MAPS: Pathologist-level cell type annotation from tissue images through machine learning

Muhammad Shaban[1,2,3,4,*], Yunhao Bai[5,*], Huaying Qiu[6,*], Shulin Mao[6], Jason Yeung[6], Yao Yu Yeo[6], Vignesh Shanmugam[1,4], Han Chen[5], Bokai Zhu[5], Garry P. Nolan[5], Margaret A. Shipp[7], Scott J. Rodig[1,7], Sizun Jiang[4,6,8,✉], and Faisal Mahmood[1,2,3,4,✉]

[1]Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States
[2]Department of Pathology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, United States
[3]Data Science Program, Dana-Farber Cancer Institute, Boston, MA, United States
[4]Broad Institute of Harvard and MIT, Cambridge, MA, United States
[5]Department of Pathology, Stanford University School of Medicine, Stanford, CA, United States
[6]Center for Virology and Vaccine Research, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, United States
[7]Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, United States
[8]Department of Pathology, Dana Farber Cancer Institute, Boston, MA, United States
[*]Those authors contributed equally to this paper
[†]Corresponding Authors

Highly multiplexed protein imaging is emerging as a potent technique for analyzing protein distribution within cells and tissues in their native context. However, existing cell annotation methods utilizing high-plex spatial proteomics data are resource intensive and necessitate iterative expert input, thereby constraining their scalability and practicality for extensive datasets. We introduce MAPS (Machine learning for Analysis of Proteomics in Spatial biology), a machine learning approach facilitating rapid and precise cell type identification with human-level accuracy from spatial proteomics data. Validated on multiple in-house and publicly available MIBI and CODEX datasets, MAPS outperforms current annotation techniques in terms of speed and accuracy, achieving pathologist-level precision even for challenging cell types, including tumor cells of immune origin. By democratizing rapidly deployable and scalable machine learning annotation, MAPS holds significant potential to expedite advances in tissue biology and disease comprehension.

Multiplexed imaging | Proteomics | Machine learning | Cell Annotation

Correspondence: ***sjiang3@bidmc.harvard.edu, faisalmahmood@bwh.harvard.edu***

Introduction

The precise delineation of cellular subtypes is crucial for elucidating structural and functional intricacies of biological tissues, within their native context. Recent advances in high-plex spatial proteomics have facilitated the simultaneous imaging of over 50 markers, thereby offering invaluable insights into protein expression and distribution within cellular and tissue architectures for phenotypic and functional investigations (1, 2). Nonetheless, accurate cellular annotation predicated on this wealth of data presents formidable challenges, stemming primarily from constraints in highly precise cell segmentation (3), lateral spillover of markers in tightly packed tissues (4), presence of tissue-level and patient-level variability, and heterogeneous expression patterns (1, 2, 5) . Existing approaches for cell annotation are contingent upon unsupervised clustering techniques, necessitating subsequent manual curation, a process that can be markedly labor-intensive and requires domain-specific expertise. Achieving higher annotation accuracies can thus be an arduous process due to the iterative steps required (6). Therefore, there is a need for automated methods that can accurately classify cells based on their spatial proteomics data. Promising automated approaches developed recently include probabilistic inferential approaches (7, 8), and convolutional neural networks (9, 10). However, these approaches may be lower in accuracy, or be computationally expensive, requiring more memory and taking longer times to train and infer.

Therefore, a computationally lightweight and fast automated cell classification method, while achieving human-level accuracy, is required to improve the efficiency and scalability of spatial proteomics data analysis. We present here **MAPS** (Machine learning for Analysis of Proteomics in Spatial biology), a machine learning package that enables highly accurate and fast cell annotation with the highest in-class performance when benchmarked across multiple spatial proteomics platforms. MAPS can significantly enhance our understanding of complex biological systems and facilitate the discovery of novel biological processes *in situ*.

Results

Development of MAPS and initial application to an in-house curated cHL MIBI dataset. Herein, we postulated that a feed-forward neural network would be an efficient and robust model for rapid and accurate cell phenotyping. This model, MAPS, predicts the cell class from a set of user-defined classes using the expression of a cell for N markers, and its area in pixels (**Fig. 1A**). MAPS employs four fully connected hidden layers with ReLU activation function and dropout layers, followed by a classification layer with softmax function. MAPS accurately predicted the cell phenotypes in healthy and diseased tissues, as exemplified by a MIBI dataset of canonical Hodgkin's Lymphoma (cHL) [1669853 cells, 13 cell types] (**Fig. 1B**, **Supp Fig. 1A**). All ground truth annotation data was generated through traditional iterative clustering and visual inspection, followed by final inspection by a board-certified pathologist (S.J.R.). All questionable clusters were subject to further clustering based on the key markers that were present, and difficult cell types,

