High-Plex Multiomic Analysis in FFPE Tissue at Single-Cellular and Subcellular Resolution by Spatial Molecular Imaging

Shanshan He[1], Ruchir Bhatt[1], Brian Birditt[1], Carl Brown[1], Emily Brown[1], Kan Chantranuvatana[1], Patrick Danaher[1], Dwayne Dunaway[1], Brian Filanoski[1], Ryan G. Garrison[1], Gary Geiss[1], Mark T. Gregory[1], Margaret L. Hoang[1], Emily E. Killingbeck[1], Tae Kyung Kim[1], Youngmi Kim[1], Mithra Korukonda[1], Alecksandr Kutchma[1], Erica Lee[1], Zachary R. Lewis[1], Yan Liang[1], Jeffrey S. Nelson[1], Giang Ong[1], Evan Perillo[1], Joseph Phan[1], Tien Phan-Everson[1], Erin Piazza[1], Tushar Rane[1], Zachary Reitz[1], Michael Rhodes[1], Alyssa Rosenbloom[1], David Ross[1], Hiromi Sato[1], Aster W. Wardhani[1], Corey Williams-Wietzikoski[1], Lidan Wu[1], Joseph M. Beechem[1*]

Author Affiliations

1 - NanoString® Technologies, Inc., Seattle, WA 98109, USA

***Corresponding author**

Joseph M. Beechem

NanoString Technologies Inc, Seattle, WA 98109, USA

jbeechem@nanostring.com

Abstract

Spatial Molecular Imager (SMI) is an automated microscope imaging system with microfluidic reagent cycling, for high-plex, spatial in-situ detection of multiomic targets (RNA and protein) on FFPE and other intact samples with subcellular resolution. The key attributes of the CosMx™ SMI platform (NanoString®, Seattle, WA) include: 1) high-plex and high-sensitivity imaging chemistry that works for both RNA and protein detection, 2) three-dimensional subcellular-resolution image analysis with a target localization accuracy of ~50 nm in the XY plane, 3) large Hamming-distance encoding scheme with low error rate (0.0092 false calls per cell per gene) and low background (~ 0.04 counts per cell per gene), 4) high-throughput (up to 1 million cells per sample, four samples per run), 5) antibody-based cell segmentation methods, and 6) compatibility with formalin-fixed, paraffin-embedded (FFPE) samples.

In this study, 980 RNAs and 80 proteins were measured at subcellular resolution in FFPE cultured cell pellets, as well as FFPE tissues from biobanked samples of non-small cell lung cancer (NSCLC) and breast cancer. Cross-platform analysis using 16 cancer cell lines validated high-correlation (R^2 ~0.77) and high sensitivity (~1.44 FPKM/TPM; roughly 1 to 2 copies of RNA per cell) when compared to RNA-seq. Real-world archived NSCLC FFPE tumor sections revealed greater than 94% cell detection efficiency for RNA, despite the low RNA quality QV200 20% to the medium quality 65%. The accuracy of protein expression measurements was independent of the level of multiplexing, as demonstrated by the linear behavior of nested multiplexing panels ($R^2 > 0.9$). At 980-plex RNA detection, data analysis allowed identification of over 18 distinct cell types, at least 10 unique tumor microenvironment neighborhoods, and over 100 pairwise ligand-receptor interactions. Data from 8 NSCLC samples comprising over 800,000 single cells and ~260 million transcripts are released into the public domain (www.nanostring.com) to allow for extended data analysis by the entire spatial biology research community.

Keywords

spatial molecular imaging, RNA and protein profiling, multiomic profiling, spatial single-cell analysis, subcellular resolution, 3D mapping, FFPE, fresh frozen, *in-situ* hybridization, cell atlas, cell-cell interaction, spatial biomarkers

Introduction

Understanding the spatial distribution of RNA and protein in tissues has the potential to greatly expand our knowledge of all aspects of life science research (1-6). Fluorescence in-situ hybridization (FISH) and immunohistochemistry (IHC) are the most traditional technologies to assess the spatial distribution of RNA and protein in fixed tissue samples (7-10). However, these technologies can only detect a handful of targets at a time. In addition, high-plex, high sensitivity multiomic detection (RNA and proteins simultaneously) in formalin-fixed, paraffin-embedded (FFPE) has not yet been accomplished. Current multiplex IHC (mIHC) techniques (11) involve long staining and assay times, are not easily automated, and often require extensive assay optimization, hindering their establishment as routine methods.

For high-throughput high-plex RNA profiling with single-cell resolution, the most commonly used technology to date is single-cell RNA-seq (scRNA-seq), such as Drop-seq, inDrop, and Chromium™ (10X Genomics) (12-15). These technologies require tissue dissociation and isolation of suspension of cells to be encapsulated into nanoliter droplets with hydrogel beads bearing barcoding DNA primers. Although these methods enable whole transcriptome RNA profiling at single-cell resolution, the lack of spatial information limits the information content of this approach to generating a very detailed “parts-list” of the tissue.

Recently, multiple spatially resolved technologies have been developed that attempt to maximize the number of markers observable at the same time. These methods can be classified into two categories: 1) Profilers that enable spatially resolved high-plex data, often at the whole transcriptome level, from small sub-regions of the tissue using next-generation sequencing as a readout. Profilers are not single-cell resolved and often have RNA capture reagents arranged in a pre-defined grid-like pattern on which tissue sections are mounted and 2) Imagers that provide true single-cell and subcellular resolution using in-situ reagents, but at a plex-level often much smaller than the whole transcriptome.

Profiling-based techniques include Spatial Transcriptomics (16), Slide-seqV2 (17), pixel-seq (18), dbit-seq (19), and Digital Spatial Profiling (DSP) (20). These techniques offer the ability to perform highly multiplexed profiling based on the final detection readout using next-generation sequencing. These techniques are limited, however, by not being resolved at the single-cell level and often suffer areas of tissue “dead space”, where no measurements are performed. Although the active RNA capture regions (“spots”) can be precisely organized, placement of the tissue sample is random (onto the spots) with limited control in the selection of regions to be analyzed, ignoring the morphological information in the tissue. The RNA capture area spot size introduces analysis issues; for example, if the spot size is set large (50 µm) for assay, many cells are randomly selected by the spot location on tissue rather than directed by the morphology of the sample. If the spots are small (2 µm or less), the number of captured

transcripts is very low and the spots are grouped together for analysis, leading to the same issues as larger spots. In addition, many RNA profiling methods capture RNA molecules via the poly-A tail, which is inefficient for the subsequent sequencing process since poly-A transcripts are dominated by highly expressed genes. As spot-based methods rely on the mobility of RNA molecules in the tissue sample, they struggle to work with FFPE samples without requiring separate chemistries for different sample types. The in-situ solution offered by the GeoMx® Digital Spatial Profiling (DSP) overcomes many of these issues, offering a method for highly multiplexed spatial profiling of RNA and proteins on FFPE and frozen samples (20, 21).

Single-cell high-plex imagers use sequential cycles of probe hybridization and imaging and offer the potential to combine the benefits of single-cell RNA-seq analysis with added spatial resolution at single-cell or even subcellular resolution. Recent single-cell high-plex imaging technologies include MERFISH (22, 23), Molecular Cartography™ (24), FISSEQ (25), and seqFISH+ (26). Despite isolated proof-of-concept experiments demonstrating RNA plexing levels in the thousands (26), most molecular imaging systems routinely plex from 100 (24) to 500 (22) targets. Also, these methods are all optimized for fresh-frozen samples with very low efficiencies on FFPE tissue samples.

For protein detection, CODEX and InSituPlex methodologies increase target plex by labeling antibodies with unique oligonucleotide tags that are identified with fluorescence detection during amplification of the probe sequence or through cyclical hybridization of fluorescently labeled reverse-complement sequences. Although cyclic IF and ion-beam imaging have enabled high-plex imaging of dozens of proteins *in situ,* both approaches are limited in the maximal plex number due to on-instrument time or availability of metal isotopes.

Here, we describe the chemistry and applications of a Spatial Molecular Imager (SMI) to address many of the current unmet needs in spatial high-plex profiling of RNA and protein expression. SMI is a completely automated and integrated platform comprising chemistry, hardware, and software that enables highly sensitive, multiplex spatial profiling of RNA and proteins in FFPE tissues at single-cell and subcellular resolution.

1. Overview of Spatial Molecular Imager (SMI) Chemistry and Workflow.

Spatial molecular imager (SMI) is developed to perform multiple nucleic-acid hybridization cycles of fluorescent molecular barcodes to enable in-situ measurement of RNA and proteins on intact biological samples at subcellular resolution. The SMI chemistry is an enzyme-free, amplification-free, hybridization-based single-molecule barcode sequencing methodology, which can be applied directly to intact FFPE and fresh-frozen tissues on standard glass slides for pathology.

SMI chemistry relies on in-situ hybridization probes (ISH probes) and fluorescent readout probes (reporters) (Figure 1A) to detect RNA in the intact tissue. The ISH probes consist of a DNA of length 35-to-50 bases that will hybridize with the RNA target in the tissue, coupled with a stretch of 60-to-80 bases of DNA readout domain (for RNA identification). The readout domain consists of four consecutive 10-to-20 nucleotide sequences that allow four individual SMI imaging barcodes (reporters) to bind sequentially. For each gene, up to five RNA-detection oligonucleotide probes ("tiles") of ISH probes were designed to detect different regions of the RNA target, but each tile can independently record RNA-location identity.

Each tile of the ISH probe has its unique sequence in the target binding domain but all tiles share the same sequence in the readout domain. This design enables the highest detection sensitivity in FFPE tissue in which RNA targets may be highly fragmented. For instance, even if only one of the tiles successfully binds to the target, the RNA target can be detected at the readout step. Each reporter construct contains the controlled number of 15-to-60 dyes (dependent on desired sensitivity) assembled with fluorophore-conjugated oligos with photocleavable (PC)-linkers. All reporters are single-color containing one of the four fluorophores: Alexa Fluor-488, ATTO 532, Dyomics-605, or Alexa Fluor-647. The key advantages of SMI reporter chemistry include high signal-to-noise ratio (SNR) detection over the background for accurate spot calling and fast fluorescent signal quenching by ultraviolet (UV) cleavage of PC-linkers (Figure 1B).

SMI utilizes standard sample preparation methods performed for FISH on FFPE tissue sections to expose RNA targets, followed by the introduction of fluorescent bead-based fiducials that are fixed to the tissue to provide an optical reference for cyclic image acquisition and registration. Following hybridization of ISH probes, slides are washed, assembled into a flow cell, and placed within a fluidic manifold on the SMI instrument for RNA readout and morphological imaging. In SMI high-plex RNA assay readout, the tissue is hybridized with 16 sets of fluorescent reporters sequentially; each reporter set contains four single-color reporter pools. Reporters specifically bind to ISH probes during the 16 rounds of reporter hybridization according to the barcode assigned to each gene (Figure S1). After the incubation of each set of reporters, high-resolution Z-stacked images are acquired for downstream analysis. Prior to the incubation with the next set of reporters, PC-linkers are cleaved by UV illumination and fluorophores released from reporters are removed by washing (Figure 1B).

SMI encoding scheme is designed to assign a unique barcode to each target transcript from a set of 64-bit barcodes (4 color reporters in each readout round over 16 readout rounds) with Hamming distance 4 (HD4) and Hamming weight 4 (HW4). Every barcode is separated by a Hamming distance of at least 4 from all other barcodes to maximally suppress RNA decoding error. Every barcode has a constant Hamming weight of 4, in which each target is "on" in 4 rounds and "off" in 12 rounds during the 16 rounds of reporter hybridization. This "on" and "off" signal

barcode design allows for continued expansion to even higher plex, since only a subset of RNAs is "on" in any given cycle. For each reporter hybridization round, a single reporter can bind to one of the reporter-landing domains on the ISH probe of the gene. (Figure 1 and Figure S1). The 64-bit encoding scheme with the HD4 and HW4 yields 1,210 barcodes, from which a subset of 980 barcodes is selected to detect 960 target genes and 20 negative probe controls (Table S1). Negative control probes are modeled after synthetic sequences from the External RNA Controls Consortium (ERCC) set (22). Since they target alien sequences which are non-existent in human tissue, negative control probes serve as non-target controls for quantification of non-specific ISH probe hybridization. Up to five RNA-detection oligonucleotide probes/tiles are designed per gene and negative control. The colors of these barcodes are randomly assigned to targets to avoid any color-code induced bias. The remaining 230 barcodes were left as blank controls for misidentification quantification of reporter readout (Table S2).

A key component of SMI platform is the integrated and fully automated fluidic and imaging capability. SMI features a large scan area (range 16 mm^2 to 375 mm^2) on each tissue slide and supports up to 4 slides simultaneously. The run-time per sample is dependent on the number of flow cells utilized, the area imaged per sample, and the net-plexity of the assay. The on-instrument time for the commercial-release SMI instrument will range from 4 samples per day for ~16 mm^2 tissue area to ~0.5 per day for 100 mm^2 samples. The ability to measure four slides per single SMI run, with each slide allowing multiple samples to be placed onto a single slide (375 mm^2 active imaging area), allows for extreme flexibility in how to maximize the throughput of the SMI system.

The RNA and protein SMI imaging barcodes camera "spots" are located (in 1 of the 4 color channels) and fit in the individual image using a 2D parabola. Note that the nature of these imaging signals represents a "deterministic super-resolution imaging reagent signal", and hence each SMI imaging barcode can be located well below the diffraction limit of the imaging system. Images from different cycles and different colors are registered using an affine transform to align the fluorescent bead reagents added to each tissue slide. The positions of the imaging spots that contribute to an individual RNA target call contribute together to generate a very rich super-resolution data matrix to generate the final X, Y, and Z coordinates of the analyte. The localization accuracy for analyte detection is within a radius of ~50 nm in the XY plane, even when averaged over the timescale for the entire experiment (Figure S2). Z-axis localization is dependent upon the number of optical Z-stacks taken during data collection and was not a key component of this study.

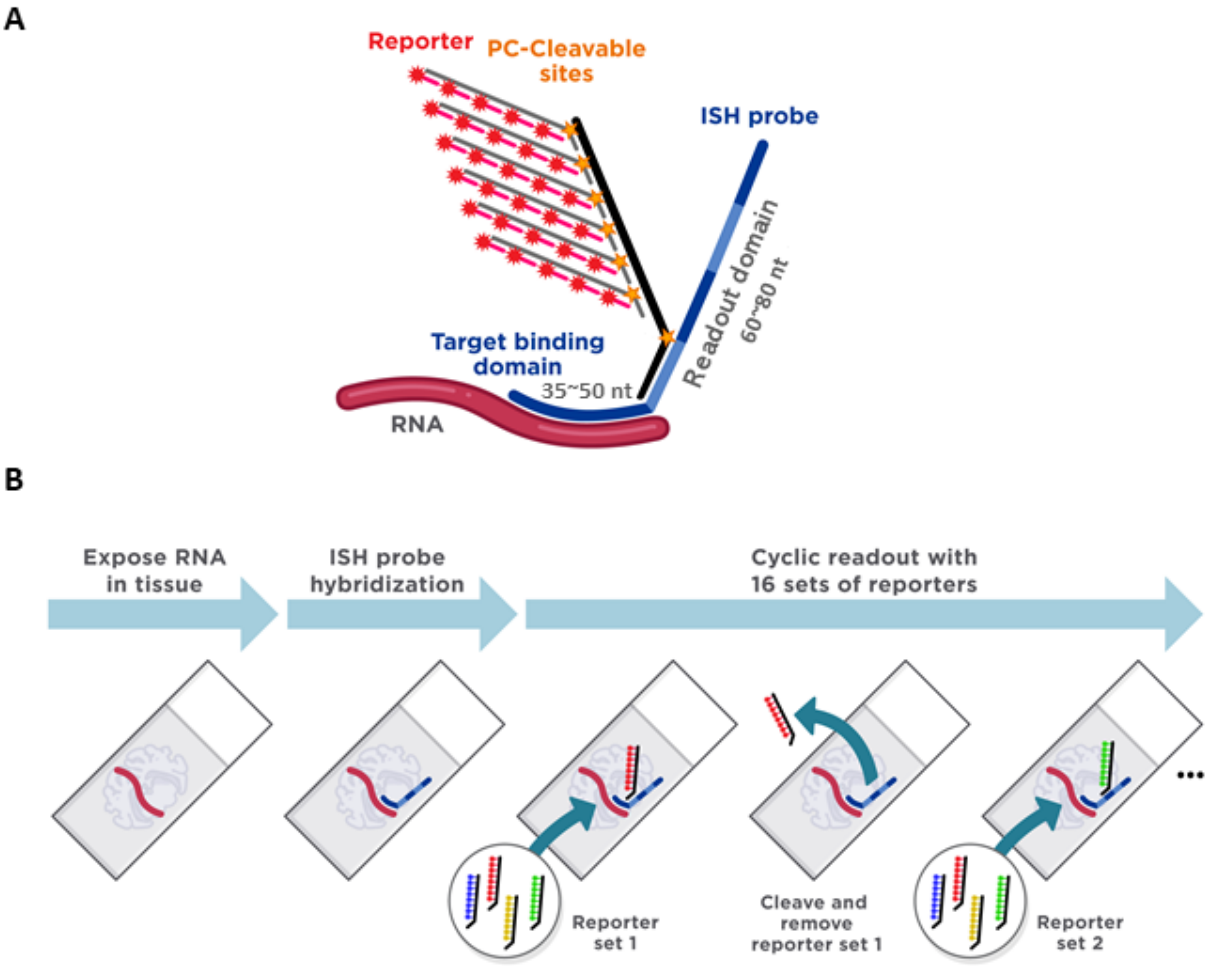


Figure 1. Overview of CosMx SMI Chemistry and Workflow. A. Schematic description of SMI ISH probe and reporter design. ISH probe consists of the target-binding domain and readout domain. The target-binding domain is a 35-to-50 nt DNA sequence that hybridizes with target RNA. The readout domain is a 60-to-80 nt DNA sequence that contains 4 consecutive 10-to-20 nt reporter-landing domains, where each landing domain can be hybridized with a unique reporter. With a 64-bit SMI barcoding design, there are a total of 64 unique reporter-landing sequences. Each reporter is a single-color branched structure assembled with oligos that are conjugated with one of the four fluorophores and will be detected as one of four colors (blue, green, yellow, or red) in SMI images. Each reporter has a controlled number of 15-to-60 dyes with 6 PC-cleavable sites where the fluorescence signals are efficiently quenched by UV cleavage and wash prior to each cyclic reporter readout. **B.** Illustration of SMI RNA assay workflow. FFPE slide undergoes standard tissue preparation to expose RNA targets for ISH probe hybridization. The sample is assembled into a flow cell and loaded onto an SMI instrument for cyclic readout with 16 sets of reporters. As each reporter set contains 4 reporters with 4 different fluorophores, a total of 64 unique reporters are used in SMI assay to bind to the different reporter landing domains on ISH probes. Following each set of reporter hybridization, high-resolution Z-stacked images are acquired, followed by cleavage and removal of fluorophores from the reporters prior to the incubation with the next set of reporters.

2. Three-Dimensional Mapping of RNA and Cell Segmentation Enables Accurate Molecular Profiling at Single-Cell and Subcellular Resolution.

