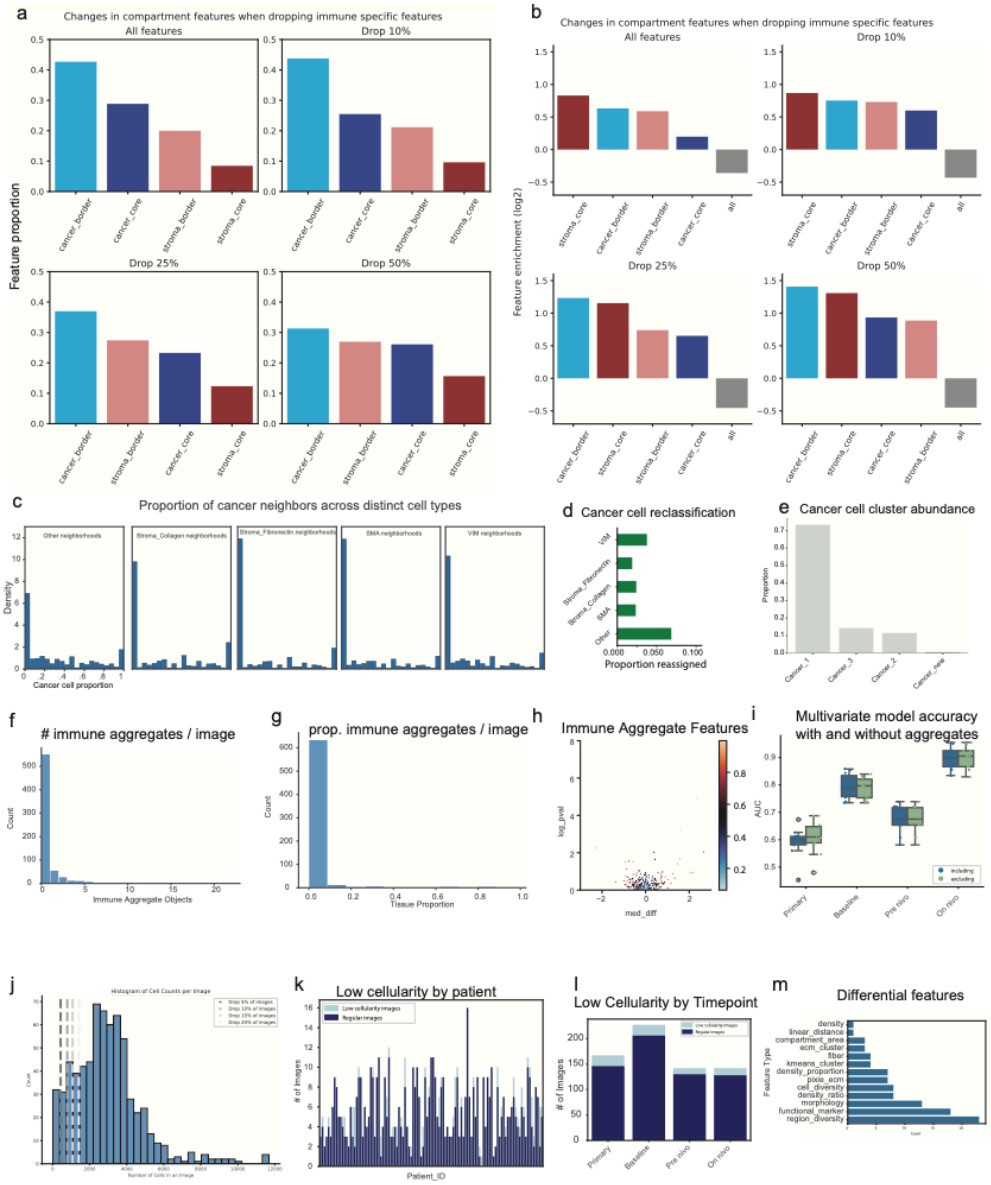

Figure 8. Evaluation of potential analytical confounders

a. Enrichment of compartment features in total feature count as the number of immune related features is removed

b. Enrichment of compartment features associated with outcome as the number of immune related features is removed

c. Proportion of Cancer cells in the neighborhood surrounding each cell type

d. Proportion of each cell type that meet criteria for being reassigned as cancer cells

e. Number of cells in each cancer subtype

f. Number of immune aggregates per image
g. Proportion of image composed of immune aggregates
h. Volcano plot showing significance (unpaired two sided t-test with equal variance, y axis) and effect size (difference in medians, x-axis) derived from immune aggregates
i. Predictive models with immune aggregate features included
j. Density of individual images
k. Proportion of low cellularity images across patients
l. Proportion of low cellularity images across timepoints
m. Top features that differ between low-cellularity images and all other images

Box plot: Lower bound is 1st quartile, center is median, upper bound is 3rd quartile, whiskers extend to 1.5∗IQR beyond bound.