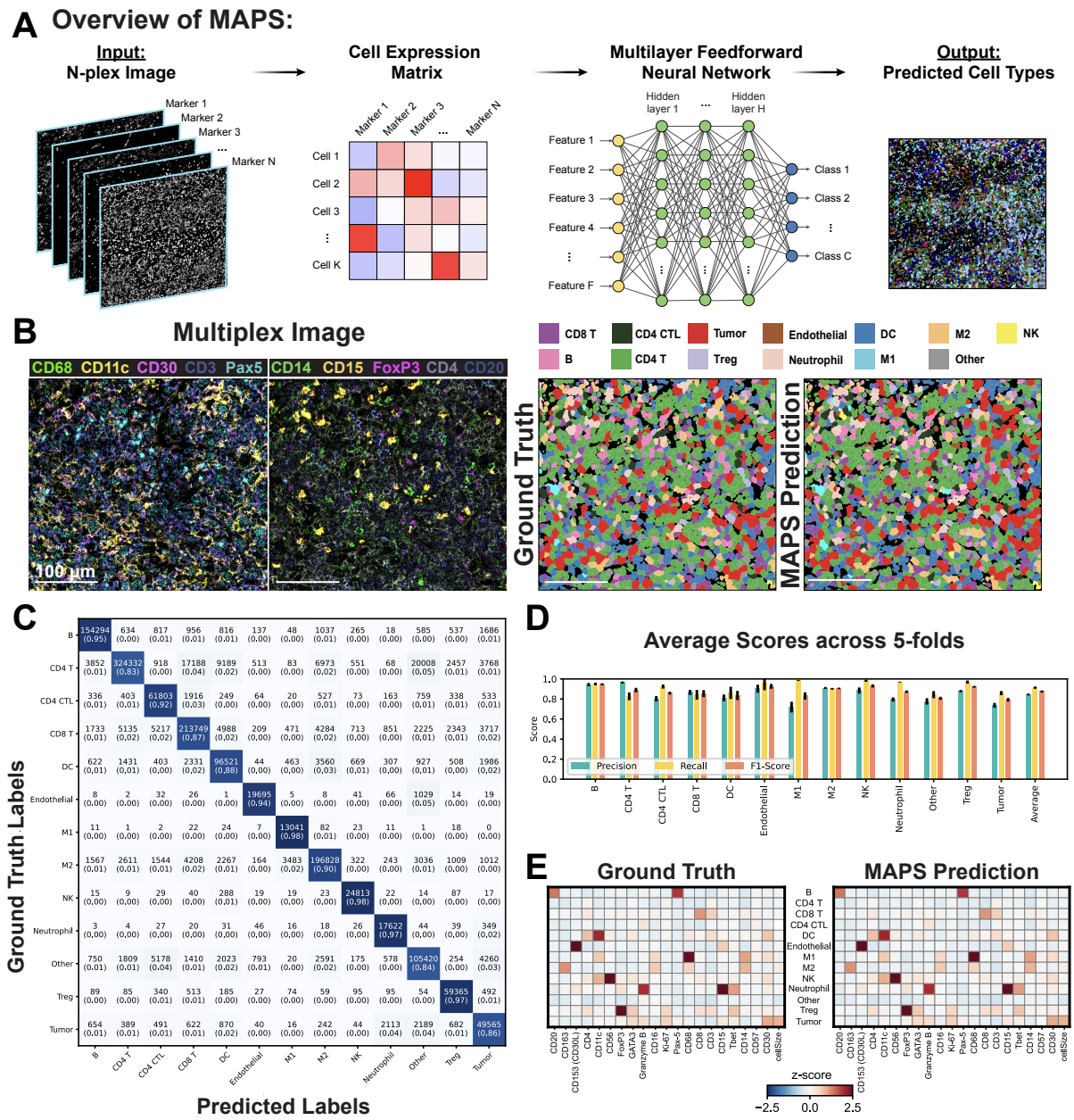

Figure 1: Overview of MAPS architecture and its performance on cHL1 (MIBI) dataset across 5-folds cross validation. (A) Schematic of MAPS for machine learning based cell phenotyping. MAPS takes a multiplex image as input and converts it into a cell expression matrix after preprocessing, which is then fed into a feedforward neural network for cell type prediction from a predefined list of classes. (B) A representative FOV of a multiplexed image used for cell phenotyping. Cell phenotype maps generated via manual annotation (Ground Truth) and MAPS (MAPS Prediction) are shown for visual comparison. (C) Confusion matrix of MAPS predictions. Numbers in parentheses indicate the percentage of cells with respect to total cells in the corresponding row/class. (D) Average precision, recall, and F1-score of MAPS predictions across five folds. Error bars represent ±1 standard deviation. (E) Average cell marker expression matrix for each cell type generated using ground truth labels (left) and MAPS prediction (right).

such as Reed–Sternberg tumor cells in cHL, were then subject to manual inspection and further annotation as needed to generate the "ground truth" reference (**Supp Fig. 1B**, further expanded in Material & Methods). Next, we evaluated MAPS performance using the following metrics: confusion matrix, precision, recall, and F1-scores from a 5-fold validation (**Figs. 1C & D**; see Material & Methods for more details). The mean cell expression matrix of various phenotypic markers for each cell type in the ground truth and predictions had high concordance (**Fig. 1E**). MAPS demonstrated consistently high accuracy in predicting the cell type from spatial proteomics datasets.

Benchmarking comparisons of MAPS against other methods and on other spatial proteomics data. We sought next to demonstrate real world practicality of MAPS, and its performance against other state-of-the-art approaches, ASTIR (7) and CellSighter (10). We collected and annotated in-house data from 1) MIBI on cHL using a first cohort (cHL 1; 1669853 cells), 2) MIBI on cHL using a second co-

hort (cHL 2; 192795 cells), and 3) CODEX on cHL (145161 cells). MAPS, ASTIR and CellSighter were trained on the same ground truth data generated on the aforementioned dataset in the same manner (see Material & Methods), and the resulting phenotype maps visualized (**Fig. 2A** and **Supp Fig. 2A**). The analysis of the F1 scores across all cell types indicated the consistently highest performance of MAPS across all datasets, followed by CellSighter and ASTIR (**Fig. 2B**, **Supp Fig. 2B & C**). Given the high performance of MAPS and CellSighter, we next computed precision-recall curves and average precision per class to gain further insights on the model differences (**Supp Fig. 2D**). MAPS consistently outperformed CellSighter on all three datasets for all cell types, with average precision per class ranging from 0.82 to 0.99 for MAPS, and 0.39 to 0.93 for CellSighter. MAPS demonstrated consistent performance across all datasets, while CellSighter performed better on cHL (CODEX) than the other two datasets (cHL 1 MIBI and cHL 2 MIBI). We postulated that this performance difference could be attributed to the different data-splitting strategies. The cHL (CODEX) dataset, consisting of a large single tissue image from one patient, was split at the cell level, which can lead to bias and overfitting in the machine learning model. This is because adjacent cells in the same image may have been split between the training and validation sets, potentially leading to high overlap in the distribution of cells in training and validation sets. This can artificially result in higher performance in the validation set, which may not generalize well to new samples. In contrast, the other two datasets consist of multiple regions and patient cases, and were split at the case level, which prevents this issue of information leakage, thus resulting in a more realistic real-world performance. Details of these datasets are further elaborated in the Material & Methods.

Given potential differences in optimal hyperparameter selection between models outside this study's scope, we further benchmark MAPS performance on the same public dataset from a colorectal cancer (CRC) study using CODEX (11). We used the reported CellSighter results from this same CRC CODEX data. Quantification of the average F1 score across all cell types showed the highest performance for MAPS across all datasets, followed by CellSighter and ASTIR, respectively (**Fig. 2B**). Detailed delineation of F1 score on each cell type showed the same trend in performance level (**Supp Fig. 2B** and **2C**).

MAPS is highly efficient in computational resource usage. Given how neural network models can be resource intensive, we next quantified the level of computational resource usage between MAPS and CellSighter. Here, we used the cHL (CODEX) dataset due to its relatively small size yet diverse number of cell type representations. We observed comparable total run time and GPU memory utilization between MAPS and ASTIR, with substantially higher values for CellSighter. Memory utilization was similar between MAPS and CellSighter, with lower values for ASTIR (**Fig. 2C**). Our results highlight the well-balanced computational efficiency and rapid performance of MAPS, relative to its top-in-class accuracy for cell type annotation.

Discussion

In this study, we introduced a new method, MAPS, for pathologist-level accuracy in cell annotation from spatial proteomics data. Our results demonstrate that MAPS outperforms existing state-of-the-art methods in terms of both accuracy and computational efficiency while showing cross-platform compatibility (MIBI and CODEX). Specifically, MAPS achieved a significantly higher F1-score, precision, and recall compared to other methods, demonstrating its ability to accurately predict cell types from spatial proteomics data.

One of the key strengths of MAPS is its use of a feed-forward neural network architecture, which allows for the efficient processing of high-dimensional spatial proteomics data. Additionally, the use of ReLU activation functions introduces non-linearity, improving the ability of the model to capture complex relationships between the input features and the cell types. The inclusion of dropout layers during training also helps to prevent overfitting and improves generalization performance.

Another important feature of MAPS is its ability to handle imbalanced datasets, a common feature of spatial proteomics data due to the unequal distribution of cell types within tissues. Our results show that MAPS outperforms other methods in terms of average precision, a metric that is particularly useful for imbalanced datasets.

In addition to its superior performance, MAPS is also computationally efficient, with a training time that is orders of magnitude faster than existing supervised methods. This makes it a valuable tool for large-scale analysis of spatial proteomics data, where computational efficiency is crucial. We designed MAPS to be incorporated into current spatial proteomics workflows, where it can accelerate the confident annotation process from a smaller, curated "ground truth" dataset (**Supp Fig. 1B**).