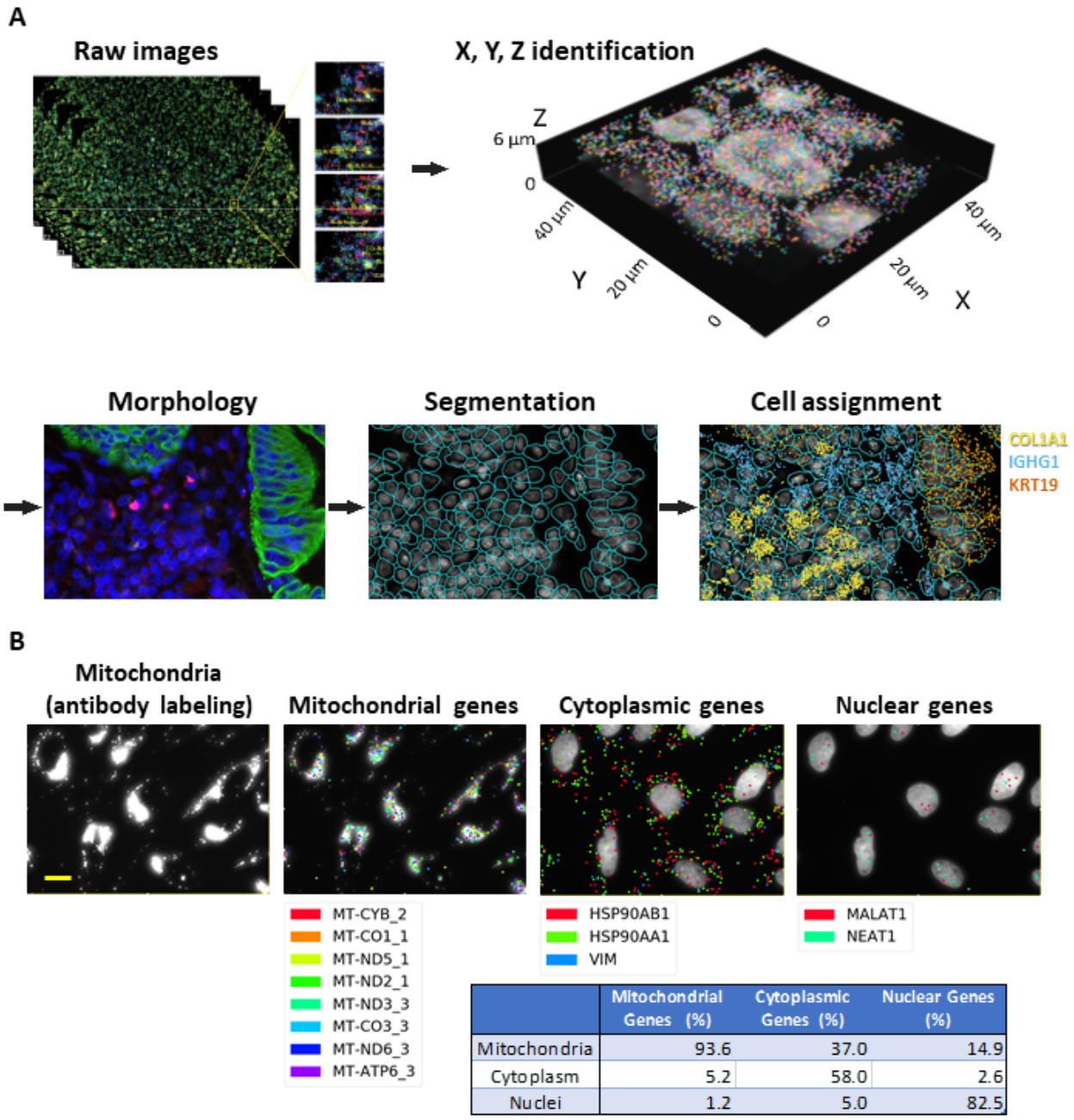
The combination of SMI chemistry, hardware, and software enables high-plex spatial RNA profiling within single cells with very high sensitivity. For spatial profiling, the data of fluorescent signals in Z-stacked raw images were transformed into decoded RNA transcript information at their registered spatial location in three dimensions (3D) and assigned to segmented single cells (Figure 2A). During the primary image processing, the fluorescence signals in raw images were transformed into digital output comprising of detected "spots" localized in X-, Y-, and Z-dimensions for subsequent RNA decoding. A spot is identified in an image as an isolated fluorescence signal with the intensity much higher than its neighbors. In biological tissues, however, dense fluorescent spots can be detected at a small spatial region, which may limit the accuracy of resolving each spot. To solve this potential issue, we developed custom image analysis algorithms to process 3D multi-channel image stacks obtained in each field of view (FOV). The key objective of this analytical method is to reduce the multi-dimensional image stack to a single list of individual reporter binding events. This process was performed across all FOVs, concurrently with image acquisition during cyclic reporter readout. All spots pertaining to a given FOV were collated into a single list of XYZ locations of all individual reporter binding events, which was used in the next step to decode the gene-specific barcodes formed by these reporter binding events.

The decoding algorithm enables the detection of as many transcripts as possible in crowded reporter binding events while limiting the rate of error calls that can be generated due to the presence of multiple transcripts in close vicinity. In this algorithm, each unique XYZ location with at least one reporter binding event is considered a 'seed' and used to construct a neighborhood in which gene-specific barcodes were searched. Two passes through of the data with the seed-centered neighborhood search are used to obtain a list of potential transcripts with their spatial locations. In the first pass-through of the data, the neighborhood search was limited to a radius of 0.5 pixels (90 nm). Because every gene barcode has 4 "on" spots, any seeds with fewer than 4 unique reporter-probe binding events in the neighborhood were removed from being considered for transcript decoding due to their inability to form a complete gene-specific barcode. All possible four reporter combinations of unique reporter probes in a seed's neighborhood such that at least one of the four reporter binding events was present at the seed location were then matched with gene-specific barcodes to detect the presence of a gene in a seed's neighborhood. If more than one gene was detected in a seed's neighborhood, the seed and all the transcripts detected in its neighborhood were dropped from further analysis.

Any reporter binding events used in making transcript calls were removed from the dataset, and the process was repeated with a slightly increased search radius of 0.75 pixels to try to recover any potential transcript calls that may have been lost by local tissue motion. The transcript calls generated by these two passes through the data were further filtered to remove any potential duplicate calls or calls in neighborhoods with a high probability of making a transcript call by random chance. These various filtering steps are crucial to address potentially elevated error-call rates and duplicate calls of individual transcripts, ensuring that at the completion of the transcript decoding process, we only retain high-confidence transcript calls while maximizing the number of decoded transcripts.

Following RNA decoding and 3D spatial registration of target signals, the cell segmentation process was performed to define cell boundaries based on tissue morphology antibody markers (Figure 2A). In our cell segmentation process, tissue was stained with a nuclear dye (DAPI) and labeled with morphology markers, such as antibodies for membrane (CD298), epithelial cells (pan-cytokeratin [PanCK]), and T cells (CD3). Defining accurate cell boundaries by segmentation is critical to data quality because it affects the spatial assignment of transcripts to specific cells. It is challenging to achieve high accuracy on tissue sections in which cells are tightly packed, having shared 3D boundaries, or unevenly stained with morphology markers. To address this issue, we developed a cell segmentation pipeline that combines image preprocessing with machine learning for better accuracy (Figure S3 and S4). The signals of membrane channels were combined and normalized to the range of the nuclear channel. Then, image subtraction was performed between the nuclear channel and the normalized membrane image. After these preprocessing steps, the images were fed into pre-trained Cellpose neural network models (18).

In the registered image, each transcript location was mapped to the corresponding cell where the transcript was located as well as within the cellular compartment (*e.g.,* nuclei, cytoplasm, membrane). To demonstrate accurate RNA detection at the subcellular resolution, we designed a set of SMI probes targeting mitochondrial genes, nuclear genes, and cytoplasmic genes. The cell compartments in U2OS cells were labeled with fluorescent probe-conjugated antibodies for mitochondria and CD298 as well as stained with DAPI for nuclei (Figure 2B). Our RNA profiling data shows that 93.6% of detected mitochondrial genes specifically colocalized with the antibody-based mitochondria segment, 82.5% of nuclear genes strictly located within nuclei, and 58.0% cytoplasmic genes were localized surrounding the nuclei. Given the unique spatial structure of mitochondria, which is reticulated throughout the cytoplasm, this type of co-localization between cytoplasmic and mitochondria compartments is expected. Thus, these results indicate high accuracy in RNA detection and its spatial assignment at subcellular resolution. Spatial molecular profiling using this cell segmentation pipeline was also performed on tissue sections of the FFPE human lung (Figure 2A).



	Mitochondrial Genes (%)	Cytoplasmic Genes (%)	Nuclear Genes (%)
Mitochondria	93.6	37.0	14.9
Cytoplasm	5.2	58.0	2.6
Nuclei	1.2	5.0	82.5

Figure 2. RNA Profiling at Single-Cell Subcellular Resolution Using Three-Dimensional Mapping of RNA and Morphology-Based Cell Segmentation. A. Analysis workflow to transform raw images to decoded RNA transcripts at single-cell subcellular resolution. The workflow includes: 1) three-dimensional primary image processing to identify and register reporter spots, 2) decoding of reporter spots to RNA transcripts with registered X, Y, Z spatial location, 3) outlining of nuclei and cell boundaries with DAPI and antibodies after cyclic reporter readout for morphology-based cell segmentation, and 4) assigning RNA transcripts to

single cells. As an example of the final output of the analysis workflow, three identified genes (COL1A1 [yellow], IGHG1 [cyan], and KRT19 [orange]) in FFPE human lung tissue were overlayed with the segmented cells based on their registered spatial information. **B.** Demonstration of subcellular resolution in U2OS cells. A 980-plex panel including transcripts that are located in mitochondria, nuclei, and the cytoplasm. These genes were read out on fresh fixed U2OS cells with SMI RNA assay, followed by labeling with morphology antibodies for CD298 (blue) and mitochondria (red), as well as DAPI stain for nuclei (magenta). The table shows the quantification of transcripts detected in these three cell compartments.

3. SMI 980-plex RNA Panel and Cross-Platform Validation.

We designed a panel targeting 960 genes to investigate the biology of single cells or subcellular compartments. Among those genes, 211 genes were selected to assess the cell type, and the remaining genes were designed to capture critical cell states, cell-cell interactions, and hormone activity (Figure S5, list of genes in Table S1). In addition, 20 ERCC negative probes containing hybridization regions that are not complementary to human genome or transcriptome are included in the panel (27). The panel design and gene selection process are fully described in the Materials and Methods.

To ensure that this panel was capable of quantitating RNAs in a wide range of biological samples, we profiled the SMI 980-plex panel on 16 different cell lines, spanning a broad range of biology and expression patterns. We compared SMI RNA counts to bulk RNA-seq FPKM values (fragments per kilobase per million mapped reads) published on the Cancer Cell Line Encyclopedia (CCLE) (28) and TPM (Transcripts Per Kilobase Million) published as part of the NCI-60 Human Tumor Cell Lines Screen (29, 30). Segmented regression was fit predicting raw SMI counts from RNA-seq (Figure 3A). For the average cell line, the breakpoint in the segmented regression occurred at an RNA-seq value ranging from 0.36 to 5.2 (median: 1.443) (Figure 3B). Above this breakpoint, SMI counts linearly tracked RNA-seq counts with high correlation (Figure 3C). The results show adequate concordance between SMI counts and bulk RNA-seq values, indicating that the panel can measure a wide range of biological processes (Figure 3C).

Using SMI image processing methodology, thousands of transcripts per cell were detected in the tested FFPE cell pellets with a maximum of 4,500 transcripts resolved in a single cell. SMI RNA profiling also features a cell drop-out rate of less than 3%. Background signals for the 20 ERCC control targets were extremely low (~ 0.04 counts per gene per cell). Our single-cell distribution analysis revealed that even low expressers were detected well above the background (Figure S6), confirming the high sensitivity of SMI RNA assay.

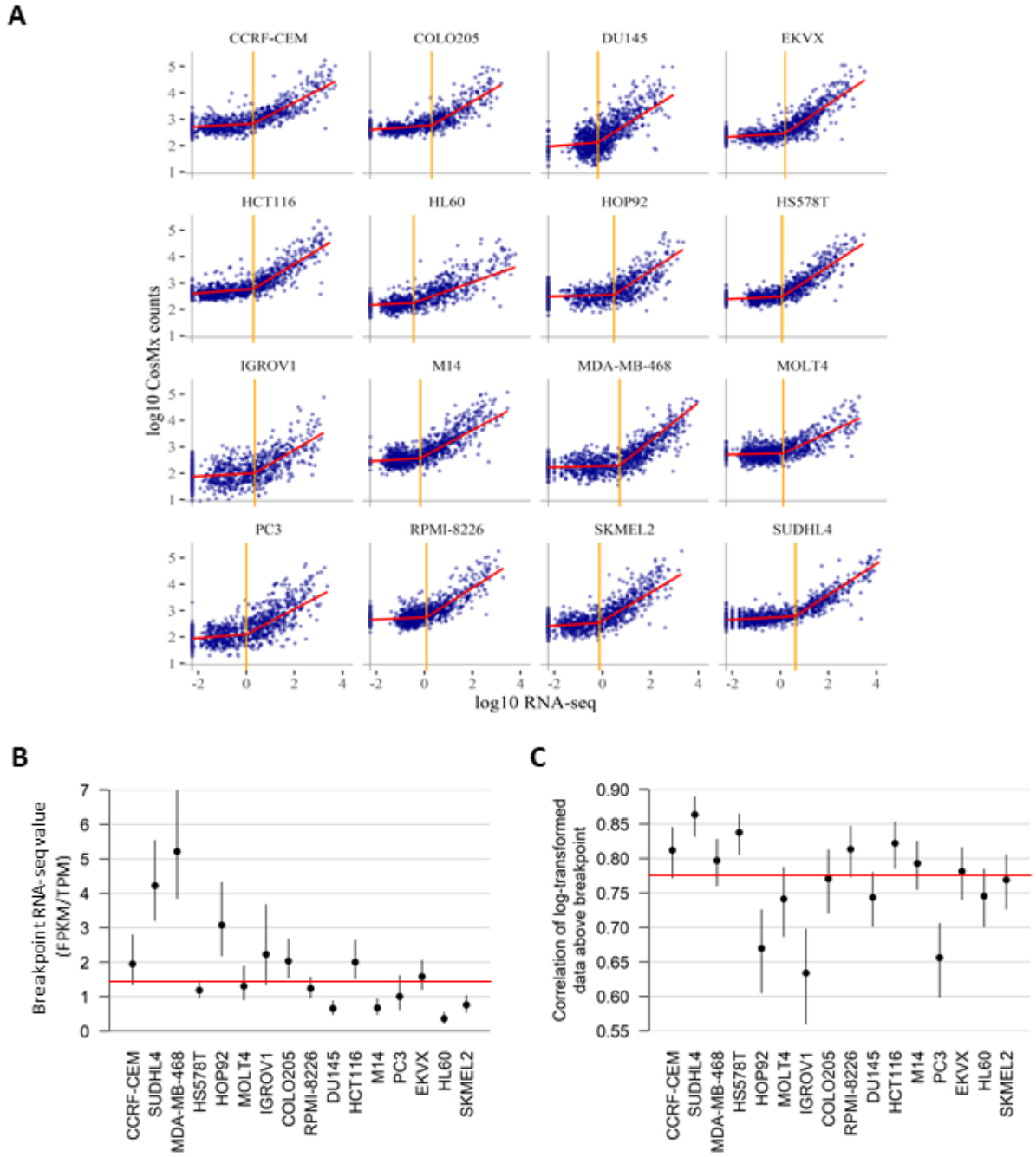


Figure 3. Comparison of SMI Data to RNA-seq Data Demonstrates Concordance Between the Two Platforms. A. For 16 cell lines, RNA-seq values plotted against raw SMI counts for 16 cell lines. RNA-seq values are in TPM for cell lines from the NCI-60, and in FPKM for cell lines from the CCLE. Orange lines show breakpoints from segmented regression. Red lines show segmented regression fit. **B.** Estimated breakpoints from each cell line in FPKM or TPM depending on cell line origin. The red line shows the median value. **C.** The correlation between log-transformed RNA-seq and SMI data for genes with RNA-seq values above the breakpoint. The red line shows the median value.

4. High-Plex Spatial RNA Detection for Cell Typing and Cell Interaction in FFPE NSCLC Tissue.

In order to understand how SMI performs on "real-world" FFPE samples, the 980-plex assay was run on 8 FFPE slides from five non-small cell lung cancer (NSCLC) tumors (Table 1A, Table S3). All samples were from tumor tissues archived in 2017 and 2018, the typical age for many archived cancer samples available to researchers (Table S4). The RNA in these samples ranged from "too degraded to sequence" to "medium quality to sequence", using either RIN index or DV200 scoring (Table 1B, Figure S7). These data, therefore, represent a range of sample types, including the most difficult class of archived material any researcher is likely to examine.

Measurements of these 8 slides represent a total of 800,327 cells, of which nearly 96% of cells passed quality control, yielding 769,114 analyzable cells (Table 1A). An average of 260 transcripts per cell were detected, and 850 of the 960 potential genes in this panel were measured (Table 1A). A first step in the analysis of these data sets is to identify the high-level cell types of as many of the spatially resolved cells as possible. Single-cell expression profiles were derived by counting the transcripts of each gene that fell within the area assigned to a cell by the segmentation algorithm. Less than 4% of cells contained fewer than 20 total transcripts, which were omitted from further analysis as this class of cells behaves anomalously when projected into lower dimensions (*e.g.*, UMAP). A normalized expression profile was defined for each cell by dividing its raw counts vector by its total counts. A separate UMAP projection was computed for each tissue.

Cell type was determined by comparing individual cells' expression profiles to reference profiles for different cell types (scRNA-seq and bulk RNA-seq of flow-sorted blood and stroma databases) (31), assigning each cell to the cell type under which its profile was most likely. The likelihood was defined using a negative binomial distribution, with mean defined by a cell type's reference profile plus expected background, and with a size parameter set at 10 to allow for extensive overdispersion. Reference profiles of immune and stromal cell types were taken from previous work (31). New reference profiles were defined with the mean expression profiles of $PanCK^+$ clusters in the UMAP projection. These new clusters were all $PanCK^+$ and included five tumor-specific clusters and one cluster shared across all tumors. The tumor-specific clusters were labeled "tumor" and the shared cluster was labeled "epithelial", and this interpretation was confirmed by a pathologist review.

The matrix of single-cell gene expression profiles was analyzed using UMAP, and each cell was assigned to a cell type as described above (Figure 4A). It should be emphasized that for classic single-cell studies, this UMAP-Cell-Type combination plot represents the essential information content of the experiment. However, for spatially resolved studies, this basic data type is the very beginning of a rich spatial analysis, as each cell in this UMAP-Cell-ID representation now has high-resolution X, Y, and Z spatial coordinates associated with it. For

example, we observed that in these tumors, B-cells typically gathered in dense clusters accompanied by T cells (Figure 4B). Plasmablasts also gathered densely, often proximal to smaller numbers of T cells. Macrophages both gathered in small clusters and trafficked diffusely throughout tumors. Neutrophils were usually found filling large vacancies within tumors, accompanied by very few other immune cells (Figure 4B).

This spatial information also allows detailed analysis of cell "neighborhoods". We defined a "neighborhood matrix" encoding the number of each cell type among each cell's 200 closest neighbors (Figure 4C). Neighborhood matrices can be tailored to answer a broad range of different biological questions. For example, neighborhoods could be defined over smaller or larger distances, or a neighborhood matrix could encode average gene expression profiles or average gene expression profiles within specific cell populations.

Once a neighborhood matrix was defined, it was subjected to traditional single-cell analyses. A UMAP projection of our neighborhood matrix shows the diverse microenvironment states within these tumors (Figure 4D). For example, it shows neighborhoods of almost pure tumor cells with very low levels of macrophage and T-cell infiltration. Other neighborhoods were dominated by single-cell types, including macrophages, neutrophils, plasmablasts, and myeloid dendritic cells (mDCs). Other neighborhoods held distinct mixtures of immune populations, such as B-cells with T-cells, macrophages with T-cells and plasmacytoid dendritic cells (pDCs), and macrophages with neutrophils and infrequent lymphoid cells. Some of these neighborhoods were specific to single tumors, while others were shared across tumors. Finally, by clustering this neighborhood matrix, we partitioned the tumor microenvironment into distinct niches. Plotting niches in physical space clarified the spatial organization within and the contrasts between these tumors (Figure 4E).

Studies across larger numbers of samples require sample-level summary statistics. Using the cell types derived from the gene expression matrix, we found these tissues to differ in the relative abundances of the immune cell population within each tumor (Figure S8A). Using the niches derived from the neighborhood matrix, we identified additional differences in the multi-cellular niches comprising their microenvironments (Figure S8B). Niche abundances expand the information beyond what cell-type abundances alone can provide. For example, samples Lung 5 and Lung 6 have similar macrophage abundances (7% and 8% of cells, respectively); but only Lung 6 contains the macrophage-dominated niche (9% of cells).

To define more nuanced sample-level summaries, we scored each cell for the number of tumor cells among its 100 closest neighbors, a metric of how much it has invaded into the tumor instead of remaining confined to the stroma. Contrasting this invasiveness score across cell types and across tumors revealed differences within and between tumors (Figure S9). For example, in

Lung 6, macrophages were primarily surrounded by non-tumor cells, while neutrophils were more likely to be surrounded by tumor cells.

By contrasting the gene expression matrix and the neighborhood matrix, we examined further advanced questions for every gene, cell type, and neighborhood characteristic: "How does this cell type change expression in response to this neighborhood characteristic?" and "How does this dependency vary across tissues?" To answer these questions, we investigated the changes of gene expression in macrophages between niches in Lung 6 (Figure 4F). More than 43% of genes (415 of 960 genes) had expression changes between niches with false discovery rates below 0.05. The most statistically significant gene was SPP1 (global *p*-value of 5×10^{-61}). SPP1 has been shown to mediate macrophage polarization and up-regulate PD-L1 expression (21). Plotting macrophage SPP1 expression across the physical space of Lung 6 demonstrated two clear subpopulations of macrophages (Figure 4G). SPP1-high macrophages dominate the tumor interior and the upper half of the stroma, while SPP1-negative macrophages dominate the lower half of the stroma and the long thrust of diverse immune cells cutting deep into the tumor. Spatial expression analysis of SPP1 and HLA-DQA1 transcripts together (Figure S10) revealed these genes to be expressed in mostly mutually exclusive regions of the tumor, suggesting an antigen-presentation role for the SPP1-negative macrophages. Plotting the density of macrophage SPP1 expression across tumors identified changes between tumors and between niches (Figure S11). For example, for every tumor, macrophages in the diverse "immune" niche had lower SPP1 expression than in almost any other niches.

With the spatial map of cell types in place, we turned to interrogate the interactions between tumor cells and T cells. First, we annotated 100 canonical ligand-receptor (LR) partners within the 980-plex panel (Figure 5A). Within our panel, many LR pairs relevant to the tumor immune interface can be found, including various immune checkpoints such as PD-L1/PD-1 and CTLA4/CD86. To understand how these interactions are changing across space and between samples, we devised a novel computational method to search for coordinated LR expression in neighboring cell types (Figure 5B). Using this method, we discovered 16 LR pairs that were enriched at the tumor-T cell interface in at least one of five lung tumors (Figure 5C). Many of these interactions were present in only a subset of the tumors. PD-L1/PD-1 (CD74/PDCD1) exhibited a higher interaction score across Lungs 5, 9, 12, and 13, but remained lower in Lung 6. Notably, HER2 (ERBB2) shows a similar profile across tumors. However, a member of the same receptor family, EGFR, maintained a higher interaction score in Lung 6 and decreased interaction in all the other lung tumors. This between-tumor variability in LR signaling strength is consistent with the known variability in tumor response to immune checkpoint inhibitors and EGFR inhibitors.

Table 1. SMI Sample Summary. Aggregated Data from the eight NSCLC samples.

A

Category	Sample Summary
Tissue type	FFPE human lung
Panel	980 plex
Number of slides analyzed	8
Total tissue area analyzed (μm^3)	753,480,217
Number of Field of Views (FOVs)	233
Total number of cells	800,327
% Cells passed QC (≥ 20 transcripts)	96.1
Number of cells analyzed	769,114
Transcripts detected	262,649,897
% of transcripts assigned to cells	79.2
Cellular transcripts/μm^3	0.446
Mean transcripts/cell	260
Mean negatives/cell/target	0.0429
Genes detected	850
% Genes detected	88.5
Mean false call/cell/target	0.0092

B

Sample	FFPE section thickness (µm)	RNA yield (ng)	RIN	DV200 (%)	RNA quality based on DV200*	Cells passed QC for SMI (%)
Lung 5	20	195	unmeasurable	21	too degraded	97.2
Lung 9	20	3,036	2.3	65	medium	94.1
Lung 12	20	2,029	2.4	64	medium	96.6
Lung 13	2x20	460	1.8	23	too degraded	98.1

**Evaluating RNA Quality from FFPE Samples* (32)

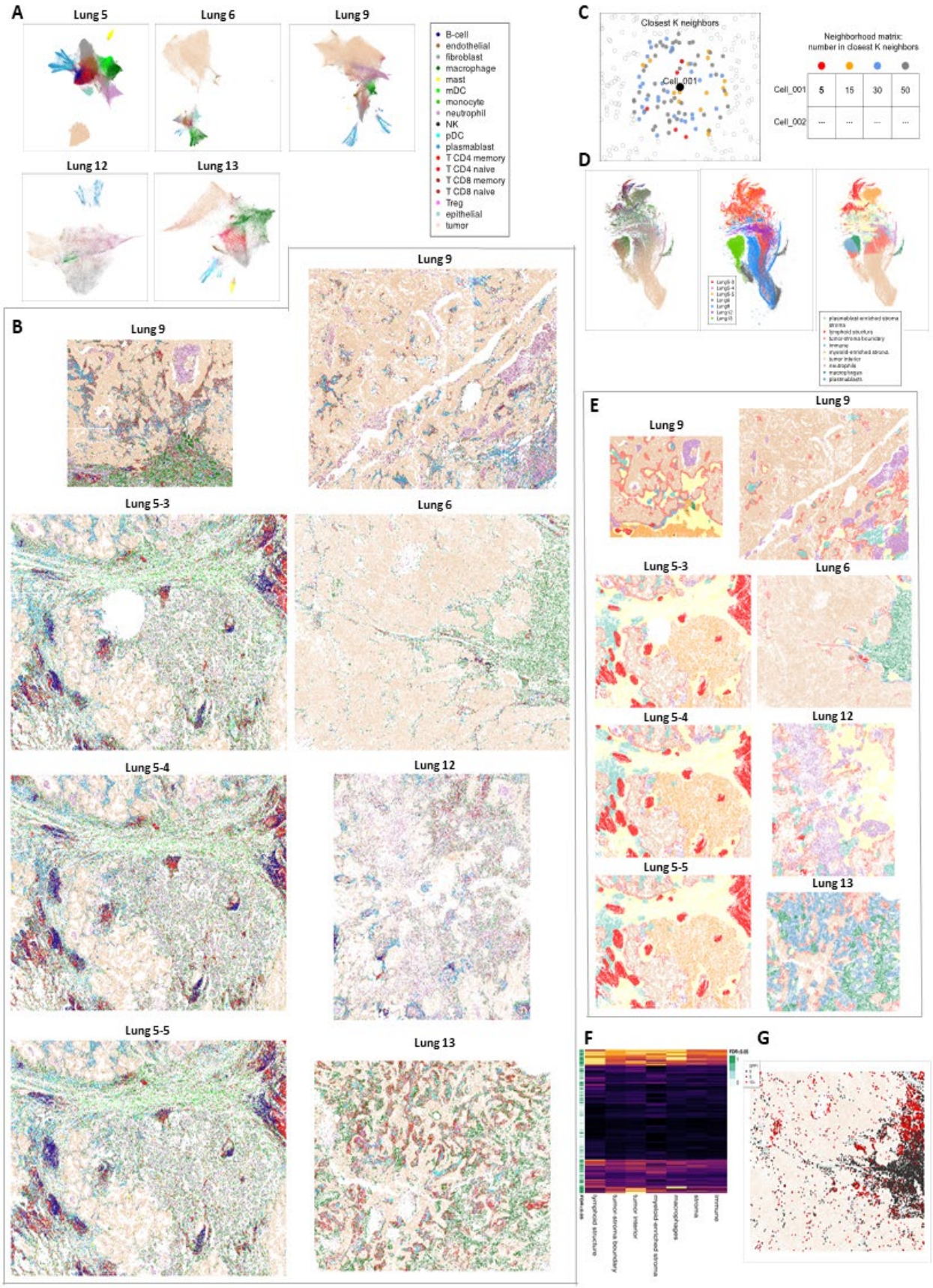


Figure 4. Spatial RNA Detection to Identify Cell Types and Cell-Cell Interactions in FFPE Human NSCLC Tissue. **A.** UMAP projections of each tissue based on the gene expression matrix. Color denotes cell type. **B.** Cells in physical locations (X, Y coordinates). Color denotes cell type. **C.** Definition of a neighborhood matrix. For each cell, the nearest neighbors are identified, and a summary of those neighbors is recorded. Here, the abundance of each cell type was taken. This operation is performed for all cells, defining a matrix of cells and neighborhood characteristics. **D.** UMAP projection and clustering of cells based on the neighborhood matrix. Left: colored by cell type. Center: colored by tissue. Right: colored by clustering of neighborhood data or "niche". **E.** Spatial arrangement of niches. For each cell, the frequency of each cell type among its 200 closest neighbors is recorded. Cells are then clustered into "niches" based on this data. Cells are shown in their physical locations and colored by their niche. **F.** Mean gene expression of macrophages in each niche in Lung 6. The green sidebar shows statistical significance from a global likelihood ratio test for uniform expression across all niches. **G.** Macrophage expression of SPP1 in Lung6. In this tissue, macrophages located inside the tumor express SPP1, as shown in the upper half of the dense macrophage region. In contrast, the macrophages located along the vasculature and lymphoid cells in the tumor rarely express SPP1. Macrophages are shown in bold points, SPP1 expression in macrophages is shown on a color scale of black to red.

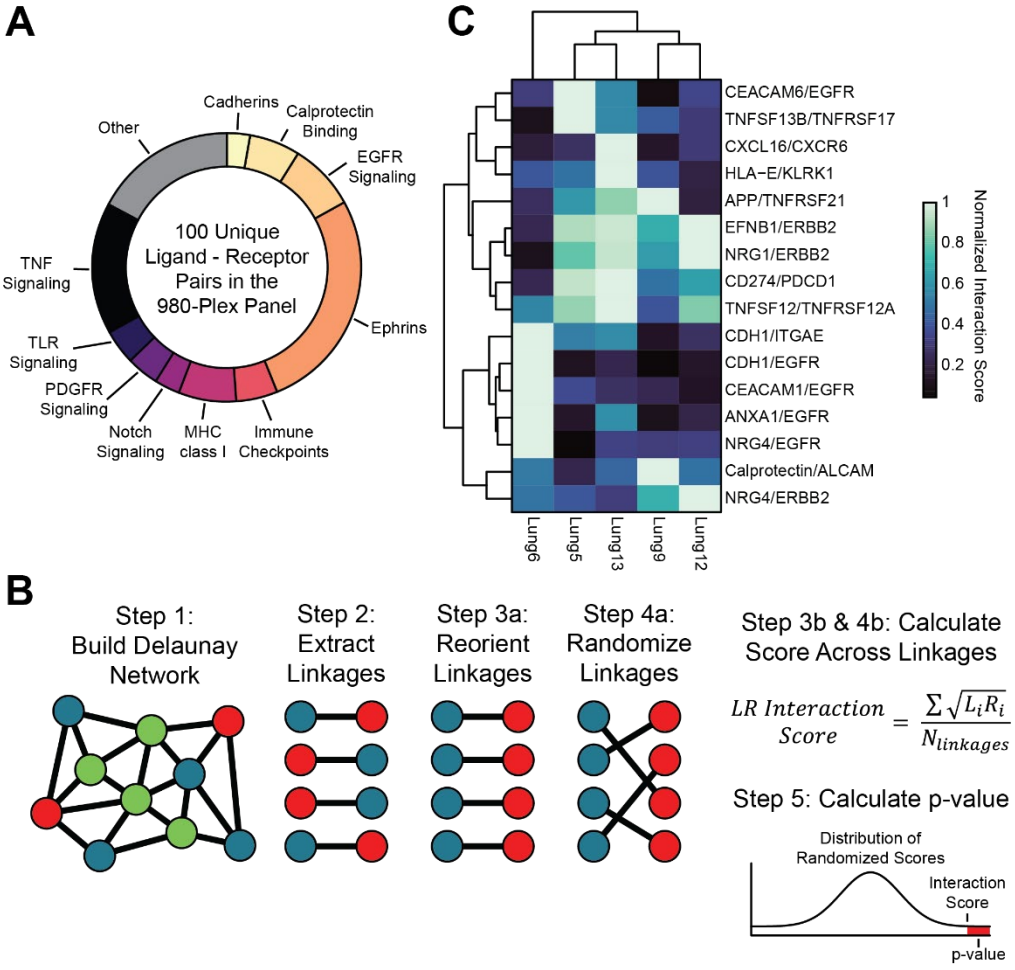


Figure 5. Paired Ligand-Receptor Expression Between Interacting Tumor and T Cells Varies Across Tumors. **A.** 100 unique ligand-receptor pairs are included in the SMI 980-plex panel. These interactions fall into functional categories depicted here in relative proportion. **B.** Ligand-receptor (LR) interactions are scored by first building a Delaunay network based on cellular spatial locations. The desired linkages are extracted from the network and used to calculate the LR interaction score, which measures pairwise LR expression between specified cell types. Linkages are reoriented to differentiate between ligand and receptor expressing cells. This score is compared to a distribution of scores produced using randomized linkages. Such a comparison can determine if the given configuration of interactions between specified cell types significantly enriches for pairwise LR expression. **C.** 16 ligand-receptor pairs exhibited spatial significance in at least one of five different lung cancer tumors. The interactions scores for these pairs scaled as each score has a maximum of 1 across all tumors.

5. Reproducibility Across Serial Sections of FFPE NSCLC Tumor.

To obtain a better understanding of the reproducibility of the SMI platform, three serial sections from a NSCLC tissue were profiled. Although these 3 serial sections would not contain the same individual cells, they had nearly identical tissue architecture (see Lung 5-3, Lung 5-4, Lung 5-5 in Figure 4), allowing comparisons of the same tissue region across slides. This experiment offers an opportunity to observe technical variability that tests all aspects of the SMI platform: independent tissue preparation, cyclic chemistry, imaging, primary-secondary-tertiary data processing, and analysis in solid tissue with minimal biological variability.

As a first step, we examined the total-slide integrated RNA expression profile from each section by acquiring the total counts of each gene across all cells. These whole-slide integrated bulk profiles, each from 94,977 to 105,903 cells, were highly concordant: the lowest correlation between the log-scale expression profile of any pair of the three replicates was 0.996 (Figure S12).

To demonstrate reproducibility on a smaller spatial scale than a whole section, we partitioned two replicate (rep) sections into 74 grid squares each (Figure 6A, top). For this analysis, we used slides Lung 5 rep 3 and Lung 5 rep 5, which were the most spatially well-aligned pair. Grid squares contained between 600 and 2,000 cells; six grid squares intersecting a tissue-hole in rep 3 were discarded (Figure 6B). The total expression profile of each grid square was acquired to produce a gene expression matrix of 960 genes across 74 grid squares in each replicate. The correlation of 960-gene expression profiles between matching grid squares was high: the average square had a correlation of 0.96 between the log-transformed expression profiles of two replicates, and 95% of all squares had correlations between 0.87 and 0.94 (Figure 6B). The lowest correlation occurred in a square where the biological structure in serial sections differed: in Lung 5 rep 3, the square contained a small part of a tertiary lymphoid structure, while in Lung 5 rep 5 it did not.

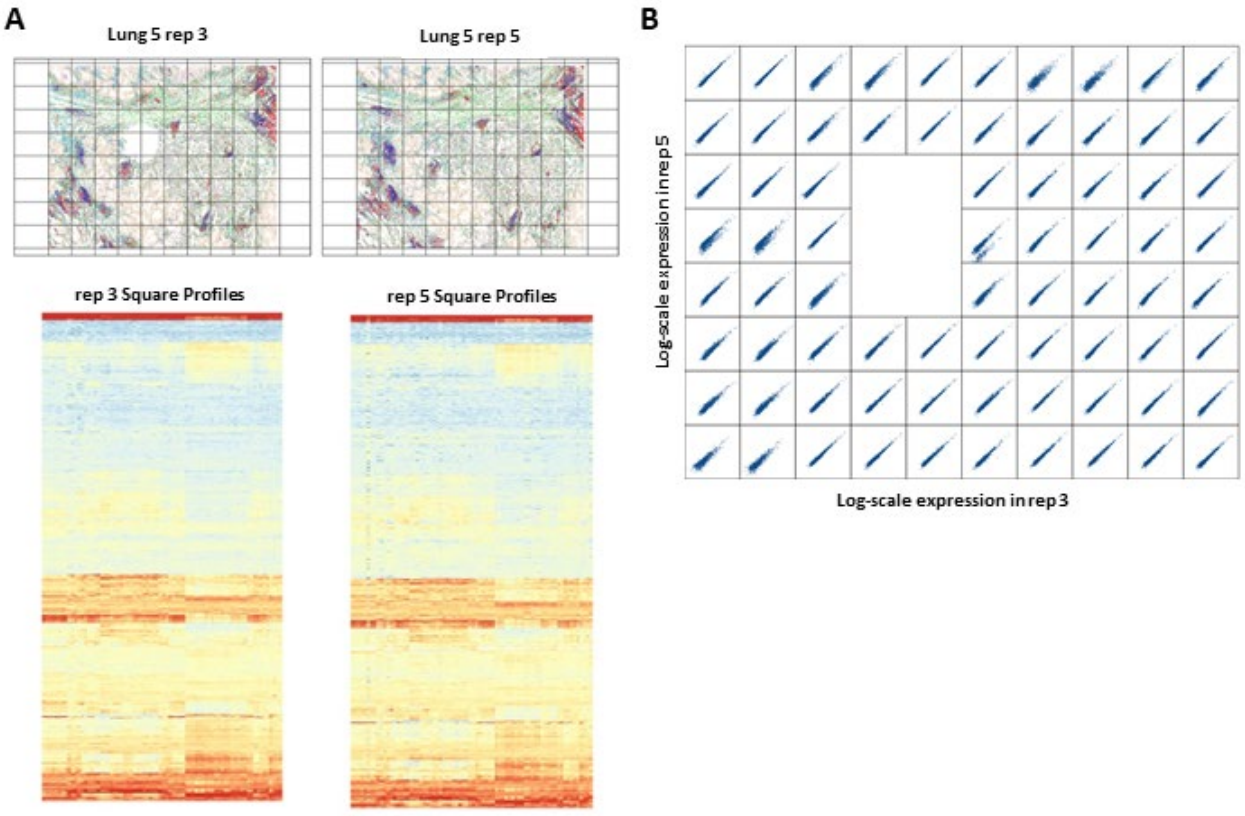


Figure 6. Concordance Between Serial FFPE Lung Sections Over a Spatial Grid. A. Each serial section of FFPE lung tissue (Lung 5 replicates 3 and 5) was partitioned into a grid. Squares held between 600 and 2,000 cells. **B.** Concordance between the 980-gene expression profiles of matching grid squares. Six squares overlapping a hole in rep 3 are excluded. To the right of the hole, a square with worse concordance is visible; this square contains part of a tertiary lymphoid structure in the rep 3 slide but not in the rep 5 slide.