In conclusion, our results demonstrate that MAPS is a highly accurate and computationally efficient method for cell annotation from spatial proteomics data. Its superior performance and efficiency make it a valuable tool for the analysis of large-scale spatial proteomics datasets, with potential applications in a wide range of biological and biomedical research areas. We additionally release the MAPS package on github (https://github.com/mahmoodlab/MAPS), along with all the data generated for this paper, including the spatial proteomics data generated here and associated ground truth annotations. We are confident such a resource can be leveraged upon by the community to accelerate future approach for cell type annotation in tissue spatial-omics, and beyond.

Materials & Methods

Section 1: dataset acquisition.

Human Tissue Acquisition and Patient Consent. Formalin-fixed paraffin-embedded (FFPE) excisional biopsies from 23 patients with newly diagnosed cHL, and one reactive lymph node were retrieved from the archives of

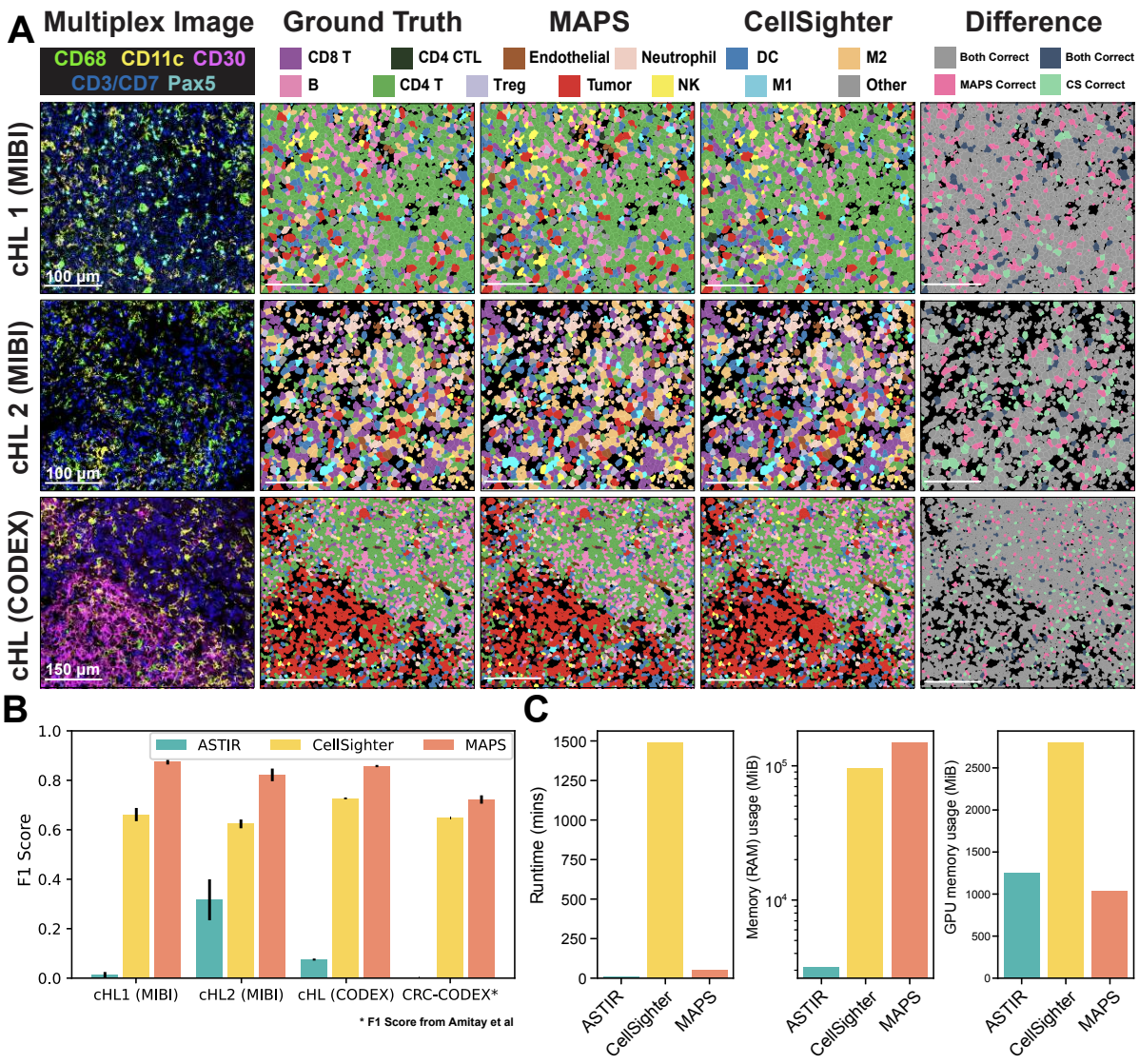

Figure 2: Visual and quantitative comparison of MAPS performance with its counterparts. (A) Comparison of MAPS and CellSighter performances across four multiplex image datasets. The last column indicates differences in cell predictions between these two methods. Row 1: Representative cHL FOV from a cHL patient cohort (cHL 1) acquired via the MIBI. Row 2: Representative cHL FOV from another cHL patient cohort (cHL 2) acquired via the MIBI. Row 3: Representative cHL FOV from a separate cHL tissue acquired via the CODEX. (B) Comparison of three cell phenotyping methods across four datasets using average F1-score across 5-folds. Error bars represent ±1 standard deviation. For the publicly available CRC (CODEX) dataset, the CellSighter results were taken from the original publication (10). ASTIR was not applied to the CRC dataset. (C) Evaluation of computational resources required for each of the models tested on the cHL (CODEX) dataset. Left: Total runtime (including model training and cell type inference). Middle: Median memory (RAM) usage, the y-axis here is on a log logarithmic scale. Right: Median GPU memory usage.

Brigham and Women's Hospital (Boston, MA) with institutional review board approval (IRB# 2010P002736). All tumor regions were annotated by V.S. and S.J.R.

Antibody Conjugation and Panel. Lanthanides conjugated antibodies for MIBI were acquired as previously described (12) using the Maxpar X8 Multimetal Labeling Kit (Fluidigm, 201300) and Ionpath Conjugation Kits (Ionpath, 600XXX) with slight modifications to manufacturer protocols. In short, 100 μg BSA-free antibody was first washed with the conjugation buffer, then reduced using 4 μ M (final concentration) of TCEP (Thermo Fisher Scientific, 77720) to reduce the thiol groups for 30 min in a 37 °C water bath. The reduced antibody was mixed and incubated with Lanthanide-loaded polymers for 90 min in a 37 °C water bath, then washed for 5 times with an Amicon Ultra filter (Millipore Sigma, UFC505096). Resulting conjugated antibodies were then buffered with at least 30% v/v Candor Antibody Stabilizer (Thermo Fisher Scientific, NC0414486) including 0.02% w/v of sodium azide, and stored at 4 °C until usage.

Oligo conjugation to antibodies for CODEX was performed as previously described (5). In short, 100 μg BSA-free antibody was reduced using 2.5 mM of TCEP at RT for 30 min to reduce the thiol groups. Maleimide-labeled oligos are resuspended in High-salt Buffer C (1 M NaCl) and incubated with the reduced antibodies at RT for 2 h. The resulting conjugated antibodies are then washed for 3 times in high salt PBS (0.9 M NaCl) in a 50 kDa centrifugal column (Sigma, UFC505096), buffered with at least 30% v/v Candor Anti-

body Stabilizer (Thermo Fisher Scientific, NC0414486) supplemented with 0.02% w/v of sodium azide, and stored at 4°. The antibody panels can be found in **Supplementary Table 1**.

Gold Slide Preparation. The protocol of preparing gold slides has been described previously (13–15). In short, Superfrost Plus glass slides (Thermo Fisher Scientific, 12-550-15) were first soaked and briefly supersonicated in a ddH_2O diluted with dish detergent, cleaned by using Microfiber Cleaning Cloths (Care Touch, BD11945) then rinsed in flowing water to remove any remaining detergent. After that, the slides were air-dried with a constant stream of air in the fume hood. The coating of 30 nm of Tantalum followed by 100 nm of Gold was performed by the Microfab Shop of Stanford Nano Shared Facility (SNSF) and New Wave Thin Films (Newark, CA).

Coverslip and Slides Vectabonding. To introduce positive charges for better adhesion of tissue sections onto the surface, pre-cleaned 22x22 mm glass coverslips (VWR, 48366-067) or the e-beam coated gold slides were silanized by VECTABOND Reagent (Vector Labs, SP-1800-7) per the protocol from the manufacturer. The slides were first soaked in neat acetone for 5 min, then transferred into 1:50 diluted VECTABOND Reagent in acetone and incubated for 10 min. After that, slides were quickly dipped in ddH_2O to quench and remove remaining reagents, then tapped on Kimwipe to remove remaining water, the resulting slides were air-dried then stored at room temperature.

MIBI Retrieval and Staining Protocol . The procedure of a general MIBI staining is similar to previously described (6, 13, 16). The FFPE block was sectioned onto Vectabond-treated gold slides by 5 μm thickness. The sections then went through a standard deparaffinization and antigen retrieval process. In brief, slides with FFPE sections were first baked in an oven (VWR, 10055-006) for 1 hour at 70 °, then were transferred into neat xylene and incubated for 2x 10 min. Standard deparaffinization was performed with a linear stainer (Leica Biosystems, ST4020) in the following sequence: 3x neat xylene, 3x 100% EtOH, 2x 95% EtOH, 1x 80% EtOH, 1x 70% EtOH, 3x ddH_2O, 180 s each step with constant dipping, then rest in ddH_2O. Antigen retrieval was then performed at 97 ° for 10 min with Target Retrieval Solution (Agilent, S236784-2) on a PT Module (Thermo Fisher Scientific, A80400012).

After PT Module processing, the cassette with slides and solution was left on the benchtop until it reached room temperature. After a quick 1x PBS rinse for 5 min, the sections were blocked by BBDG (5% NDS, 0.05% sodium azide in 1x TBS IHC wash buffer with Tween 20), then stained at 4 ° in an antibody cocktail for overnight (**Supplementary Table 1**). Subsequently, the samples were quickly rinsed with 1x PBS, then fixed by the Post-fixation buffer (4% PFA + 2% GA in 1x PBS buffer) for 10 min, then quenched with 100 mM Tris HCl pH 7.5, before undergoing a series of dehydration steps on the linear stainer (3x 100 mM Tris pH 7.5, 3x ddH_2O, 1x 70% EtOH, 1x 80% EtOH, 2x 95% EtOH, 3x 100% EtOH, 60 s for each step), before store in a vacuum desiccator until acquisition.

CODEX Retrieval and Staining Protocol . The procedure for CODEX staining is similar to previously described (17). A cHL FFPE section was mounted on a No.1 glass coverslip pre-treated with VECTABOND Reagent (Vector laboratories, SP-1800-7) as described above, and deparaffinized by heating at 70 °C for 1 hour, followed by two 15-min soaks in a xylene bath. The tissue was then manually rehydrated in 6-well plates by incubating in 2x 100% EtOH, 2x 95% EtOH, 1x 80% EtOH, 1x 70% EtOH, and 3x ddH_2O, for 3 min each with gentle rocking. Heat-induced antigen retrieval (HIER) was performed in a coverslip jar containing 1x Dako pH 9 Antigen Retrieval Buffer (Agilent, S2375) while using a PT module filled with 1x PBS; the PT module was set to prewarm to 75 °C, heat to 97 °C for 20 min, before cooling to 65 °C. After HIER, the tissue was washed in CODEX hydration buffer (Akoya Biosciences, 232105) 2x for 2 min and incubated in CODEX staining buffer (Akoya Biosciences, 232106) for 20 min. The tissue was then transferred to a humidity chamber to block with 200 μL of BBDG while being photobleached with a custom LED array for 2 hours (see below), then stained at 4 °C in an antibody cocktail overnight.

The blocking buffer was prepared by combining 180 μL of BBDG block, 10 μL of oligo block, and 10 μL of sheared salmon sperm DNA. The BBDG block was prepared by mixing 5% donkey serum, 0.1% Triton X-100, and 0.05% sodium azide in 1x TBS IHC Wash buffer with Tween 20 (Cell Marque, 935B-09). The oligo block was prepared by mixing 57 different custom oligos (IDT) to create a master mix with a final concentration of 0.5 μM per oligo. The sheared salmon sperm DNA was used directly from its original 10 mg/ml stock (ThermoFisher, AM9680). To create a humidity chamber, an empty pipette tip box was filled with ddH_2O and wet paper towels and then placed on top of a cool box (Corning, 432021) containing an ice block. Two happy lights (Best Buy, 6460231) were leaned against either side of the humidity chamber, and an LED grow light (Amazon, B07C68N7PC) was positioned above. Staining antibodies (**Supplementary Table 1**) were prepared while blocking.

After overnight antibody staining, the tissue was washed 2x in CODEX staining buffer for 2 min each. Subsequently, it was fixed with 1.6% paraformaldehyde (PFA) with gentle rocking for 10 min; the PFA solution was made by diluting 16% PFA with CODEX storage buffer (Akoya Biosciences, 232107). The tissue was then washed 3x in 1x PBS, incubated in cold 100% methanol for 5 min on ice, and washed 3x with 1x PBS again. All steps except the methanol incubation were performed in 6 well plates with gentle rocking. The tissue was then fixed with CODEX final fixative for 20 min at RT in a humidity chamber; the final fixative was prepared by mixing 20 μL of CODEX final fixative (Akoya Biosciences, 232112) in 1000 μ L of 1x PBS. Finally, the tissue was rinsed 3x in 1x PBS and stored in 1x PBS at 4 ° until CODEX image acquisition.

MIBI-TOF Imaging. Datasets were acquired on a commercially available MIBIscope^{TM} System from Ionpath (Production) equipped with a Xenon ion source (Hyperion, Oregon Physics). The typical running parameters on instruments are listed as following:

- Pixel dwell time: 2 ms
- Pixel dwell time: 2 ms
- Image area: 400 μm x 400 μm
- Image size: 512 x 512 pixels
- Probe size: 400 nm
- Primary ion current: 4.9 nA on a builtin Faraday cup (or the "Fine" imaging mode)
- Number of depths: 1 depth

After acquisition, images were extracted with the toffy package (toffy notebook 3b). Detailed pre-processing is mentioned in the sections below.