6. SMI Chemistry Can Be Immediately Applied to Oligo-labeled Antibody Detection to Generate High-Plex Protein Imaging.

The same SMI encoded detection chemistry utilized for RNA can be applied to protein detection by conjugating the oligonucleotide-readout sequences to antibodies (Figure S13). This simple expansion of SMI chemistry for antibody-based detection is enabled by the extremely small size of the oligonucleotides required for SMI 64-bit encoding: only 60-to-80 base pairs are required to encode all of the multiplexing.

For the 80-plex protein panel described here, a set of 129 barcodes with Hamming distance 5 and Hamming weight 4 were generated using the same initial set of 64-bit barcodes used for RNA. The barcodes were designed to maximize the bit distribution over 16 hybridization

rounds and to minimize bit overlap across barcodes. A subset was selected to detect 78 target proteins, with the remaining 51 left as blank controls for misidentification quantification. For cell morphological visualization (*e.g.*, specific cell type markers such as CD45, PanCK, and membrane marker CD298), three reporter sequences were assigned to three antibodies with sequences that are orthogonal to the reporters within the 64-bit barcode set.

Using modified 64-bit SMI oligonucleotide barcodes, we performed site-specific conjugation of 80 antibodies against immune cell activation states and drivers of cancer progression (Table S5). Among these, 78 of these antibodies carried an encoded barcode to be read out over four detection events across 16 rounds of reporter hybridization. Three antibodies (PanCK, CD45, and CD298) were conjugated to a distinct set of oligonucleotide landing sites for non-encoded morphological visualization during a single hybridization event with a distinct set of fluorescent reporter oligonucleotides. Each antibody had been previously reviewed by a pathologist and validated on GeoMx DSP (20).

FFPE breast cancer tissue was prepared under standard IHC conditions with an overnight antibody incubation. Following washes and fiducial application, the tissue sample in the flow cells was placed on the SMI instrument. Protein localization patterns and relative expression levels were read out on the SMI instrument using the same cyclic chemistry as the RNA readout assay. The primary data output is decoded OME-TIF files, showing the localization pattern and relative intensity of each protein target in the assay (Figure 7).

The OME-TIF files were subset using cell segmentation masks to yield cell-by-cell expression patterns (Figure 7B). Protein localization patterns for a subset of targets, including CD45 and Histone H3, were cross validated by comparing the IF signals from antibodies detected using traditional IHC or DAPI stain with a single target detected per channel to the computationally decoded protein maps (Figure S14). All other localization patterns were further validated by a pathologist's review of decoded protein-localization patterns using appropriate immune and oncologic control tissues.

While this identical chemistry has been successfully utilized to plex over 900 RNA-targets, the ability to multiplex proteins needed to be established. To establish this capability, we performed SMI protein assays across various plex levels on 35 cell lines. We accomplished this plex-testing, by making "nested-multiplexed" assays. In a nested-multiplexed assay, we compare a 25-plex result on multiple FFPE cell lines with a 50-plex assay, where the 50-plex assay consists of the same targets as the 25-plex assay, but with 25 additional new targets. In this manner, one can graph the measured counts of the 25 targets measured in the 25-plex assay along the x-axis, and the counts of the same 25-targets when measured as part of the 50-plex assay. The measured counts should be independent of the level of plexity of the final assay. The nested protein assays at 25-plex, 50-plex, and 79-plex yielded highly concordant results over all FFPE cell lines (Figure

S15). These data suggest that protein detection on SMI can likely achieve well over 79-plex and may scale in a manner very similar to that of the RNA-based assay.

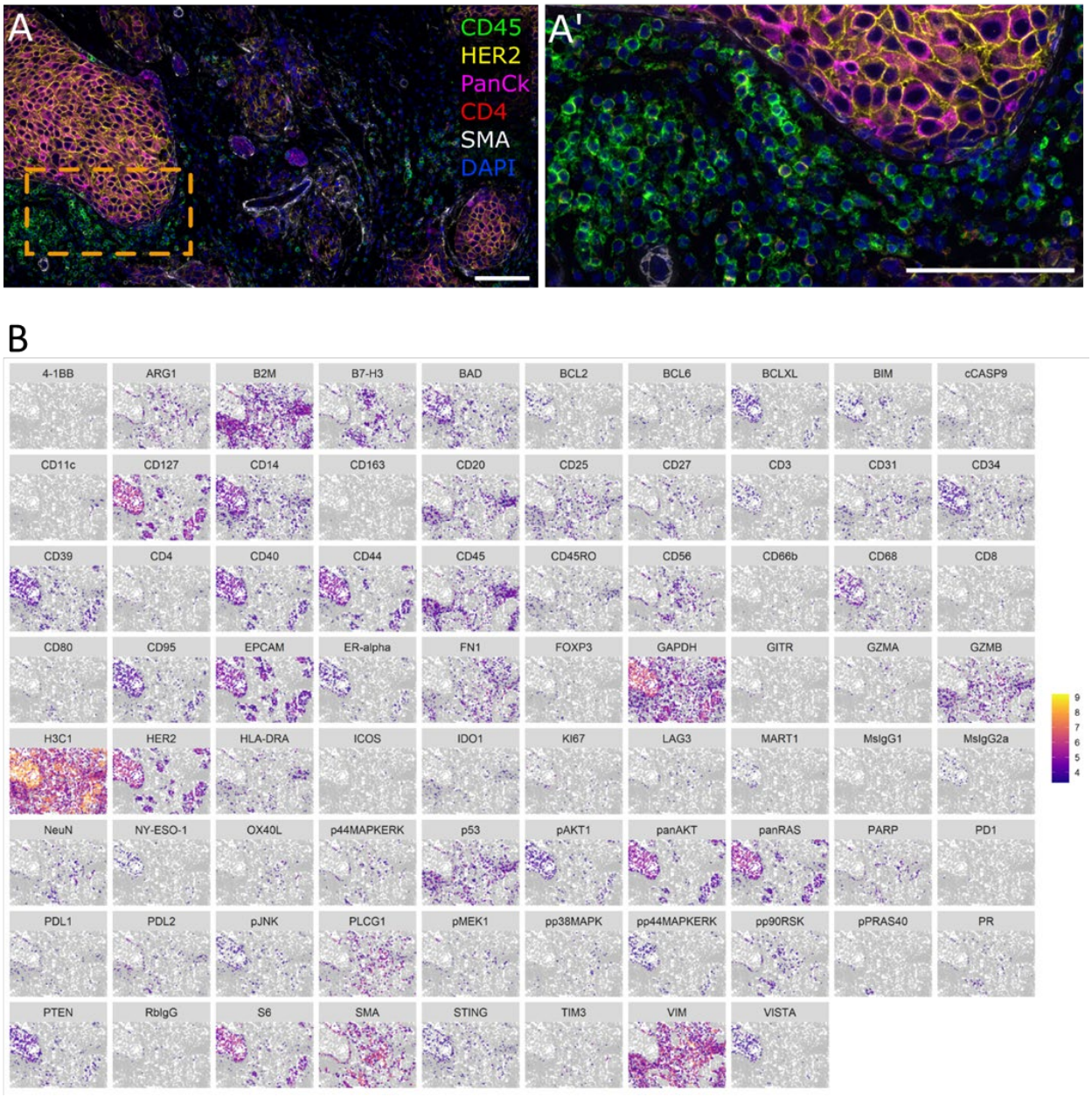


Figure 7. Spatial Subcellular Protein Analysis on SMI. A. Multi-channel overlay of five protein targets detected in a breast cancer biopsy (HER-2 positive invasive carcinoma) from an 80-plex assay. **A'.** Enlargement of boxed region in (A). Each decoded marker is visualized along with DAPI (blue) and the morphological markers PanCk (magenta) and CD45 (green). Scale bars: 100 µm. **B.** Single-cell expression

profiles across 78-encoded protein targets in the sample shown in (A), colored according to the log2 transformation of the SNR, with a threshold for visualization at SNR > 10.

Discussion

Real-world FFPE cancer samples are challenging to image at high-plex in a multiomic manner using conventional methods. As shown in Table 1A, over 96% of the 800,327 total cells in the samples passed quality control and were analyzable at 980-plex RNA-level with subcellular spatial resolution. Especially noteworthy, this high efficiency of detection was independent of the overall RNA integrity of the samples. Samples Lung 5 and Lung 13 had QV200 values under 30% and are classified as "too degraded" to even attempt to perform bulk RNA-sequencing (32), while Lung 9 and Lung 12 would be graded as having "medium" RNA quality. The fact that SMI had over 94% cell detection efficiency over this wide range of QV200 (and RIN scores) reflects the ability of the very small hybridization footprint of SMI chemistry (~50 bp hybridization zone and 60-to-80 bp readout zone) to quantitate very fragmented RNA patterns where the QV200 scores enter the "too degraded to sequence" classification.

The ultra-small readout zone also permits the very simple "migration" of the SMI chemistry from detecting in-situ RNAs, to detecting oligo-labeled antibodies. Our 60~80-bp oligonucleotide-labeled antibodies are easily conjugated and purified and retain their original specificity and sensitivity. Every antibody in this SMI protein panel migrated directly from the extensively validated antibody panels developed for the GeoMx Digital Spatial Profiler (20). All of DSP labeled antibodies successfully migrated to an SMI oligo-labeled antibody format. To the best of our knowledge, this represents the highest plex protein imaging experiment carried out on FFPE samples. Another unique feature of the SMI platform is the ability to work with standard FFPE samples and utilize standard, automatable FFPE sample preparation methods. SMI utilizes a sample preparation protocol nearly identical to standard low-plex in-situ hybridization, which does not require tissue clearing and/or expansion. Sample preparation for an SMI measurement takes a total of 1 day, with less than 60 min hands-on time. All of these procedures can be automated on a standard tissue sample processor such as the Leica Bond system (33). The simplicity and automated features of SMI sample preparation save time and labor and keep the tissue preparation highly consistent and reproducible (see Reproducibility Figure 6A and B). The SMI chemistry utilizes a controlled number of 15-to-60 fluorescent dyes encoded into a single 64-bit imaging barcode, providing sufficient fluorescence signal-to-noise to quantitate a single in-situ molecule of RNA or protein even in the presence of the severe autofluorescence of FFPE tissue samples.

SMI has demonstrated simultaneous detection of up to 980-plex RNA and up to 80-plex protein, which is the highest plexity of spatial RNA and protein profiling on FFPE tissue available

(to date) with single-cell (Figure 4) and subcellular resolution (Figure 2B). Increases in the plex of both RNA and protein assays are possible under this encoding scheme but have not yet been demonstrated. High accuracy is ensured by the error-robust color encoding scheme with large Hamming distance (HD ≥ 4) between targets. In addition to the high sensitivity and specificity of RNA and protein detection, the very low (< 4%) cell drop-out rate makes the data from almost every cell available for analysis, which has a distinct advantage over gene dropout in droplet-based single-cell sequencing methods (21).

SMI accomplishes RNA and protein multiplexing using a true 64-bit encoding method. This contrasts with techniques that are cycle non-encoded reagents for RNA and protein imaging. For instance, cyclic immunofluorescence methods (34) are limited to the number of fluorescent channels multiplied by the number of reagent imaging cycles, which increases linearly in the number of cycles of reagents (*e.g.*, 16 imaging rounds x 4 channel detection = 64-plex capability). SMI uses *encoded* barcoded antibodies and the same 16 imaging rounds with 4 channel detection, Hamming distance 4 and Hamming weight 4 yields a 1,210-plex capability. This class of encoded barcode technology provides a chemistry that can be essentially unlimited in plex with no fundamental change in instrumentation.

The multiomic capability of SMI is important for a number of key features. Firstly, while high-plex RNA images supply incredible information content concerning the overall activity of individual cells, the images themselves do not reveal detailed tissue architecture as well as protein images; for instance, comparing a typical RNA tissue image (Figure 2A) with a typical protein tissue image (Figure 7A). The protein-based images clearly reveal overall tissue architecture in exquisite detail, whereas the RNA-image reveals “punctate-spots” distributed in a much more uniform pattern. SMI is capable to combine both the robust RNA information content with the tissue architecture in the image.

Simultaneous RNA and protein detection on a single slide is also required for accurate assignment of RNA to single cells, especially in real tissue materials, where cells are of very divergent sizes and densities. Cell segmentation methods relying on nuclei staining to predict cell boundaries are particularly challenged by low- and high-density tissue cell environments. SMI uses a nuclei stain and universal membrane antibody-stain cocktails, along with state-of-art cell segmentation algorithms, to accurately outline the cell boundary. This is essential for high-quality single-cell data which enables downstream analysis including cell typing, cell state, and cell-cell interactions (see Figures 4 and 5).

The spatial imaging data from five tumors provides more richness than can be fully explored in a single manuscript. For experts in immuno-oncology, the images in Figure 4 will suggest insights and questions far beyond those discussed here. We expect that similar datasets

may often lead to multiple publications, as new analysts discover additional insights. We demonstrated a narrow application of spatial differential expression: we looked for spatial dependency in a single cell type in a single sample. Using our data, similar analyses may still be performed for 959 other genes in any of 18 cell types and in any of the 5 tumors. These results could then be contrasted across samples. Gene expression could be contrasted with more nuanced spatial variables, such as the expression profile among cells' neighbors, or in the expression profile of only a cell's neighbors of a given cell type. Experts in a given cell type or gene will find much to explore in these data.

With the advent of spatial molecular imaging using technologies such as the SMI platform, we can now directly test for the enrichment of pairwise ligand-receptor expression at the interface of interacting cells. Here, we find that out of unique ligand-receptor pairs, 16 were significantly enriched at the interface between tumor and T cells in at least one of five tumors. As expected, the interaction scores for each of these pairs varied across the tumors. This approach can be employed for any pair of cell types, and contrasts in ligand-receptor interactions can be drawn between different samples as well as between spatial contexts within a single sample.

We have not attempted a comprehensive analysis of this data in this first manuscript describing the SMI technology. For that reason, we are placing the raw and processed data described in this study into the public domain (http://nanostring.com/CosMx-dataset), where interested scientists from around the world can explore the data. Spatial imaging technologies such as SMI have the potential to greatly enhance the field of Spatial Biology. When SMI is combined with high-plex profiling technologies (such as Digital Spatial Profiling), a truly comprehensive spatial investigation of tissues can be accomplished.

Materials and Methods

FFPE Microarray

Custom FFPE cell pellet arrays of 16 cell lines for RNA assay and 35 cell lines for protein assay were made with A-FLX™ FFPE CELL PELLET by Acepix Bioscience, Inc. All cell lines were originally sourced from ATCC. The 16 cell lines for RNA assay include CCRF-CEM, COLO205, DU145, EKVX, HCT116, HL60, HOP92, HS578T, IGROV1, M14, MDA-MB-468, MOLT4, PC3, RPMI-8226, SKMEL2, SUDHL4. The 35 cell lines for protein assay include SHSY5Y CA, A431 CA, NCIH2228, DBTRG05MG, SW48, NB4, U251MG, U87MG, SUPB15, THP1, WSUNHL, SKMEL2, SKMEL5, SKBR3, RAMOS, RPMI8226, SUDHL6, RI1, RAJI, MDAMB468, HUT78, HUH7, HL60, HCT116, HCC78, H596, A431, HEK293 ICOS, HEK293 PD1-overexpressing, HEK293 CTLA4-overexpressing, HEK293 GITR-

overexpressing, HEK293 LAG3-overexpressing, HEK293 PDL2-overexpressing, and HEK293 PDL1-overexpressing. The breast cancer biopsy used for protein assay was CTRL301 tissue microarray supplied by US Biomax, Inc. The FFPE sample used was a malignant breast biopsy from a 61-year-old female with invasive carcinoma (Her-2 3+), T2N0M0 at IIA grade.

SMI Instrument

SMI utilizes standard sample preparation methods typical for FISH or immunohistochemistry on FFPE tissue sections, with the introduction of fluorescent bead-based fiducials that are fixed to the tissue to provide an optical reference for cyclic image acquisition and registration. For sample-processing details, see the "RNA assay FFPE tissue prep by manual method" section (below).

Following hybridization of ISH probes or antibody incubation, slides were washed and the coverslip with a defined height (50-to-100 µm) spacer was applied to the slide. The slide plus coverslip constitutes the flow cell, which was placed within a fluidic manifold on the SMI instrument for analyte (RNA or protein) readout and morphological imaging.

The in-situ chemistry and imaging analyses were performed on a prototype instrument with a custom large FOV, high numerical aperture, and low aberration objective (and associated optics) that permit more than 6,000 cells to be imaged per FOV from a flow cell assembled onto a standard glass slide. The prototype system can run up to 4 flow cells simultaneously and can interleave fluidics and optical-imaging operations to maximize throughput.

The optical system has an epi-fluorescent configuration that is based on a custom water objective with a numerical aperture (NA) of 0.82 and a magnification of 13X. The FOV size was customized to 0.7 mm x 0.9 mm. Illumination is widefield with a mix of lasers and LEDs that allow imaging of Alexa Fluor-488, Atto 532, Dyomics Dy-605, and Alexa Fluor-647 as well as cleaving of photolabile-dye components. The camera used is FLIR BFS-U3_200S6M-C based on the IMX183 Sony industrial CMOS sensor, and the sampling at the image plane is 180 nm/pixel. An XY stage moves the flow cell above the objective lens, and a Z-axis motor moves the objective lens.

The fluidic system uses a custom interface to draw reagents through a flow cell using a syringe pump. Reagent selection is controlled by a shear valve (Idex Health & Science). A flow sensor between the flow cell and the syringe pump was used for flow rate feedback (Sensirion AG). The fluidic interface includes a flat aluminum plate that is in direct contact with the flow cell. The temperature of this metal plate was controlled to regulate the reporter hybridization temperature between 20 to 35°C. The enclosure around the instrument was also maintained at a constant temperature using a separate thermoelectric cooler.

SMI In-Situ Hybridization (ISH) Probe Design

The ISH probes were designed to bind in-situ mRNA targets (Figure 1A). From 5' to 3', they each comprised a 35- to 50-nucleotide target-complementary sequence followed by four consecutive 10-to-20 nucleotide readout sequences corresponding to four "on" bits (HW4) assigned to each target. The target binding sequences in ISH probes were developed by a probe design pipeline that optimizes sensitivity and specificity for mRNA transcripts. The process begins with an exhaustive evaluation of all possible contiguous 35-50 nucleotide sequence windows for each mRNA target. This large pool of possible probe candidates was first filtered for ideal intrinsic characteristics including melting temperature (Tm), GC content, secondary structure, and runs of polynucleotides. Probes satisfying these parameters were further screened for homology to the full transcriptome of the parent organism utilizing the Basic Local Alignment Search Tool (BLAST) from the National Center for Biotechnology Information (NCBI). Preference was given to probes covering known protein coding transcripts, lying within coding regions, and maximizing the coverage of the isoform repertoire. Final panel candidates were further screened for intermolecular interactions with other probes in the candidate pool including potential probe-probe hybridization as well as minimizing common sequences between probes. Up to five oligonucleotide RNA detection probes were designed per target mRNA. Negative control probes were modeled after synthetic sequences from the External RNA Controls Consortium (ERCC) set (27). These negative ISH probes were designed to contain the same intrinsic characteristics and subjected to the same inter-/intra-molecular interaction screens as the primary panel of probes.

The 10-to-20 nucleotide readout sequences based on the 64-bit barcode design were filtered based on having a GC fraction greater than 35% to minimize the cross-hybridization between reporter probes and the junctions between readout sequences, as well as to maintain a Hamming distance of four between sequences. Also, readout sequences contained only bases A, C, and T to maximize binding kinetics between reporters and ISH probes.