CODEX Imaging. A black flat bottom 96-well plate (Corning, 07-200-762) was used for the reporter plate, where each well represented an imaging cycle. Each well was filled with 240 μL of plate master mix, containing DAPI nuclear stain (7000003, Akoya) (1:600) and CODEX assay reagent (Akoya Biosciences, 7000002) (0.5 mg/ml), as well as two fluorescent oligonucleotides (5 μL each) on the Cy3 and Cy5 channels. Blank channels were also included in the first and last wells, with plate master mix substituted for fluorescent oligonucleotides. The plate was then sealed with aluminum film and stored at 4 °C until CODEX image acquisition.
Prior to CODEX image acquisition, the tissue coverslip and reporter plate were placed into the CODEX microfluidics instrument. The coverslip was stained with 750 μL nuclear stain solution for 3 min before being washed by the fluidics device; the nuclear stain solution was prepared by mixing 1 μL of DAPI nuclear stain in 1500 μL of 1x CODEX buffer. CODEX imaging was operated under a 20x/0.75 objective (CFI Plan Apo λ, Nikon) mounted to an inverted fluorescence microscope (Keyence, BZ-X810) connected to the CODEX microfluidics instrument and CODEX driver software, and the DAPI stain was used to set up imaging areas and z planes. Each imaging cycle contained three channels - DAPI, Cy3, Cy5 - and images taken on the first and last cycles were used as blanks for background correction. Multiplexed images were stitched and background corrected using the Singer software (v1.0.7) from Akoya.

Section 2: Dataset Pre-processing.

Channel Crosstalk Removal. Similar to fluorescence imaging, mass-spectrometry imaging such as MIBI also has channel crosstalk due to the formation of adducts (14) or isotopic impurity of the elemental labels used. Thus, Rosetta algorithm was applied to extracted raw images to remove noise from channel crosstalk in a manner similar to flow-cytometry data (toffy notebook 4a). In addition to that, as background signals from bare slides and organic fragments can be partially reflected by gold and "Noodle" background channels, those counts were also removed with a fine-tuned coefficient matrix along with channel crosstalk. This step was performed with a local implementation of toffy package with minor modification.

Image Denoising. Image noise in multiplex images is a well-known issue caused by various factors such as instrumentation, tissue quality, and non-specific binding of antibodies. To tackle this challenge, a deep learning-based method is proposed that poses image denoising as a background-foreground segmentation problem. In this approach, the real signal is considered as foreground, while the noise is considered as background. The proposed method uses a supervised deep learning-based segmentation model, UNET (18), to segment the foreground from the given image. To train the model, ground truth is generated using a semi-supervised kNN-based clustering method (19). The kNN-based clustering method helps to generate reliable ground truth for the model training. Once the model is trained, it is applied to all markers in all images to obtain predicted foreground segmentation maps. These segmentation maps are then multiplied with the original images to get rid of noise and obtain clean images.

Cell segmentation. Cell segmentation of the MIBI cHL datasets was performed with a local implementation of deepcell-tf 0.6.0 as described (3, 20). Histone H3 channel was used for the nucleus, while the summation of HLA-DR, HLA1, Na-K-ATPase, CD45RA, CD11c, CD3, CD20, and CD68 was used as the membrane feature. Signals from these channels were first capped at the 99.7th percentile before input into the model.
Cell segmentation of the CODEX cHL dataset was performed using a local implementation of deepcell-tf 0.12.2. Segmentation was done using DAPI as the nuclear channel and a summation of CD4, CD7, CD15, CD30, CD11b, CD20, CD45RA, CD45RO, CD31, Podoplanin, and HLA-DR as the membrane features to ensure ideal segmentation of all cell types in the singular field of view.
The deepcell-tf version used to generate the final segmentation mask, along with the detailed parameters for cell segmentation are summarized in **Supplementary Table 2**.

Image Intensity Normalization. Due to instrumental limitation, the FOV that MIBI routinely acquired is only 400x400 μm size, stitching to achieve large tissue acquisition, and thus the across FOV difference is unavoidable. To compensate for the inter FOV difference, a set of scripts were developed and integrated into the data processing pipeline. Briefly, in a stitched run, the average Histone H3 counts under cell segmentation masks of each FOV were calculated, then, all FOVs Histone H3 counts were normalized towards the highest counts, while other channels were multiplied by the same coefficient. Additional flattening based on the Histone H3 counts were also used to avoid boundary effects and image biases. The code and parameters used are available in the analysis pipeline section.

Image to Cell Expression Matrix and across-runs normalization. The counts of each channel inside each cell segmented mask were summed up and then divided by the cell size to create the cell expression matrix based on normalized stitched TIFs along with their segmentation mask. To avoid the across-runs derivation, the median value of per cell Histone H3 of each run was calculated, then all runs medians of Histone H3, along with all other channels counts were normalized towards the highest Histone H3 median value of that MIBI dataset. The code and parameters used are available in the analysis pipeline section.

Generation of Cell Phenotyping Ground Truth. Cell phenotyping on the cHL MIBI datasets was accomplished through an iterative clustering and annotating process. The clustering was performed with FlowSOM (21) on the cHL 1 dataset and Leiden (22) on the cHL 2 dataset. The cHL 1 dataset was initially clustered with CD11c, CD14, CD15, CD153, CD16, CD163, CD20, CD3, CD30, CD4, CD56, CD57, CD68, CD8, FoxP3, GATA3, Granzyme B, and Pax-5 to capture most of the cell types present in the data. The resulting clusters were then manually annotated by examining the predominantly enriched markers of each cluster, which was done by plotting Z-score and mean expression heatmaps across all clusters and the phenotypic markers used. Clusters with a clear enrichment pattern were annotated. Next, with Mantis Viewer (23), the assigned annotation was confirmed by mapping the annotation to each cell and overlaying the raw images of the enriched markers for visual inspection. Due to noise in the data, there were certain clusters with unclear enrichment patterns. These clusters were assessed based on the phenotype marker enrichment patterns and subjected to further clustering and visual inspection. This interactive process was repeated until no useful information could be further extracted, and the remaining cells with no clear enrichment pattern were assigned as "Others". For the cHL 1 dataset, 1538433 out of 1669853 cells (92.2%) were assigned a final annotation.

Cell phenotyping on the cHL CODEX dataset was performed through an iterative process using Rphenoannoy (R implementation of PhenoGraph) and FlowSOM (21, 22) to cluster on CD30, CD20, CD2, CD7, CD8, CD57, CD4, Granzyme B, CD56, FoxP3, CD11c, CD16, CD206, CD163, CD68, CD15, CD11b, Cytokeratin, Podoplanin, CD31, MCT, and a-SMA. The resulting stratified cell clusters and corresponding enriched phenotypic markers were then visualized with Z-score and mean expression heatmaps. Cells were then individually mapped back to the original tissue images in QuPath 0.2.0-m1 to validate marker enrichment. Clusters with clear enrichment patterns for a particular cell type were annotated accordingly. Clusters with unclear or partially correct enrichment patterns were further clustered using FlowSOM based on a curated subset of phenotypic markers present on these unclear populations. Multiple iterations of clustering and annotation were performed until signal-noise ratio was too low to confidently distinguish the phenotype of the remaining cells, which were assigned as “Others”. 140,053 out of 145,161 cells (96.5%) were assigned a final annotation. All final annotations were assessed by S.J. and S.J.R (a board certified hematopathologist).

Section 3: Datasets Overview.

Our study utilized four different datasets for cell phenotyping in Classical Hodgkin’s lymphoma and CRC. The cHL 1 and cHL 2 datasets were acquired using Multiplexed Ion Beam Imaging (MIBI) and contained cells from 13 and 12 different phenotypes, respectively. The cHL CODEX and CRC CODEX datasets were acquired using Co-detection by Indexing (CODEX) and contained cells from 16 and 14 different phenotypes, respectively. The datasets had varying numbers of cells, protein/functional markers, and levels of class imbalance, and were split into five-folds for cross-validation with the mentioned method.

cHL 1 and cHL 2 (MIBI) Dataset. The cHL 1 and cHL 2 (MIBI) Datasets are two in-house datasets used in our study for cell phenotyping in cHL. Both sets of samples were stained with the same batch of antibody cocktail (**Supplementary Table 1**) with 46 protein/functional markers, and acquired using Multiplexed Ion Beam Imaging (MIBI). cHL 1 Dataset contains 1,669,853 cells from 18 cHL patients and 1 control rLN, while cHL 2 Dataset has over 230,000 cells from six FOVs - five from cHL patients and one from a control rLN. When training the proposed method, 5 markers from the cHL 1 dataset were dropped due to poor staining quality, while all 46 markers remained in the training set of cHL 2. To evaluate the performance of our proposed method, both datasets were split into 5 folds for multi-fold training and validation of the proposed method, and under both cases, the FOVs of the control cases were part of the training set in each fold.

cHL (CODEX) Dataset. The cHL (CODEX) dataset is another in-house dataset that was acquired using Co-Detection by Indexing (CODEX), a multiplex imaging technique that allows for simultaneous detection of over 50 markers. The dataset consists of a single large FOV containing over 143,000 cells. The cells in the cHL (CODEX) dataset are classified into 16 different cell phenotypes, and each class has an average of 8000+ cells. The multiplex FOV in this dataset consists of 49 markers, which include different markers than those used in the cHL 1 (MIBI) and cHL 2 (MIBI) datasets (see **Supplementary Table 1** for more details). To evaluate the performance of MAPS, we randomly split the cells in the cHL (CODEX) dataset into five folds using stratified sampling to ensure a balanced number of cells in each fold for each class.

CRC CODEX Dataset. The CRC CODEX dataset (DOI: 10.17632/mpjzbtfgfr.1) is a public dataset that we used in our study to evaluate our proposed method for cell phenotyping (11). It consists of more than 200,000 cells from 14 different classes, with a large variation in the number of cells per class, ranging from as low as 323 cells to as high as >47,000 cells.

For our study, we used the same markers and classes as described in the CellSighter paper to ensure a fair head-to-head comparison with MAPS. As there was no information available about the training and validation split in the dataset, we adopted the same five-fold cross-validation approach that we used for the above datasets in our study.

Section 4: MAPS model, training and evaluation.

Model architecture. The proposed cell phenotyping method used a feed-forward neural network to predict the cell class from a set of predefined classes (K). Let $x \in \mathbb{R}^{N+1}$ be the input data, which consists of the expression of a cell for N markers and its area in pixels. The neural network processes this input data to generate a predicted cell class y. The neural network used in the proposed method consists of four fully connected hidden layers, denoted by h_1, h_2, h_3, and h_4. Each hidden layer is followed by a ReLU activation function and a dropout layer, denoted by g_1, g_2, g_3, and g_4. The output of the last hidden layer, h_4, is fed into the classification layer, which generates the predicted cell class y. The classification layer uses a softmax function to convert the output of the neural network into a probability distribution over the predefined classes. Let W_i and b_i denote the weights and biases of the i^{th} layer of the neural network, respectively. Then the output h_i of the i^{th} hidden layer can be written as:

$$h_i = g_i(W_i h_{i-1} + b_i)$$

where $h_{i-1} \in \mathbb{R}^{512}$ is the output of the $(i-1)^{th}$ hidden layer or the input x for $i = 1$, and g_i is the activation function for the i^{th} layer, which is the ReLU function in this case. The dropout layers are not included in this equation, as they only modify the output of the hidden layers during training, and do not affect the final output of the neural network. The classification layer computes the predicted cell class y as follows:

$$y = \underset{k}{\operatorname{argmax}} \operatorname{softmax}(W_c h_4 + b_c)_k$$

where W_c and b_c are the weights and biases of the classification layer, and $\operatorname{softmax}$ is the softmax function that converts the k^{th} output into a probability distribution over the predefined classes (K). The predicted cell class y is the class with the highest probability.

Training details. For the training of the proposed method, the Adam optimizer with a learning rate of 0.001, batch size of 128 and a dropout probability of 0.10 was used for all datasets. The number of training epochs varied for each dataset due to the varying sizes of the datasets. The larger datasets, such as cHL 1 (MIBI) dataset, have more optimization steps in each epoch as compared to the smaller datasets. Specifically, the model was trained for 100 epochs on the cHL 1 (MIBI) dataset, and for 500 epochs on all other datasets. During training, if the validation loss did not decrease for a certain number of epochs, the training was stopped to save time and the model with lowest validation loss was selected as the best model for inference.

Section 5: Evaluation Across Methods.

To evaluate the performance of the proposed method, we employed several evaluation methods. Firstly, we used the confusion matrix to visualize the performance of the model. The confusion matrix displays the number of true positive, false positive, true negative, and false negative predictions made by the model. From the confusion matrix, we calculated the precision, recall, and F1-score metrics. Precision measures the proportion of true positive predictions made by the model out of all the positive predictions made, while recall measures the proportion of true positive predictions made out of all the actual positive instances in the dataset. The F1-score is the harmonic mean of precision and recall and is a balanced measure of both metrics.
Additionally, we used the average precision metric, which measures the area under the precision-recall curve. This metric is particularly useful for imbalanced datasets, where there are more negative instances than positive ones. The average precision metric takes into account the precision and recall values at various thresholds and provides a summary of the model's overall performance.
Finally, we also used the mean cell expression matrix to visualize the expression levels of different markers in the different cell types predicted by the model. This matrix provides a summary of the mean expression levels of each marker in each cell type and can help to identify differences in marker expression between different cell types when compared with the cell expression matrix generated using ground truth labels.

Comparisons With Other Methods. We compared our proposed method with two existing cell phenotyping methods, namely ASTIR and CellSighter. The code for both ASTIR and CellSighter methods is publicly available for reproducibility and comparison purposes.

ASTIR. ASTIR is a probabilistic model for cell phenotyping that uses deep recognition neural networks to predict cell types without requiring labels for each cell (7). Instead, ASTIR only requires a list of protein markers for each expected cell type within a dataset. The method is based on the assumption that each cell type can be characterized by a unique combination of protein markers, and that the expression levels of these markers can be used to classify cells into their respective types. We reported results of the ASTIR method on three in-house datasets. For each dataset, our experts defined the list of protein markers for each cell type. We evaluated the results using five-fold cross-validation, using exactly the same folds as in the proposed method, for a fair head-to-head comparison.

CellSighter. The CellSighter is a deep learning based supervised cell classification method (10). Unlike ASTIR and the proposed method which works on cell expression matrices, CellSighter takes image, cell segmentation mask, and cell to class mapping as input. To evaluate the performance of CellSighter, we re-trained it on the same three in-house datasets

using the same 5-fold cross validation splits as in the proposed method. This ensures a fair comparison between the methods. We obtained the CellSighter results on the publicly available CRC CODEX dataset from the paper to avoid any re-training bias while comparing it with the MAPS results on the same dataset.