SMI Reporter Design and Assembly

The SMI reporter is a defined 15-to-60-dye DNA construct assembled from three oligonucleotide motifs: dye (RPD), sub-branch (RPU) and nanoBarCode (nBC) (Figure S16). All oligos were made from DNA amidite (ChemGenes), while the RPU and nBC also contain PC-linker (Glen Research) for photocleavage. The RPD contains 15 nucleotides with a 5'-amino modifier used to conjugate Alexa Fluor-488, ATTO 532, Dyomics-605, or Alexa Fluor-647 fluorophores. The RPU contains a repeated 10-to-20-nucleotide nBC motif on the 5' end, followed by a PC-linker and a repeated 10-to-20-nucleotide RPD motif toward the 3' end. The nBC contains a 10-to-20 nucleotide ISH-probe motif at the 3' end, followed by a PC-linker and a repeated RPU motif at the 5' end. Each RPD has one corresponding RPU and 16 nBCs for a total of 64 reporters. Lyophilized RPD and RPUs were normalized to 1.5 mM and 0.5 mM in Tris EDTA (TE) pH 8, respectively. The nBCs were normalized to 50 µM in TE. All oligos were stored at 4°C prior to use.

Reporter assembly occurs via a two-stage process involving RPD-RPU hybridization and nBC hybridization (Figure S16). The 980-plex RNA and 80-plex protein readout requires 16 sets of pools. Each set contains four reporters with four different fluorophores. The pools were made by diluting the 1 µM stock to a 5 nM/probe solution in 8.75X SSPE, 0.5% Tween-20 (%w/v) and 0.1% ProClin™ 950. Pool identity and cleavage were assessed using biotinylated targets on a streptavidin-covered slide (Schott). Pools were stored at 4°C until ready to use.

Antibody Conjugation

All antibodies were sourced from vendors in a BSA- and glycerol-free format. Antibodies were quantified using UV spectrophotometry and quality-checked using gel electrophoresis. Antibody heavy chains were prepared for conjugation using SiteClick™ azide modification (Invitrogen) of carbohydrate domains. Amine modified oligo tags were conjugated to antibodies using modifications of Site-Click chemistry (Invitrogen) and heterbifunctional DBCO linker (Click Chemistry Tools). The resulting conjugates were HPLC purified and normalized to 200 µg/mL as described (20).

For the protein detection assay, the oligonucleotides resembled those used for SMI RNA detection in that there were landing sites for four of the 64 nano-barcode reporters, corresponding to the 64-bit barcode (Figure S13). For imaging of tissue morphology and cell boundaries, the antibodies were conjugated to an oligonucleotide possessing a single landing site for one of four morphology reporters, which have the same overall structure as the nano-barcode reporters but bind a unique set of four sequences that are orthogonal to the 64 landing sites used in the readout assay. This allows the signals from antibodies used for morphological visualization to be eliminated prior to the high-plex RNA and protein readouts.

SMI 980-Plex RNA Panel Design

For comprehensive RNA profiling, we designed a 980-plex panel (Table S1 and Figure S5E) to investigate the biology of single cells across tumors and diverse organs. To select panel content, 749 genes were selected to capture critical cell states and cell-cell interactions. The remaining genes were selected to optimize the panel's power to distinguish between different cell types.

While cell state is a broad term encompassing a wide variety of cellular phenotypes and processes, we focused our curation on core pathways or environmental factors that are important across broad areas of physiology and disease. These included immune cell states, basic cellular processes (*e.g.,* apoptosis, autophagy), cellular structures that integrate environmental cues (*e.g.*, cytoskeleton, extracellular matrix), and stress or damage responses (*e.g.*, hypoxia,

wound healing, DNA damage response) (Figure S5). Genes for these cell states were primarily curated from the literature, using hallmark genes for each set. Genes for cellular signaling pathways, including ligands and receptors, were curated from the literature, HUGO gene families, and the KEGG BRITE Ontology. Finally, target genes for common cell signaling pathways were included when those transcriptional outputs were known with confidence. Consensus target genes were determined for the MAPK (35), NF-κB (35-38), Interferon (39), Wnt (40-47), and Hedgehog pathways (48).

Cell state and cell signaling categories were also subjected to a data-driven expression level assessment. We employed 76 individual Human Cell Landscape Datasets to score for expression level across cell types and tissues. We removed genes that were highly unlikely to be detected in any tissue types. However, we utilized subject matter expertise to selectively retain genes of high biological interest, for example, T-cell checkpoint genes. While these genes may be rarely detected in healthy tissue, they are crucial to understanding the immune system's role in disease (Figure S5).

Once genes of biological interest were selected, the remainder of the panel was chosen to power cell typing. First, literature-driven searches identified 56 highly informative immune cell markers (31, 49, 50) and 6 adipocyte markers (51-54). The remaining genes were then chosen in a data-driven manner to maximize the contrast between different cell type's expression profiles. This process was motivated by the intuition that useful cell-typing genes will have two properties: they will vary strongly across cell types, and they will have high expression levels in at least one cell type.

Specifically, for a given pair of cell types, we scored each gene's "pairwise distance" as $(|x1 - x2| / \mathrm{sqrt}(x1 + x2)) \max(x1, x2)$,

where x1 and x2 are the gene's expression levels in the two cell types. The left side of this equation was motivated by the *t*-statistic and the mean-variance relationship of the Poisson distribution. The right side was motivated by the assumption that every transcript of a gene offers a fixed amount of evidence for one cell type versus another, and that the total evidence from a gene will therefore vary in proportion to its total transcripts. While the exact statistical power conferred by a gene depends on the cell typing method is used, the above statistic formalizes the intuition that useful cell-typing genes will have two properties: they will vary strongly across cell types, and they will have high expression levels in at least one cell type.

For every gene, the pairwise distance between all pairs of cell types within each dataset was calculated. Then, the total pairwise distance between all cell types within each dataset was calculated over the genes already chosen for the panel. Cell typing genes were then chosen in a greedy manner to improve the pairwise distances between the least distant cell types across the datasets. To initially choose genes that were informative between many pairs of cell types, the

first genes were chosen to maximally increase the 40th percentile of between-cell distances. Over 224 successive iterations, this percentile was dropped to 0.005 to focus on increasing the distances between the most similar cell types.

RNA Assay FFPE Tissue Prep by Manual Method

Five-micron tissue sections were cut from FFPE tissue blocks using a microtome, placed in a heated water bath, and adhered on Leica Bond Plus Microscope slides (Leica Biosystems). Slides were then dried at room temperature overnight.

To perform in-situ hybridization on the tissue sections, the tissue slides were baked in 60°C oven overnight. The tissue sections were dewaxed in xylene (Millipore) for 5 min twice, ethanol (Pharmco) for 2 min twice, and then the slides were baked in 60°C oven for 5 min. The tissue sections were subjected to the target retrieval step using RNAscope™ Target Retrieval kit (ACD Bio) and heated at 100°C in a pressure cooker for 15-30 min. After target retrieval, the tissue sections were rinsed with diethyl pyrocarbonate (DEPC)-treated water (DEPC H_2O) (Thermofisher), washed in ethanol for 3 min, and dried at room temperature for 30 min. On the dried slide, a hydrophobic barrier line was drawn around the tissue section using ImmEdge™ Pen (Vector).

The tissue was then digested with Protease Plus (ACD Bio) spiked with proteinase K (ThermoFisher), ranging from 1 µg/mL to 5 µg/mL depending on tissue type, at 40°C for 15-30 min. The tissue sections were rinsed with DEPC H_2O twice, incubated in 1:400 diluted fiducials (Bangs Laboratories) in 2X saline sodium citrate and Tween (SSCT; 0.001% Tween-20, Teknova) for 5 min at room temperature, and washed with 1X phosphate buffered saline (PBS; ThermoFisher) for 5 min. After digestion and fiducial placement, the tissue was fixed with 10% neutral buffered formalin (NBF) for 1 min to maintain soft tissue morphology, then washed twice with Tris-glycine buffer (0.1M glycine [Sigma], 0.1M Tris base [FisherScientific] in DEPC H_2O) for 5 min, and washed with 1X PBS for 5 min. The fixed tissue was blocked using 100 mM *N*-succinimidyl (acetylthio) acetate (NHS-acetate; ThermoFisher) diluted in NHS-acetate buffer (0.1M NaP+0.1% Tween PH8 in DEPC H2O) for 15 min at room temperature and washed in 2X saline sodium citrate (SSC) for 5 min. Adhesive SecureSeal™ Hybridization Chamber (Grace Bio-Labs) was placed to cover the tissue.

NanoString ISH probes were prepared by incubation at 95°C for 2 min and immediately transferred to ice. The ISH probe mix (1 nM ISH probe, 1X Buffer R, 0.1 U/µL SUPERase•In™ [Thermofisher] in DEPC H_2O) was then pipetted into the chamber and adhesives supplied with the chambers were applied to the ports of the chambers. Hybridization occurs at 37°C overnight after sealing the chamber to prevent evaporation. After ISH probe hybridization, the tissue

sections were washed twice in a buffer composed of 50% formamide (VWR) in 2X SSC at 37°C for 25 min, washed twice with 2XSSC for 2 min each at room temperature, and then blocked with 100 mM NHS-acetate for 15 min. After blocking, the hydrophobic barrier was removed with a blade, and a custom-made flow cell was attached to the slide.

Automated RNA Assay FFPE Tissue Prep on Leica Bond RX

Same as FFPE tissue prep by manual method, tissue slides were baked overnight at 60°C to ensure tissue adherence to the positively charged glass slides (Leica Bond Plus Microscope slides, Leica Biosystems). Then the tissue was deparaffinized and digested with Protease Plus (ACD Bio) spiked with proteinase K (ThermoFisher) ranging from 1 µg/mL to 5 µg/mL depending on tissue type and prepared for heat-induced epitope retrieval (HIER) on a Leica Biosystems automated tissue handler (Bond RX, Leica Biosystems). For cell pellet array (CPA) or tissue microarray (TMA) samples, HIER requires the treatment in Leica buffer ER1 at 100°C for 8 min, while 30min treatment for tissue samples. After Leica handling, the remaining process is the same as RNA assay FFPE tissue prep by manual method described above.

RNA Isolation, rRNA Depletion, NGS Library Preparation and Sequencing

Total lung RNA was isolated from single or double 20 µm FFPE curls using RNeasy FFPE Kit (Qiagen) and digested with proteinase K for 30 min. Lung RNA was quantified using an Agilent RNA 6000 Nano Kit (Agilent) according to the manufacturer's instructions. DV200 and RIN scores were determined using Agilent BioAnalyzer 2100 Expert Software B.02.09. For DV200, Region 1 selection was from 200 nt to about 8,000 nt using Smear Analysis.

RNA samples were processed using NEBNext® rRNA Depletion Kit (Human/Mouse/Rat) (New England BioLabs), NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (New England BioLabs) with varying inputs (117 ng to 1 µg). PCR enrichment of cDNA was performed using NEBNext Multiplex Oligos for Illumina Set 2 (New England BioLabs). After 12 cycles of PCR elongation, NGS libraries were quantified using an Agilent High Sensitivity DNA Kit (Agilent), and libraries with low quantification were subjected to additional elongation cycles. NGS libraries were purified using Agencourt AMPure XP (Beckman Coulter), and double-sided size selection was performed on a single sample as described by Bruinsma *et al.* (55). NGS library quality was assessed using Agilent Tapestation High Sensitivity D1000 (Agilent) and quantified using a Qubit 4 Fluorometer (ThermoFisher Scientific). NGS libraries were diluted and sequenced using a NextSeq 2000 (Illumina) and NextSeq 1000/2000 P2 Reagents (100 Cycles) v3 (Illumina) according to the manufacturer's instructions.

Automated Protein Assay Sample Preparation on Leica Bond RX

Deparaffinization and antigen retrieval were performed automatically on a BOND RX machine (Leica Biosystems) using Bond Dewax Solution (30 seconds at 72°C) and Bond ER Solution 1 (20 minutes at 100°C). Samples were blocked for 1 hour with blocking buffer W (NanoString). Oligonucleotide-conjugated primary antibodies were pooled and diluted uniformly to 100 ng/ml in blocking buffer. Samples were incubated with primary antibodies overnight at 4°C. Then, samples were incubated with fiducials (0.00025%, Bangs Laboratories) for 5 minutes. To secure antibodies and fiducials, the samples were fixed with 4% paraformaldehyde for 10 minutes at room temperature and then washed in 1 x PBS. Prior to cyclic encoded protein detection, slides were incubated with *N*-succinimidyl (acetylthio)acetate (NHS-acetate, Thermo Scientific, diluted in 0.0932 M Na2HPO4 + 6.8 mM NaH2PO4 + 0.1% Tween pH 8) for 15 minutes on the SMI instrument.

Cyclic RNA Readout on SMI Instrument

Processed tissue was assembled into the flow cell and loaded onto the SMI instrument. The tissue was washed with 1.5 mL of reporter wash buffer to remove air bubbles in the flow cell. The Reporter Wash Buffer consisted of 8.75XSSPE, 0.5% Tween 20, 0.1 U/µL SUPERase•In RNase Inhibitor (20 U/µL), 1:1000 Proclin 950 and DEPC-treated water. A low-resolution image of the whole slide was acquired to allow the user to access the tissue, and then FOVs were placed at the areas of interest.

To start the cyclic RNA readout, 100 µL Reporter 1 was flowed in at 200 µL/min and incubated for 15 minutes. After incubation, 1 mL of Reporter Wash Buffer was flowed into the flow cell at 750 µL/min to wash out the reporter probes that did not hybridize. Following reporter wash, 100 µL imaging buffer was flowed into the flow cell prior to imaging. Imaging buffer consisted of 80mM of glucose, 0.6 U/mL of pyranose oxidase from Coriolus sp., 18 U/mL of catalase from bovine liver, 1:1000 Proclin 950, 500 mM Tris-HCl Buffer pH 7.5, 150 mM sodium chloride, and 0.1% Tween 20 in DEPC-treated water. After eight Z-stack images (0.8 µm step size) of each FOV were acquired, fluorophores on the reporter were UV cleaved and washed off with 200 µl strip wash buffer. The strip wash buffer consisted of 0.0033XSSPE, 0.5% Tween 20, and 1:1000 Proclin 950. This fluidic and imaging procedure was repeated for the 16 reporter pools, and the 16-round of reporter hybridization can be repeated multiple times for increased sensitivity.

After all cycles were completed, the tissue was subjected to morphology stain workflow on the same instrument. The blocking buffer (Buffer W, NanoString) was incubated in the flow

cell for 30 min, then the tissue was incubated with a 4-antibody cocktail (CD298, CD45, CD3, PanCK) diluted in Blocking Buffer for 1 h. After the incubation, the tissue was washed with 8 mL of Reporter wash and then 100 µL of imaging buffer was flowed into the flow cell, prior to collecting the antibody-labeled images of the tissue.

Cyclic Protein Readout on the SMI Instrument

The on-instrument SMI protein assay readout was performed as described for RNA, with three exceptions. The readout was only 16 rounds of hybridization with no repeated cycling. Morphology visualization was performed using oligonucleotide conjugated antibodies (Figure S13), as described in the Antibody Conjugation section. Tissue morphology and cell membrane morphology were visualized following hybridization of a specific pool of nano-barcode reporters to the oligonucleotide-conjugated antibodies. The protein images were captured with the implementation of a high dynamic range imaging protocol to accommodate a large range of expression variation without camera sensor saturation.

Primary Data Processing

Primary data processing is a standard image analysis of a three-dimensional multi-channel image stack obtained at each FOV location. The objective of this analysis is to reduce the multi-dimensional image stacks to a single list of individual reporters seen in the specific binding event. This process was performed in parallel across all FOVs and occurs in-line with data acquisition. The image processing comprises three main steps – registration, features detection, and localization.

3D rigid image registration was performed with the use of fiducial markers embedded within the tissue sample. The fixed image reference was established at the start of the experiment prior to reporter hybridization. Subsequent image stacks, shifted by stage motion, were matched to this reference using phase correlation. Individual channels within the image were aligned to each other through the application of a pre-calibrated affine transformation.

The RNA image analysis pipeline focuses on identifying diffraction-limited features that represent the fluorescence response from a single molecule. Once the image stacks were registered, a 2D Laplacian of Gaussian (LoG) filter was applied to each channel to remove background and enhance the encoded reporter signatures. The kernel size and standard deviation of the filter were matched to the expected reporter point spread function. Post filtering, potential reporter locations were identified as local maxima using a 3D nearest neighbor search. Only local maxima greater than a channel-specific threshold were retained for further localization.

Thresholds for each channel were predetermined empirically based on the signal-to-noise ratio (SNR) of the fluorescent channel. The retained maxima were assigned a confidence value determined by the intensity of the reporter signal.

Sub-pixel localization of each feature was obtained by fitting a 2D polynomial to the maxima and analytically solved for the sub-pixel maxima locations in X and Y. The final reporter signature locations in X, Y, and Z axes along with the assigned confidence were recorded in a list assigned to the specific reporter binding event. After the acquisition was completed, all features pertaining to a given FOV were collated into a single list that forms the basis of the secondary analysis.

The protein image analysis pipeline uses locally adaptive thresholds to segment regions that range from single diffraction-limited features to large contiguous clusters within the FOV. The output of the primary data analysis is a series of 2D binary maps that represent encoded reporter signal in each hybridization round.

Secondary Analysis and Decoding

The imaging data converted to a table of XYZ locations of all individual reporter binding events was used to determine the presence of individual transcripts. To start this process, each unique location with at least one reporter binding event is considered a 'seed', and all neighboring locations to each seed with at least one reporter binding event were determined. The neighbor search is limited to a radius of 0.5 pixels (90 nm) in a first pass through the data. Any seeds with fewer than 4 unique reporter probe binding events in the neighborhood were removed from being considered for transcript decoding due to their inability to form a complete gene-specific barcode. All possible four reporter combinations of unique reporter probes in a seed's neighborhood such that at least one of the four reporter binding events was present at the seed location were then matched with gene-specific barcodes to detect the presence of a gene in a seed's neighborhood. If more than one gene was detected in a seed's neighborhood, the seed and all the transcripts detected in its neighborhood were dropped from further analysis.

All the seeds (or transcripts) retained after this step went through another filtering step. In this step, any seeds with a high probability of making a transcript call by random chance (and hence the transcripts detected in their neighborhoods) were then dropped from further analysis in first pass through the data. Given that a set number of barcodes (*e.g.,* 980) out a pool of all possible four reporter barcodes that can be generated using 64 unique reporters (465,920) were used to denote gene types, there is a non-zero probability that any random combination of four unique reporters can match a gene-specific barcode. This probability is further increased when the unique reporter binding events in a seed's neighborhood can be used to generate many

potential four reporter barcodes. To ensure high confidence in the presence of a true transcript for every target call made, any seeds (and associated transcript calls) with a random transcript call probability of more than 2% were dropped from further analysis in first pass through the data. The reporter binding event locations that contributed to making transcript calls were used to estimate a centroid location for all transcripts retained after this filtering step. All reporter binding events that contributed to making retained transcript calls were then removed from the original imaging dataset and this modified dataset was used as a starting point for a second pass through the data.

The second pass through the data repeated all the steps above albeit with an increased target search radius of 0.75 pixels (135 nm). The rationale behind this increase in radius was to try and recover any potential transcript calls that may have been lost due to local tissue motion during reagent cycling. After the conclusion of two passes through the data, we have a list of potential transcript locations in each FOV. However, as we consider each unique reporter binding event location as a 'seed' for transcript calling, there is a potential for duplicate calling of each individual transcript, artificially inflating the total count of transcripts detected in a FOV. For instance, if all four reporter binding events contributing to a transcript call occur at slightly different locations, the same transcript could be counted four times. To prevent this from happening, for each gene, all transcript calls made are filtered to ensure that there is no other transcript call present within a radius of 0.75 pixels from each transcript's estimated location. Whenever multiple transcript calls are found within this search radius, the transcript call with the highest number of reporter binding events contributing to it is retained and others discarded.

Cell Segmentation Algorithm

Accurate cell segmentation that assigns transcripts to cell locations is critical to data quality, but it is very challenging for tissue sections where cells are tightly packaged with shared 3D boundaries and uneven morphology staining. To address this issue, we have developed a cell segmentation pipeline combining image preprocessing and machine-learning techniques (Figure S3). Deep convolutional neural network such as Cellpose has shown good performance overall (56). This algorithm models cell intensities as heat diffuses from cell centers and extends to the boundaries. Spatial gradients were derived from this diffusion map and fed into a U-Net neural network architecture to train using annotated cell boundaries. The network was then used to predict the gradient values of input fluorescence images. The spatial vector field of the gradient values predicted by the network was further converted into cell mask via watershed algorithm (57). The usage of a spatial vector field as an intermediate representative of cell objects resulted in high detection of real cell boundaries over other structures, which would often lead to false detections using conventional methods, while other machine learning methods that rely on

distance to nuclei as penalty function in cell mask generation, often produce circular shapes from the extension of nuclei shape and not matching the real cell shape definition, creating localization errors around the cell borders.

The direct application of Cellpose is often suboptimal when raw images have issues in: 1) cells on the out of focus plane, 2) uneven staining, 3) high autofluorescence signals, and 4) low cell boundary contrast. To mitigate these issues and create a more robust pipeline, we implemented image enhancement prior to segmentation (Figure S4). Our pipeline takes tissue images stained with both nuclear and membrane markers (DAPI, CD298/PanCK/CD3) to perform rescaling, normalization, image deconvolution and boundary enhancement. Membrane channels of the tissue images were combined and normalized to the range of the corresponding nuclear channel. Image subtraction was performed between the nuclear channel and the normalized membrane images. This process increases the contrast between adjacent cells and significantly reduces the intake of tissue auto-fluorescence signal. The preprocessed images were fed into pre-trained Cellpose neural network models for both nuclear segmentation methods using nuclear channel only and cytoplasm segmentation using combined nuclear and membrane channels. The combination of nuclear and membrane segmentation modes ensures that an appropriate cell boundary definition, with which the cell boundary will be accurately detected when the membrane is present in the image. The single nuclear channel mode ensures that cells with weak membrane staining can still be segmented correctly. Results from two segmentation tasks were combined to select the best results from each mode, by analyzing intersection and union between all segmented cells. The use of both modes also enables the pipeline to perform subcellular compartment analysis. The combination of the preprocessing steps, the dual mode segmentation, and using the publicly available pretrained models have greatly improved the robustness of cell segmentation outcomes, making retraining the models unnecessary for most tissue images.

The analysis pipeline has run with two different modes of hardware processors: Cuda GPU (Ge Force RTX 2070, Quadro P4000) and parallel CPU threads, with average segmentation runtime of around 94 secs per FOV with 5 channels, 6 to 8 Z-slices and image size of 5,472 x 3,678 pixels. The final segmentation step was to map each transcript location in the registered image to the corresponding cell, as well as the cell compartment (nuclei, cytoplasm, membrane), where the transcript is located. Other features/properties generated include shape (area, aspect ratio) and intensity statistics (min, max, average) per cell.

Segmented Regression to Estimate the Relationship Between SMI and RNA-seq

Segmented regression was performed using the R package “segmented”. Log-transformed SMI counts were modeled as a function of log-transformed RNA-seq counts. To

avoid negative-infinity values after log-transformation, RNA-seq counts were thresholded below at their lowest non-zero value. Estimated breakpoints and linear trends were extracted from the segmented regression models.

Analysis of NSCLC Samples

Single-cell expression profiles were derived by counting the transcripts of each gene that fell within the area assigned to a cell by the segmentation algorithm. Cells with fewer than 20 total transcripts were omitted from the analysis. A normalized expression profile was defined for each cell by dividing its raw counts vector by its total counts. A separate UMAP projection was computed for each tissue.

After analyzing the single-cell expression data, we created a single-cell "neighborhood" matrix. This matrix specifies the number of each cell type among each cell's 200 closest neighbors in 2D physical space. This matrix was then input into the UMAP algorithm. To define "niches", this matrix was clustered using the Mclust algorithm.

To test whether each gene was differentially expressed between macrophages in different niches, a linear model was run predicting raw counts from niche. Only macrophages from Lung 6 were used in this analysis. A global *p*-value for each gene was taken using a likelihood ratio test comparing this linear model to the null model, using the R package lmtest.

Protein Expression Visualization

The primary outputs of protein data processing are protein localization maps for each antibody in the assay, as well as images of the three antibodies used for tissue morphology and cell membrane visualization. Protein expression per cell was reported as a sum of the expression values of pixels present in both the cell label and protein localization mask. The SNR was calculated by dividing the protein intensity by the global mean of the isotype control intensities for that FOV. A threshold of 10 was set for a positive SNR signal for the purpose of visualization and spatial analysis. Visualization of protein localization patterns was performed in ImageJ. Single-cell expression visualization was performed in R.

Data Availability

The data from the NSCLC is available at http://nanostring.com/CosMx-dataset.

Funding Information

Research and development reported in this publication was supported in part through a strategic development collaboration between NanoString and Lam Research (Fremont, CA).

Declaration of Interest

All authors are employees of NanoString Technologies Inc. and hold NanoString stock or stock options.

Reference

1. Garon EB, Rizvi NA, Hui R, Leighl N, Balmanoukian AS, Eder JP, et al. Pembrolizumab for the treatment of non-small-cell lung cancer. N Engl J Med. 2015;372(21):2018-28.
2. Yu H, Batenchuk C, Badzio A, Boyle TA, Czapiewski P, Chan DC, et al. PD-L1 Expression by Two Complementary Diagnostic Assays and mRNA In Situ Hybridization in Small Cell Lung Cancer. J Thorac Oncol. 2017;12(1):110-20.
3. Yu H, Boyle TA, Zhou C, Rimm DL, Hirsch FR. PD-L1 Expression in Lung Cancer. J Thorac Oncol. 2016;11(7):964-75.
4. Ting DT, Lipson D, Paul S, Brannigan BW, Akhavanfard S, Coffman EJ, et al. Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. Science. 2011;331(6017):593-6.
5. Garber K. Oncologists await historic first: a pan-tumor predictive marker, for immunotherapy. Nature Biotechnology. 2017;35(4):297-8.
6. Sokolenko AP, Imyanitov EN. Molecular Tests for the Choice of Cancer Therapy. Curr Pharm Des. 2017;23(32):4794-806.
7. Dereli AS, Bailey EJ, Kumar NN. Combining Multiplex Fluorescence in situ Hybridization with Fluorescent Immunohistochemistry on Fresh Frozen or Fixed Mouse Brain Sections. J Vis Exp. 2021(172).
8. Taube JM, Akturk G, Angelo M, Engle EL, Gnjatic S, Greenbaum S, et al. The Society for Immunotherapy of Cancer statement on best practices for multiplex immunohistochemistry (IHC) and immunofluorescence (IF) staining and validation. J Immunother Cancer. 2020;8(1).
9. Hirsch FR, McElhinny A, Stanforth D, Ranger-Moore J, Jansson M, Kulangara K, et al. PD-L1 Immunohistochemistry Assays for Lung Cancer: Results from Phase 1 of the Blueprint PD-L1 IHC Assay Comparison Project. J Thorac Oncol. 2017;12(2):208-22.
10. Udall M, Rizzo M, Kenny J, Doherty J, Dahm S, Robbins P, et al. PD-L1 diagnostic tests: a systematic literature review of scoring algorithms and test-validation metrics. Diagnostic Pathology. 2018;13(1):12.
11. Halse H, Colebatch AJ, Petrone P, Henderson MA, Mills JK, Snow H, et al. Multiplex immunohistochemistry accurately defines the immune context of metastatic melanoma. Sci Rep. 2018;8(1):11158.
12. Macosko Evan Z, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 2015;161(5):1202-14.

13. Zilionis R, Nainys J, Veres A, Savova V, Zemmour D, Klein AM, et al. Single-cell barcoding and sequencing using droplet microfluidics. Nature Protocols. 2017;12(1):44-73.
14. Wang X, He Y, Zhang Q, Ren X, Zhang Z. Direct Comparative Analyses of 10X Genomics Chromium and Smart-seq2. Genomics, Proteomics & Bioinformatics. 2021.
15. See P, Lum J, Chen J, Ginhoux F. A Single-Cell Sequencing Guide for Immunologists. Front Immunol. 2018;9:2425.
16. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 2016;353(6294):78-82.
17. Stickels RR, Murray E, Kumar P, Li J, Marshall JL, Di Bella DJ, et al. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. Nature Biotechnology. 2021;39(3):313-9.
18. Fu X, Sun L, Chen JY, Dong R, Lin Y, Palmiter RD, et al. Continuous Polony Gels for Tissue Mapping with High Resolution and RNA Capture Efficiency. bioRxiv. 2021:2021.03.17.435795.
19. Liu Y, Yang M, Deng Y, Su G, Enninful A, Guo CC, et al. High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue. Cell. 2020;183(6):1665-81.e18.
20. Merritt CR, Ong GT, Church SE, Barker K, Danaher P, Geiss G, et al. Multiplex digital spatial profiling of proteins and RNA in fixed tissue. Nat Biotechnol. 2020;38(5):586-99.
21. Qiu P. Embracing the dropouts in single-cell RNA-seq analysis. Nature Communications. 2020;11(1):1169.
22. Moffitt JR, Hao J, Wang G, Chen KH, Babcock HP, Zhuang X. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. Proceedings of the National Academy of Sciences. 2016;113(39):11046-51.
23. Moffitt JR, Zhuang X. Chapter One - RNA Imaging with Multiplexed Error-Robust Fluorescence In Situ Hybridization (MERFISH). In: Filonov GS, Jaffrey SR, editors. Methods in Enzymology. 572: Academic Press; 2016. p. 1-49.
24. Groiss S, Pabst D, Faber C, Meier A, Bogdoll A, Unger C, et al. Highly resolved spatial transcriptomics for detection of rare events in cells. bioRxiv. 2021:2021.10.11.463936.
25. Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Ferrante TC, Terry R, et al. Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. Nature Protocols. 2015;10(3):442-58.
26. Eng C-HL, Lawson M, Zhu Q, Dries R, Koulena N, Takei Y, et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. Nature. 2019;568(7751):235-9.
27. Baker SC, Bauer SR, Beyer RP, Brenton JD, Bromley B, Burrill J, et al. The External RNA Controls Consortium: a progress report. Nat Methods. 2005;2(10):731-4.
28. Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald ER, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019;569(7757):503-8.
29. Nusinow DP, Szpyt J, Ghandi M, Rose CM, McDonald ER, 3rd, Kalocsay M, et al. Quantitative Proteomics of the Cancer Cell Line Encyclopedia. Cell. 2020;180(2):387-402.e16.
30. NCI-60 Human Tumor Cell Lines Screen: National Cancer Institute; [Available from: https://dtp.cancer.gov/discovery_development/nci-60/.
31. Danaher P, Kim Y, Nelson B, Griswold M, Yang Z, Piazza E, et al. Advances in mixed cell deconvolution enable quantification of cell types in spatially-resolved gene expression data. bioRxiv. 2020:2020.08.04.235168.
32. Evaluating RNA Quality from FFPE Samples: Illumina; 2021 [Available from: https://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/evaluating-rna-quality-from-ffpe-samples-technical-note-470-2014-001.pdf.
33. BOND-III Fully Automated IHC and ISH Staining System: Leica Biosystems; 2021 [Available from: https://www.leicabiosystems.com/ihc-ish-fish/fully-automated-ihc-ish-instruments/bond-iii/.

34. Lin JR, Fallahi-Sichani M, Chen JY, Sorger PK. Cyclic Immunofluorescence (CycIF), A Highly Multiplexed Method for Single-cell Imaging. Curr Protoc Chem Biol. 2016;8(4):251-64.
35. Wagle M-C, Kirouac D, Klijn C, Liu B, Mahajan S, Junttila M, et al. A transcriptional MAPK Pathway Activity Score (MPAS) is a clinically relevant biomarker in multiple cancer types. npj Precision Oncology. 2018;2(1):7.
36. Son YH, Jeong YT, Lee KA, Choi KH, Kim SM, Rhim BY, et al. Roles of MAPK and NF-kappaB in interleukin-6 induction by lipopolysaccharide in vascular smooth muscle cells. J Cardiovasc Pharmacol. 2008;51(1):71-7.
37. Kang HB, Kim YE, Kwon HJ, Sok DE, Lee Y. Enhancement of NF-kappaB expression and activity upon differentiation of human embryonic stem cell line SNUhES3. Stem Cells Dev. 2007;16(4):615-23.
38. Hiscott J, Alper D, Cohen L, Leblanc JF, Sportza L, Wong A, et al. Induction of human interferon gene expression is associated with a nuclear factor that interacts with the NF-kappa B site of the human immunodeficiency virus enhancer. J Virol. 1989;63(6):2557-66.
39. Kitamura A, Takahashi K, Okajima A, Kitamura N. Induction of the human gene for p44, a hepatitis-C-associated microtubular aggregate protein, by interferon-alpha/beta. Eur J Biochem. 1994;224(3):877-83.
40. Kim JH, Park SY, Jun Y, Kim JY, Nam JS. Roles of Wnt Target Genes in the Journey of Cancer Stem Cells. Int J Mol Sci. 2017;18(8).
41. Jho EH, Zhang T, Domon C, Joo CK, Freund JN, Costantini F. Wnt/beta-catenin/Tcf signaling induces the transcription of Axin2, a negative regulator of the signaling pathway. Mol Cell Biol. 2002;22(4):1172-83.
42. Lustig B, Jerchow B, Sachs M, Weiler S, Pietsch T, Karsten U, et al. Negative feedback loop of Wnt signaling through upregulation of conductin/axin2 in colorectal and liver tumors. Mol Cell Biol. 2002;22(4):1184-93.
43. Yan D, Wiesmann M, Rohan M, Chan V, Jefferson AB, Guo L, et al. Elevated expression of axin2 and hnkd mRNA provides evidence that Wnt/beta -catenin signaling is activated in human colon tumors. Proc Natl Acad Sci U S A. 2001;98(26):14973-8.
44. Ramakrishnan AB, Cadigan KM. Wnt target genes and where to find them. F1000Res. 2017;6:746.
45. Barker N, van Es JH, Kuipers J, Kujala P, van den Born M, Cozijnsen M, et al. Identification of stem cells in small intestine and colon by marker gene Lgr5. Nature. 2007;449(7165):1003-7.
46. He TC, Sparks AB, Rago C, Hermeking H, Zawel L, da Costa LT, et al. Identification of c-MYC as a target of the APC pathway. Science. 1998;281(5382):1509-12.
47. Shtutman M, Zhurinsky J, Simcha I, Albanese C, D'Amico M, Pestell R, et al. The cyclin D1 gene is a target of the beta-catenin/LEF-1 pathway. Proc Natl Acad Sci U S A. 1999;96(10):5522-7.
48. Katoh Y, Katoh M. Hedgehog target genes: mechanisms of carcinogenesis induced by aberrant hedgehog signaling activation. Curr Mol Med. 2009;9(7):873-86.
49. Danaher P, Warren S, Dennis L, D'Amico L, White A, Disis ML, et al. Gene expression markers of Tumor Infiltrating Leukocytes. Journal for ImmunoTherapy of Cancer. 2017;5(1):18.
50. Nguyen QH, Pervolarakis N, Blake K, Ma D, Davis RT, James N, et al. Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. Nat Commun. 2018;9(1):2028.
51. Barneda D, Planas-Iglesias J, Gaspar ML, Mohammadyani D, Prasannan S, Dormann D, et al. The brown adipocyte protein CIDEA promotes lipid droplet fusion via a phosphatidic acid-binding amphipathic helix. Elife. 2015;4:e07485.
52. Ussar S, Lee KY, Dankel SN, Boucher J, Haering MF, Kleinridders A, et al. ASC-1, PAT2, and P2RX5 are cell surface markers for white, beige, and brown adipocytes. Sci Transl Med. 2014;6(247):247ra103.
53. Min SY, Desai A, Yang Z, Sharma A, DeSouza T, Genga RMJ, et al. Diverse repertoire of human adipocyte subtypes develops from transcriptionally distinct mesenchymal progenitor cells. Proceedings of the National Academy of Sciences. 2019;116(36):17970-9.

54. Shan T, Liu W, Kuang S. Fatty acid binding protein 4 expression marks a population of adipocyte progenitors in white and brown adipose tissues. Faseb j. 2013;27(1):277-87.
55. Bruinsma S, Burgess J, Schlingman D, Czyz A, Morrell N, Ballenger C, et al. Bead-linked transposomes enable a normalization-free workflow for NGS library preparation. BMC Genomics. 2018;19(1):722.
56. Stringer C, Wang T, Michaelos M, Pachitariu M. Cellpose: a generalist algorithm for cellular segmentation. Nature Methods. 2021;18(1):100-6.
57. cellpose-turbo/cellpose/flow2msk.py [cited 2021 Oct 14]. Available from: https://github.com/Image-Py/cellpose-turbo/blob/master/cellpose/flow2msk.py.

Supplemental Materials

Supplemental Tables

Table S1. Genes for 980-plex RNA panel

Table S2. 1,210 encoding barcodes for 980-plex RNA panel

Table S3. SMI results summary of lung tissue samples

Table S4. Lung tissue samples information

Table S5. Contents of 80-plex protein panel

Supplemental Figures

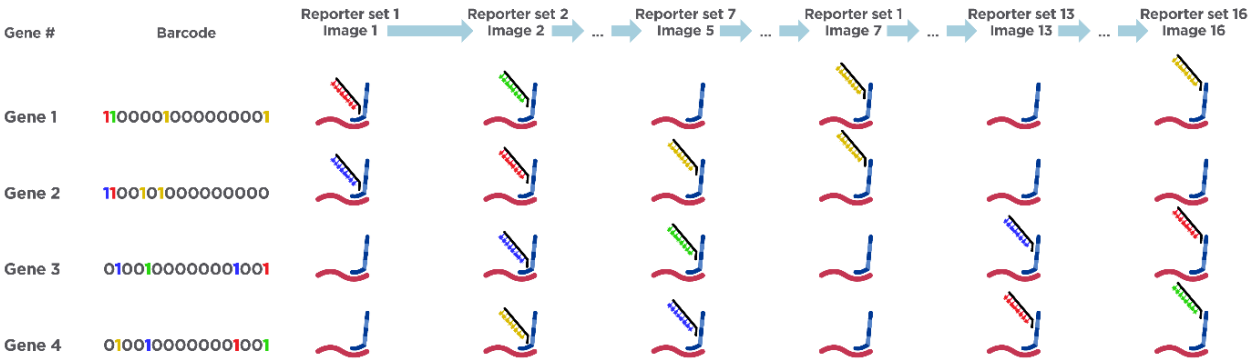


Figure S1. Schematic Depiction of SMI Encoding Design and Readout Process. Each gene is assigned with a unique 16-digit barcode with 4 "on" spots (labeled as "1") and 12 "off" spots (labeled as "0"). Each digit of the barcode indicates the presence of reporter that is associated with the target in that reporter hybridization round. "1" means that there is a reporter hybridizing to the ISH probe of the target in that hybridization round, and its color indicates the fluorophore of the hybridized reporter. "0" means that there is no reporter in that hybridization round binds to the target ISH probe, and that target should be silenced or blank in that round of imaging. For each gene, 4 reporters will bind to the 4 designated reporter landing domains on ISH probe sequentially throughout the 16 rounds of cyclic reporter readout.