Computation Resource Evaluation across Methods. To evaluate the computation resource usage of each method, we ran the three methods on a Linux platform (2x Intel Xeon 6334 'Ice Lake-SP' 3.6 GHz 8-core 10nm CPUs; 4x NVIDIA "Ampere" RTX A5000 PCI-E+NVLink 24GB ECC GPU Accelerator / Graphics Cards; 1TB DDR4 memory @ 3200MHz) using the cHL (CODEX) dataset. During model training and cell type inference of each method, we tracked their CPU, GPU, and memory (RAM) usage using "top", "ps -ef", and "nvidia-smi" commands. For the parallel methods, we recorded the resource usage of all its processes and multiplied it by the number of cores used in parallel.

Data Visualization. Single channel and multi-color images were assembled and visually inspected with either ImageJ (24), Qupath (25), and Mantis Viewer (23). Visualizations of the analysis results were either produced using Excel, or R packages 'ggplot2' and 'pheatmap'.

Data & Code Availability. All the data described in this work, including channel images and segmentation will be publiclly available on Mendeley Data upon publication. The code for anchoring analysis and data visualization can be downloaded at https://github.com/mahmoodlab/MAPS.

ACKNOWLEDGEMENTS

The authors thank members of the Jiang and Mahmood laboratories for their discussions. S.J. is supported by NIH DP2AI171139, R01AI149672, a Gilead's Research Scholars Program in Hematologic Malignancies, and the Bill & Melinda Gates Foundation INV-002704. F.M. and M.S. are funded by R35GM138216, and the Fredrick National Laboratory. G.P.N. is supported by the Rachford and Carlota A. Harris Endowed Professorship. M.A.S. and S.J.R. are supported by a Blood Cancer Discoveries Grant Program from the Leukemia Lymphoma Society, The Mark Foundation, and The Paul G. Allen Frontiers Group. This article reflects the views of the authors and should not be construed as representing the views or policies of the institutions that provided funding.

AUTHOR CONTRIBUTIONS

Conceptualization: S.J., F.M., S.J.R.
Methodology: Y.B., Y.Y.Y., H.C., B.Z., S.J.
Software: M.S., H.Q., Y.B.
Pathological Annotations: V.S., S.J.R.
Formal Analysis: M.S., H.Q., Y.B., S.M., J.Y., Y.Y.Y.
Writing – Original Draft: M.S., H.Q., Y.B., S.J.
Writing – Reviewing and Editing: all authors
Supervision: S.J., F.M., S.J.R., M.A.S., G.P.N.

CONFLICT OF INTERESTS

S.J.R. has received research support from Affimed, Merck, and Bristol-Myers Squibb (BMS), he is on the Scientific Advisory Board for Immunitas Therapeutics and part of Bristol Myers Squibb International Immuno-Oncology Network (II-ON). M.A.S. has research funding from BMS, Bayer, Abbvie, and AstraZeneca, and is on advisory boards for AstraZeneca and BMS. G.P.N., is co-founder and stockholder of IonPath Inc, which manufactures the instrument used in this manuscript, is a co-founder and stockholder of Akoya Biosciences, Inc. and inventor on patent US9909167, and is a Scientific Advisory Board member for Akoya Biosciences, Inc. The other authors declare no competing interests.