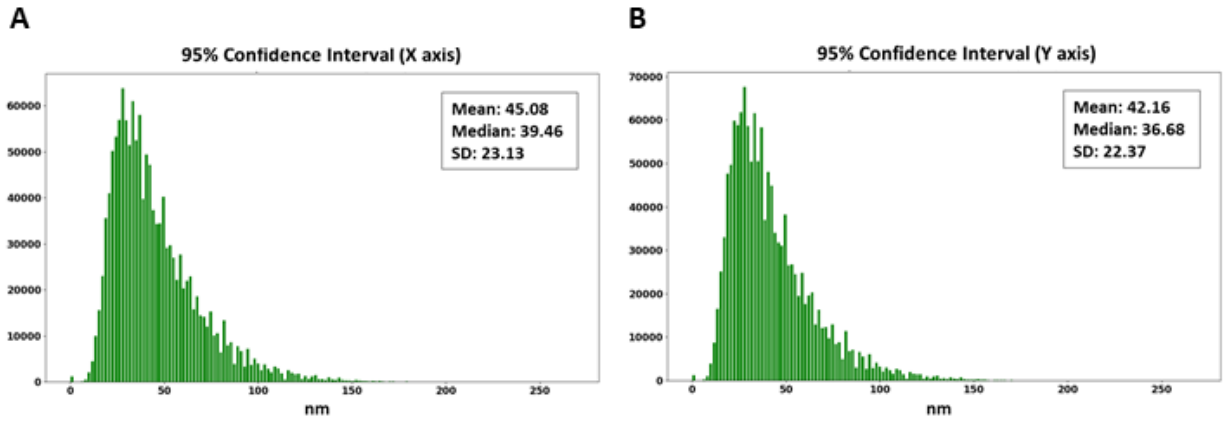


Figure S2. SMI Overall Localization Precision. The data shows an overall localization accuracy of 42-to-45 nm on averaged over the entire experiment in X and Y planes.

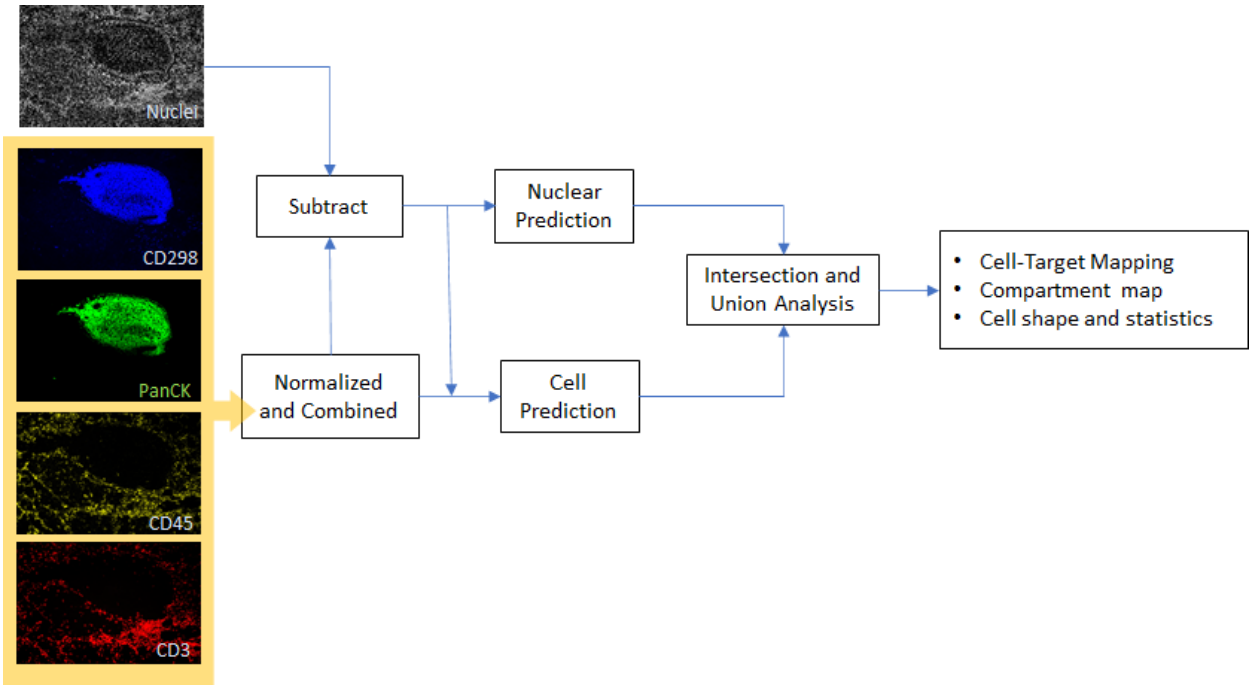


Figure S3. Cell Segmentation Pipeline. Membrane images are combined with normalization and subtracted with the nuclear image to enhance gaps between nuclei. Results are fed into the prediction network to segment the nuclei and cell boundaries. Results from segmentation are used to map the target to its corresponding cell and compartment within the cell.

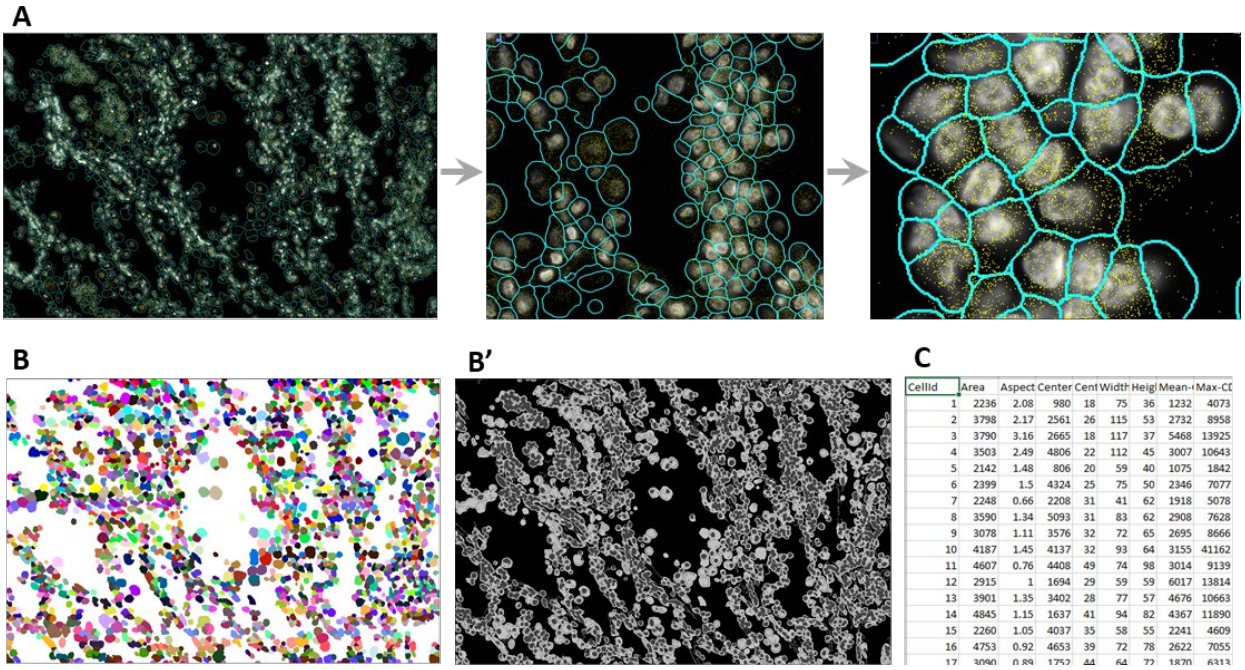


Figure S4. Cell Segmentation Results in Non-Small Cell Lung Carcinoma (NSCLC) Tissue. A. Segmentation overlay output: Segmentation boundaries (cyan) overlaid with nuclear image and transcripts (plotted in yellow) at different zoom levels. **B.** Cell label output: each cell is marked with unique cell ids (shown as different colors). **B'.** Cell compartment map: an output to show specific compartments of nuclei, membrane, and cytoplasm (marked with different gray level values). **C.** Cell statistics table output: area, aspect ratio, bounding box, and intensity statistics per channel are listed for each cell (id).

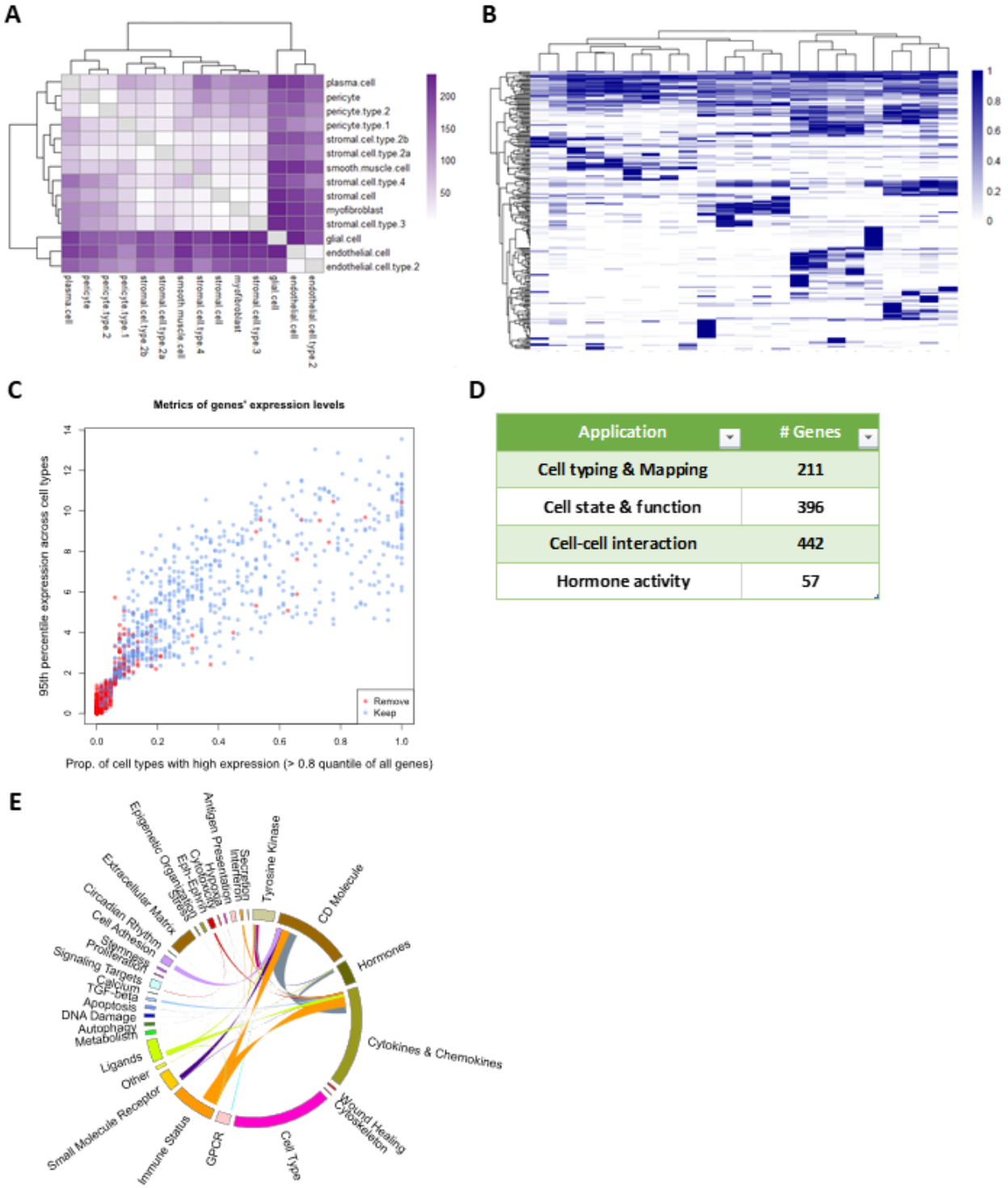


Figure S5. Content Design for a 980-lex Single Cell Panel for Measuring Cell Type, Cell State, and Cell-Cell Interactions. **A.** Example of starting point for gene selection, demonstrating which cell types were selected for additional discrimination after accounting for biologically relevant genes. **B.** Example of cell type profiles over complete gene list (per kidney scRNA-seq). Most genes discriminate between cell types

(genes on y-axis, cell types on x-axis). **C.** Using scRNA-seq average cell type profiles from 76 Human Cell Landscape datasets, we scored genes for expression level across cell types based on level of expression and consistency of expression across cell types. Genes selected based on scoring metrics are shown. Other genes were considered required based on strong biological relevance. **D.** Final gene counts by purpose, categories not mutually exclusive. **E.** Circos diagram summarizing the overlapping genes between the panel's gene sets. Cell type is a set designed to increase the contrast between cell types beyond the biological content.

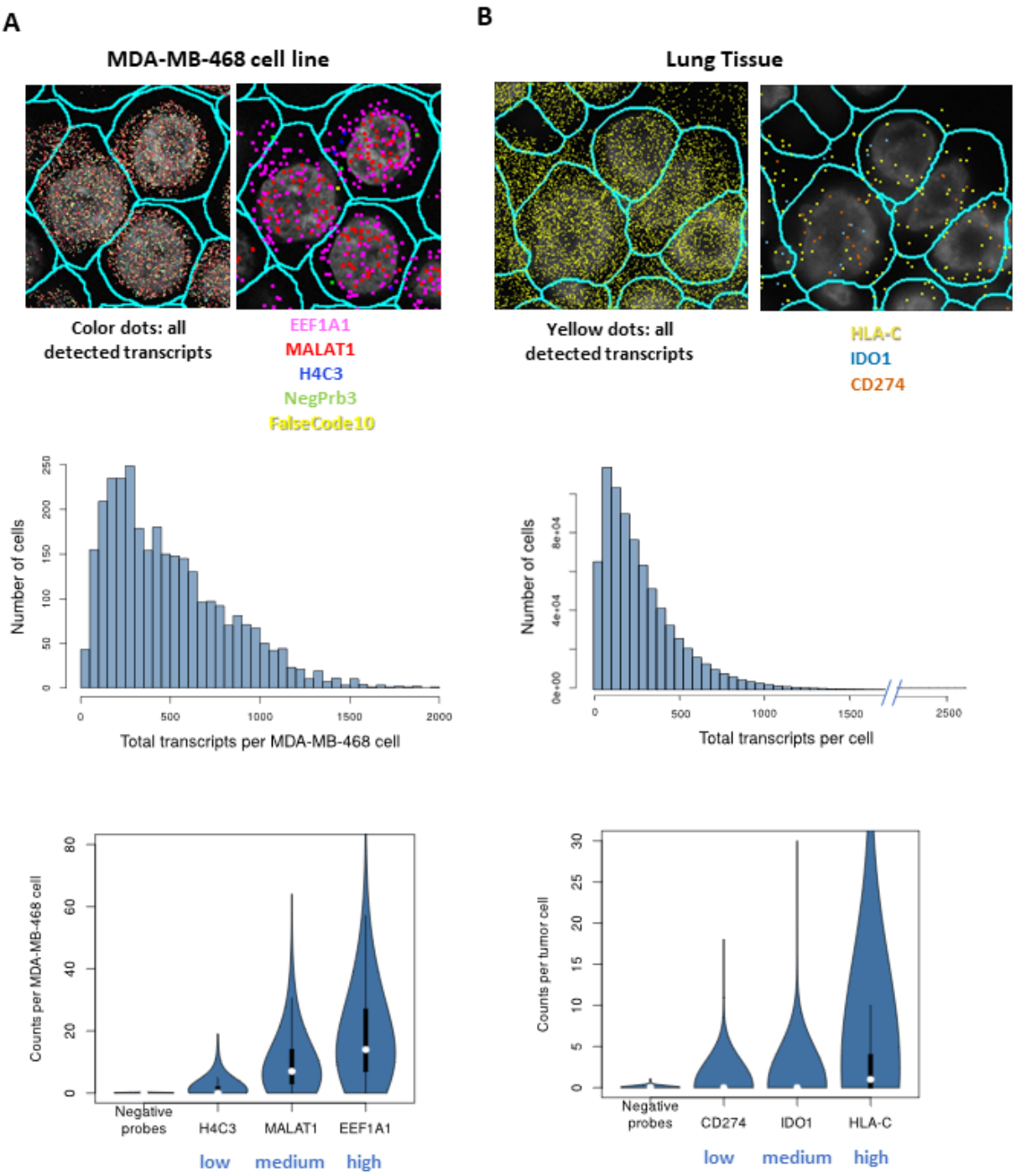


Figure S6. Single-Cell Distribution and Detection Sensitivity of Low, Medium, and High Expressers. A. FFPE MDA-MB-468 cell line. Single-cell distribution of total transcripts per cell is shown. Overlay images and average counts per cell of genes representing low (H4C3), medium (MALAT1), high (EEF-1A1) expressers in this cell line and negative probes. **B.** FFPE lung tissue. Single-cell distribution of total transcripts per cell is shown. Overlay images and average counts per cell of genes representing low (CD274), medium (IDO1), high (HLA-C) expressers in the lung tissue and negative probes.

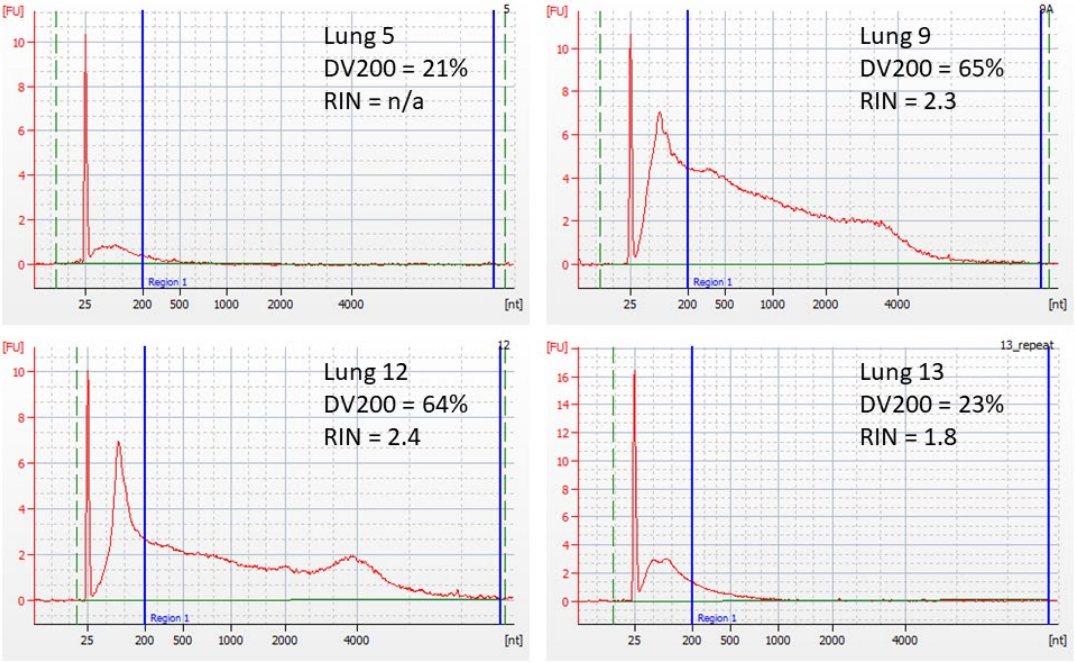


Figure S7. RNA Quality of Lung Samples. DV200 and RIN scores are determined using Agilent BioAnalyzer 2100 Expert Software B.02.09. For DV200, Region 1 selection is from 200 nt to about 8,000 nt using Smear Analysis.

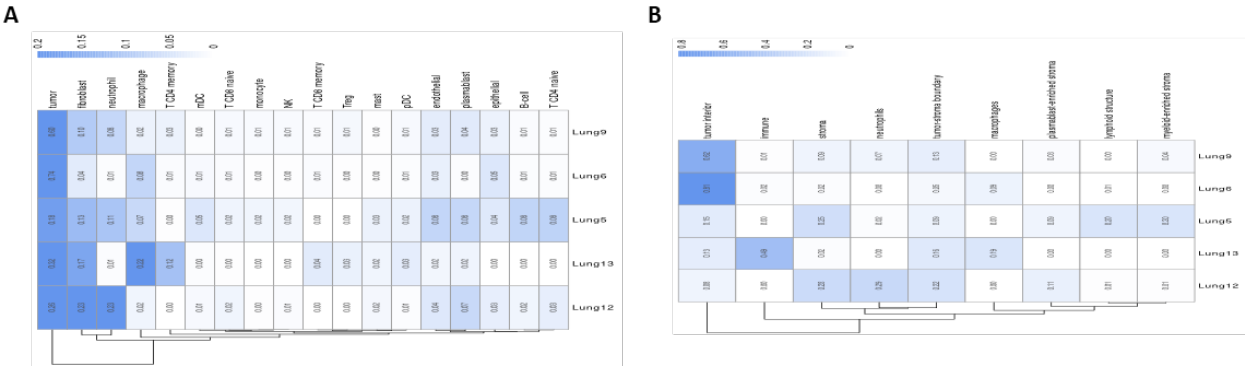


Figure S8. Extracting Tissue-Level Characteristics. **A.** Abundance of cell types within each tumor. **B.** Abundance of microenvironment niches within each tumor.

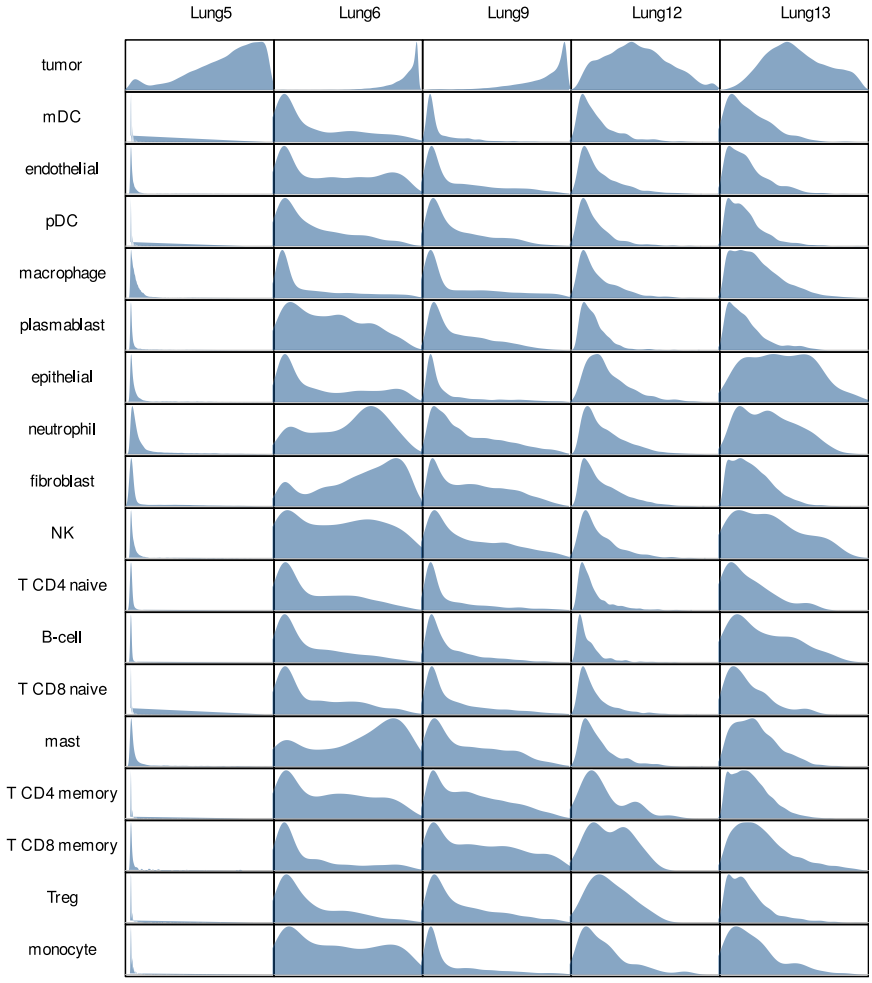


Figure S9. Invasiveness of Each Cell Type Within Each Tissue. Cells are scored for the frequency of tumor cells within their 100 nearest neighbors. Shapes show the density of the score within each cell type.

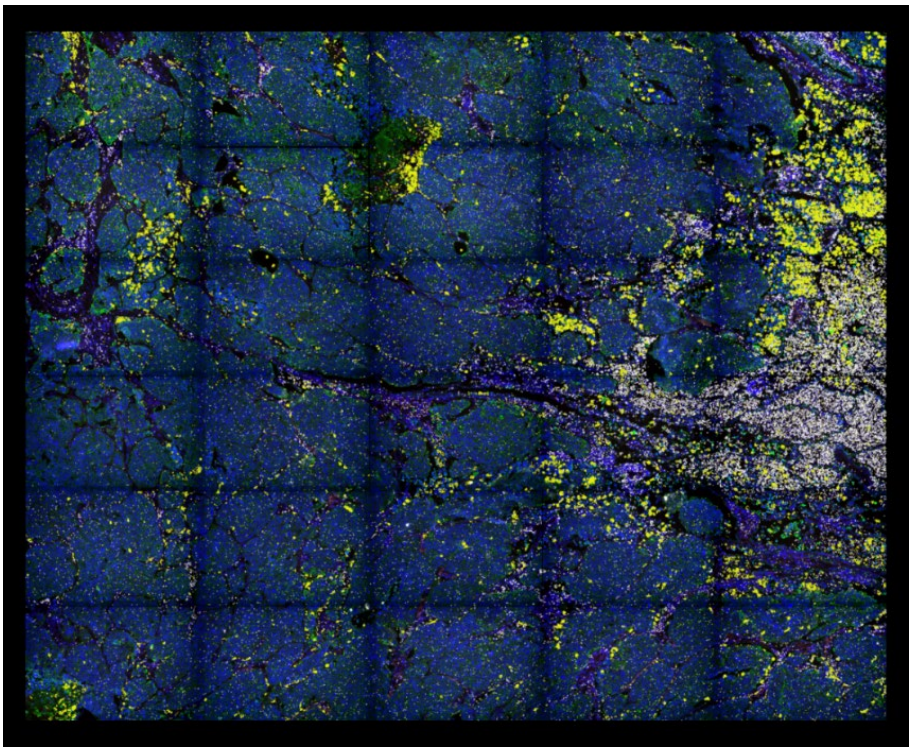


Figure S10. Highlighting Transcripts of SPP1 and HLA-DQA1 in Lung 6. SPP1 transcripts are shown as yellow dots; HLA-DQA1 transcripts are shown as white dots.

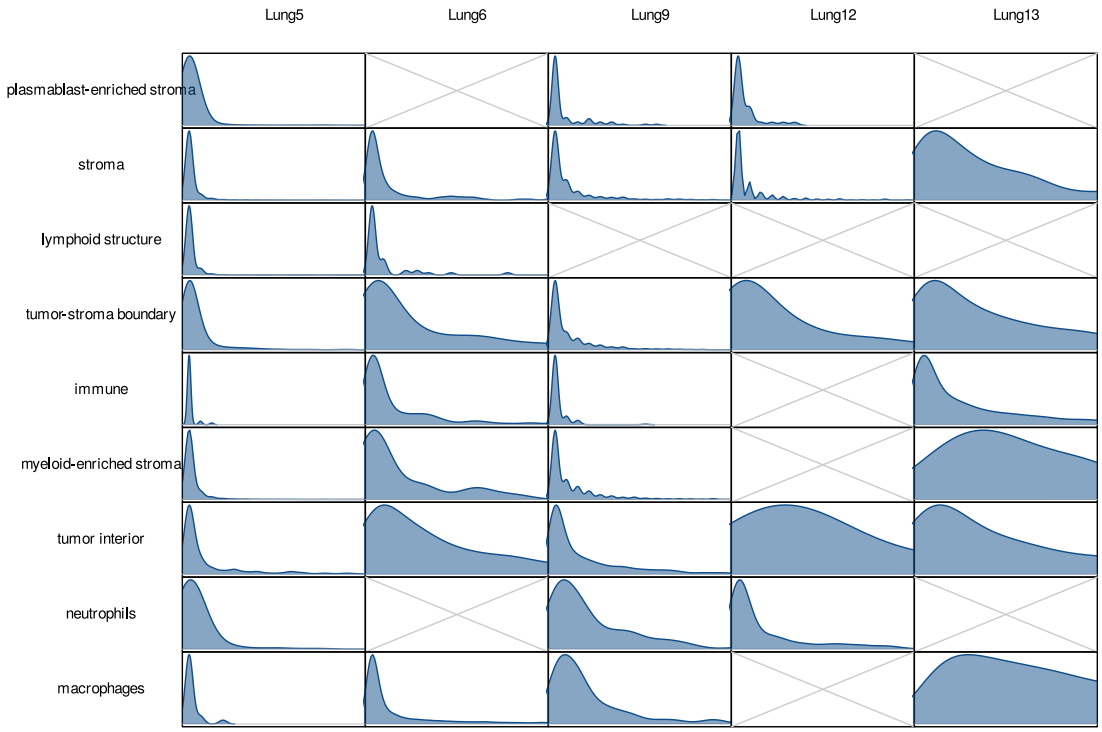


Figure S11. Expression of SPP1 Across Tumors and Niches. Polygons show densities of SPP1 transcripts per macrophage; horizontal axes are truncated at 10 counts. Only tumors and niches with > 20 macrophages are shown.

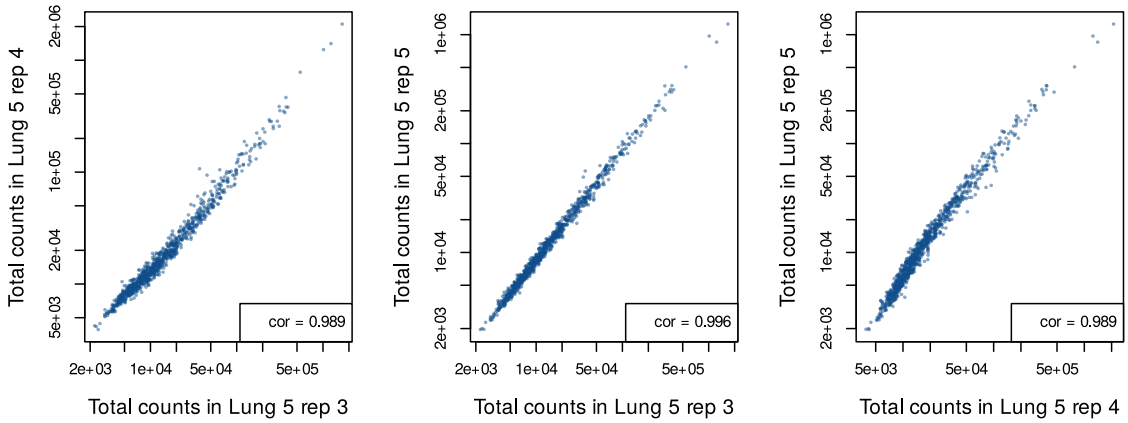


Figure S12. Concordance of Bulk Expression Profiles from Three Replicates. For each of the three replicates, total transcripts of each gene were recorded, and these bulk profiles are compared across replicates. Correlation is calculated on log-transformed data.

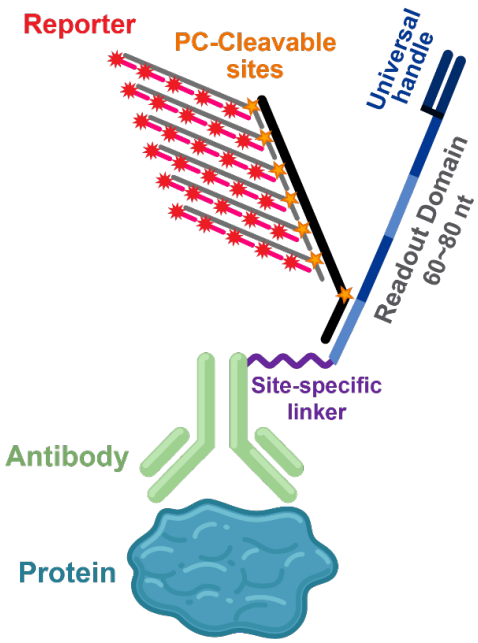


Figure S13. Protein Detection Probe. The protein detection probe is readily adapted from the RNA detection probe shown in Figure 1. For protein detection, the readout domain contains an additional universal handle and is conjugated to the antibody via a site-specific linker.

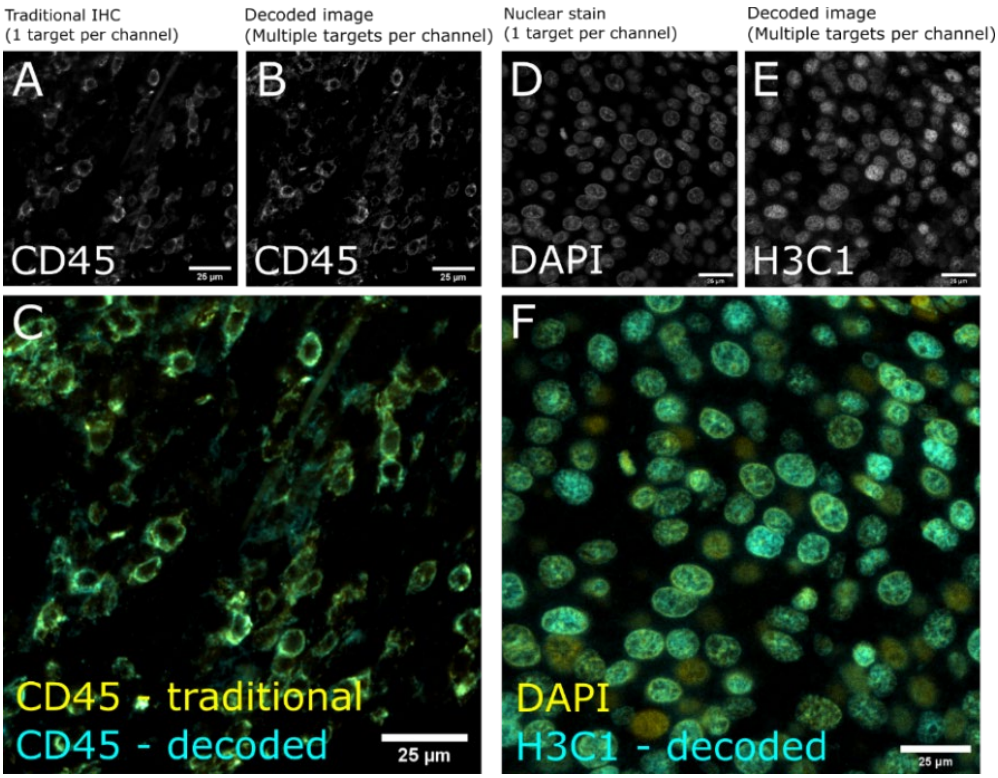


Figure S14. Concordance Between Protein Localization Patterns Detected in the 80-plex SMI Protein Assay with Ground Truth Data. A. CD45 localization detected using traditional single-channel immunohistochemistry compared with **B.** CD45 localization as decoded from the SMI 80-plex protein assay. **C.** Overlay of (A) and (B). **D.** DAPI staining of nuclei. **E.** Histone H3 localization as decoded from the 80-plex protein assay. **F.** Overlay of (D) and (E). Note that (D) is a single Z-plane while (E) is a projection of 6.4 µm.

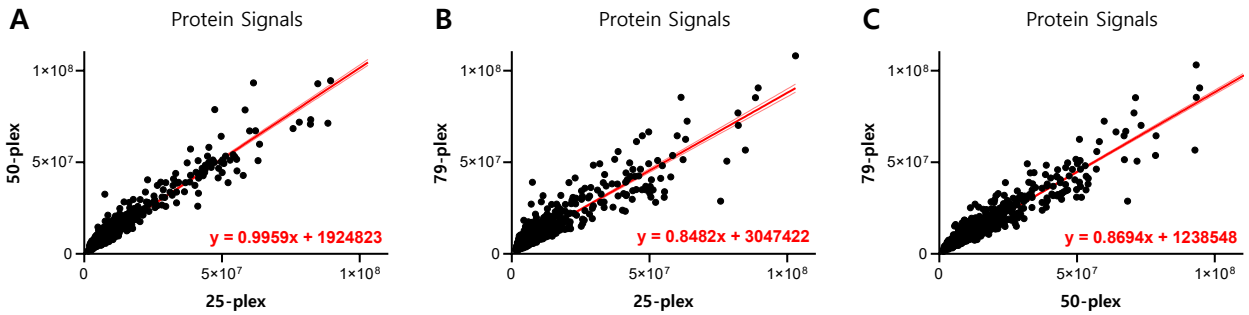


Figure S15. SMI Protein Assay Delivers Consistent Results as the Plex Size Increases. Protein expression in 35 different cell lines was evaluated by SMI protein assay. The number of antibodies simultaneously detected ranged from 25 to 79 antibodies. Protein expression values for each cell line were normalized by cell count, and pairwise comparisons of shared targets were conducted between **A.** 25-plex and 50-plex assays, **B.** 25-plex and 79-plex assays, and **C.** 50-plex and 79-plex assays. Simple linear regression lines (red) show the overall linear relationship among the plotted data points. As the difference in plex increases, Pearson R coefficients remained high ($r > 0.90$).

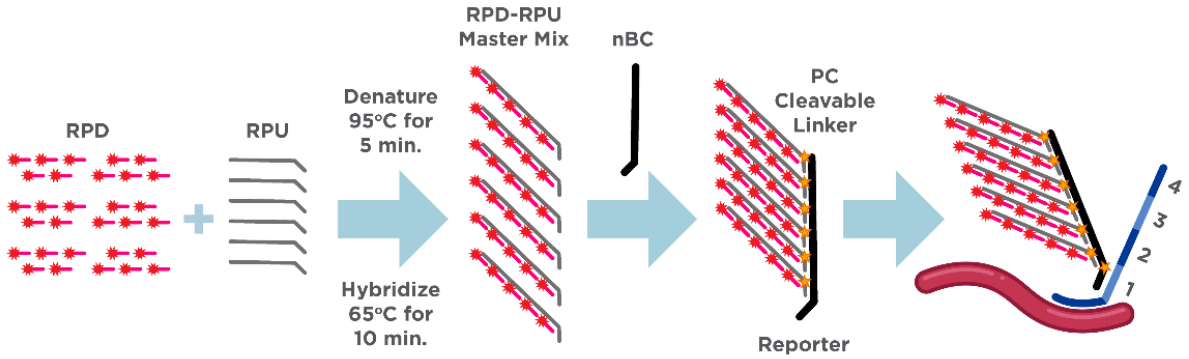


Figure S16. Reporter Readout Probe Design and Assembly. RPD and RPU are denatured to ensure all molecules are unhybridized and then allowed to hybridize to produce the RPD-RPU Master Mix. This Master Mix is hybridized to the nBC to form a single-color reporter. These reporters are pooled by the hybridization region in the probe. Each set of pools contains four reporters with four different fluorophores (*e.g.*, blue, red, yellow, green reporters for spot 1).