Reference

1. John W. Hickey, Elizabeth K. Neumann, Andrea J. Radtke, Jeannie M. Camarillo, Rebecca T. Beuschel, Alexandre Albanese, Elizabeth McDonough, Julia Hatler, Anne E. Wiblin, Jeremy Fisher, Josh Croteau, Eliza C. Small, Anup Sood, Richard M. Caprioli, R. Michael Angelo, Garry P. Nolan, Kwanghun Chung, Stephen M. Hewitt, Ronald N. Germain, Jeffrey M. Spraggins, Emma Lundberg, Michael P. Snyder, Neil L. Kelleher, and Sinem K. Saka. Spatial mapping of protein composition and tissue organization: a primer for multiplexed antibody-based imaging. *Nature Methods*, 19(3):284–295, 2022. ISSN 1548-7091. doi: 10.1038/s41592-021-01316-y.
2. Maryam Pourmaleki, Nicholas D. Socci, Travis J. Hollmann, and Ingo K. Mellinghoff. Moving Spatially Resolved Multiplexed Protein Profiling toward Clinical Oncology. *Cancer Discovery*, 13(4):824–828, 2023. ISSN 2159-8274. doi: 10.1158/2159-8290.cd-22-1015.
3. Noah F. Greenwald, Geneva Miller, Erick Moen, Alex Kong, Adam Kagel, Thomas Dougherty, Christine Camacho Fullaway, Brianna J. McIntosh, Ke Xuan Leow, Morgan Sarah Schwartz, Cole Pavelchek, Sunny Cui, Isabella Camplisson, Omer Bar-Tal, Jaiveer Singh, Mara Fong, Gautam Chaudhry, Zion Abraham, Jackson Moseley, Shiri Warshawsky, Erin Soon, Shirley Greenbaum, Tyler Risom, Travis Hollmann, Sean C. Bendall, Leeat Keren, William Graf, Michael Angelo, and David Van Valen. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nature Biotechnology*, pages 1–11, 2021. ISSN 1087-0156. doi: 10.1038/s41587-021-01094-0.
4. Yunhao Bai, Bokai Zhu, Xavier Rovira-Clave, Han Chen, Maxim Markovic, Chi Ngai Chan, Tung-Hung Su, David R. McIlwain, Jacob D. Estes, Leeat Keren, Garry P. Nolan, and Sizun Jiang. Adjacent Cell Marker Lateral Spillover Compensation and Reinforcement for Multiplexed Images. *Frontiers in Immunology*, 12:652631, 2021. doi: 10.3389/fimmu.2021.652631.
5. Darci Phillips, Christian M. Schürch, Michael S. Khodadoust, Youn H. Kim, Garry P. Nolan, and Sizun Jiang. Highly Multiplexed Phenotyping of Immunoregulatory Proteins in the Tumor Microenvironment by CODEX Tissue Imaging. *Frontiers in Immunology*, 12:687673, 2021. doi: 10.3389/fimmu.2021.687673.
6. Sizun Jiang, Chi Ngai Chan, Xavier Rovira-Clavé, Han Chen, Yunhao Bai, Bokai Zhu, Erin McCaffrey, Noah F. Greenwald, Candace Liu, Graham L. Barlow, Jason L. Weirather, John Paul Oliveria, Tsuguhisa Nakayama, Ivan T. Lee, Matthias S. Matter, Anne E. Carlisle, Darci Philips, Gustavo Vazquez, Nilanjan Mukherjee, Kathleen Busman-Sahay, Michael Nekorchuk, Margaret Terry, Skyler Younger, Marc Bosse, Janos Demeter, Scott J. Rodig, Alexandar Tzankov, Yury Goltsev, David Robert McIlwain, Michael Angelo, Jacob D. Estes, and Garry P. Nolan. Combined protein and nucleic acid imaging reveals virus-dependent B cell and macrophage immunosuppression of tissue microenvironments. *Immunity*, 55(6):1118–1134.e8, 2022. ISSN 1074-7613. doi: 10.1016/j.immuni.2022.03.020.
7. Michael J. Geuenich, Jinyu Hou, Sunyun Lee, Shanza Ayub, Hartland W. Jackson, and Kieran R. Campbell. Automated assignment of cell identity from single-cell multiplexed imaging and proteomic data. *Cell Systems*, 12(12):1173–1186.e5, 2021. ISSN 2405-4712. doi: 10.1016/j.cels.2021.08.012.
8. Weiruo Zhang, Irene Li, Nathan E. Reticker-Flynn, Zinaida Good, Serena Chang, Nikolay Samusik, Saumyaa Saumyaa, Yuanyuan Li, Xin Zhou, Rachel Liang, Christina S. Kong, Quynh-Thu Le, Andrew J. Gentles, John B. Sunwoo, Garry P. Nolan, Edgar G. Engleman, and Sylvia K. Plevritis. Identification of cell types in multiplexed in situ images by combining protein expression and spatial information using CELESTA. *Nature Methods*, 19(6):759–769, 2022. ISSN 1548-7091. doi: 10.1038/s41592-022-01498-z.
9. Maria Brbić, Kaidi Cao, John W. Hickey, Yuqi Tan, Michael P. Snyder, Garry P. Nolan, and Jure Leskovec. Annotation of spatially resolved single-cell data with STELLAR. *Nature Methods*, 19(11):1411–1418, 2022. ISSN 1548-7091. doi: 10.1038/s41592-022-01651-8.
10. Yael Amitay, Yuval Bussi, Ben Feinstein, Shai Bagon, Idan Milo, and Leeat Keren. CellSighter – A neural network to classify cells in highly multiplexed images. *bioRxiv*, page 2022.11.07.515441, 2022. doi: 10.1101/2022.11.07.515441.
11. Christian M. Schürch, Salil S. Bhate, Graham L. Barlow, Darci J. Phillips, Luca Noti, Inti Zlobec, Pauline Chu, Sarah Black, Janos Demeter, David R. McIlwain, Shigemi Kinoshita, Nikolay Samusik, Yury Goltsev, and Garry P. Nolan. Coordinated Cellular Neighborhoods Orchestrate Antitumoral Immunity at the Colorectal Cancer Invasive Front. *Cell*, 182(5):1341–1359.e19, 2020. ISSN 0092-8674. doi: 10.1016/j.cell.2020.07.005.
12. Guojun Han, Matthew H. Spitzer, Sean C. Bendall, Wendy J. Fantl, and Garry P. Nolan. Metal-isotope-tagged monoclonal antibodies for high-dimensional mass cytometry. *Nature Protocols*, 13(10):2121–2148, 2018. ISSN 1754-2189. doi: 10.1038/s41596-018-0016-7.
13. Leeat Keren, Marc Bosse, Diana Marquez, Roshan Angoshtari, Samir Jain, Sushama Varma, Soo-Ryum Yang, Allison Kurian, David Van Valen, Robert West, Sean C. Bendall, and Michael Angelo. A Structured Tumor-Immune Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam Imaging. *Cell*, 174(6):1373–1387.e19, 2018. ISSN 0092-8674. doi: 10.1016/j.cell.2018.08.039.
14. Leeat Keren, Marc Bosse, Steve Thompson, Tyler Risom, Kausalia Vijayaragavan, Erin McCaffrey, Diana Marquez, Roshan Angoshtari, Noah F. Greenwald, Harris Fienberg, Jennifer Wang, Neeraja Kambham, David Kirkwood, Garry Nolan, Thomas J. Montine, Stephen J. Galli, Robert West, Sean C. Bendall, and Michael Angelo. MIBI-TOF: A multiplexed imaging platform relates cellular phenotypes and tissue structure. *Science Advances*, 5(10):eaax5851, 2019. doi: 10.1126/sciadv.aax5851.
15. Andrew L Ji, Adam J Rubin, Kim Thrane, Sizun Jiang, David L Reynolds, Robin M Meyers, Margaret G Guo, Benson M George, Annelie Mollbrink, Joseph Bergenstråhle, et al. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell*, 182(2):497–514, 2020.
16. Xavier Rovira-Clavé, Alexandros P. Drainas, Sizun Jiang, Yunhao Bai, Maya Baron, Bokai Zhu, Alec E. Dallas, Myung Chang Lee, Theresa P. Chu, Alessandra Holzem, Ramya Ayyagari, Debadrita Bhattacharya, Erin F. McCaffrey, Noah F. Greenwald, Maxim Markovic, Garry L. Coles, Michael Angelo, Michael C. Bassik, Julien Sage, and Garry P. Nolan. Spatial epitope barcoding reveals clonal tumor patch behaviors. *Cancer Cell*, 40(11):1423–1439.e11, 2022. ISSN 1535-6108. doi: 10.1016/j.ccell.2022.09.014.
17. Sarah Black, Darci Phillips, John W. Hickey, Julia Kennedy-Darling, Vishal G. Venkataraaman, Nikolay Samusik, Yury Goltsev, Christian M. Schürch, and Garry P. Nolan. CODEX multiplexed tissue imaging with DNA-conjugated antibodies. *Nature Protocols*, 16(8):3802–3835, 2021. ISSN 1754-2189. doi: 10.1038/s41596-021-00556-8.
18. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv*, 2015. doi: 10.48550/arxiv.1505.04597.

19. Alex Baranski, Idan Milo, Shirley Greenbaum, John-Paul Oliveria, Dunja Mrdjen, Michael Angelo, and Leeat Keren. MAUI (MBI Analysis User Interface)—An image processing pipeline for Multiplexed Mass Based Imaging. *PLoS Computational Biology*, 17(4): e1008887, 2021. ISSN 1553-734X. doi: 10.1371/journal.pcbi.1008887.
20. David A. Van Valen, Takamasa Kudo, Keara M. Lane, Derek N. Macklin, Nicolas T. Quach, Mialy M. DeFelice, Inbal Maayan, Yu Tanouchi, Euan A. Ashley, and Markus W. Covert. Deep Learning Automates the Quantitative Analysis of Individual Cells in Live-Cell Imaging Experiments. *PLoS Computational Biology*, 12(11):e1005177, 2016. ISSN 1553-734X. doi: 10.1371/journal.pcbi.1005177.
21. Sofie Van Gassen, Britt Callebaut, Mary J. Van Helden, Bart N. Lambrecht, Piet Demeester, Tom Dhaene, and Yvan Saeys. FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry Part A*, 87(7):636–645, 2015. ISSN 1552-4930. doi: 10.1002/cyto.a.22625.
22. F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 2018. ISSN 1474-7596. doi: 10.1186/s13059-017-1382-0.
23. Robert Schiemann, Pier Federico Gherardini, Robin Kageyama, Mike Travers, and Lacey Kitch. Mantis Viewer (v1.2.0-beta.1). *Zenodo*, 2020. doi: https://doi.org/10.5281/zenodo.4909620.
24. Caroline A Schneider, Wayne S Rasband, and Kevin W Eliceiri. NIH Image to ImageJ: 25 years of image analysis. *Nature Methods*, 9(7):671–675, 2012. ISSN 1548-7091. doi: 10.1038/nmeth.2089.
25. Peter Bankhead, Maurice B. Loughrey, José A. Fernández, Yvonne Dombrowski, Darragh G. McArt, Philip D. Dunne, Stephen McQuaid, Ronan T. Gray, Liam J. Murray, Helen G. Coleman, Jacqueline A. James, Manuel Salto-Tellez, and Peter W. Hamilton. QuPath: Open source software for digital pathology image analysis. *Scientific Reports*, 7(1):16878, 2017. doi: 10.1038/s41598-017-17204-5.

Supplementary Figures

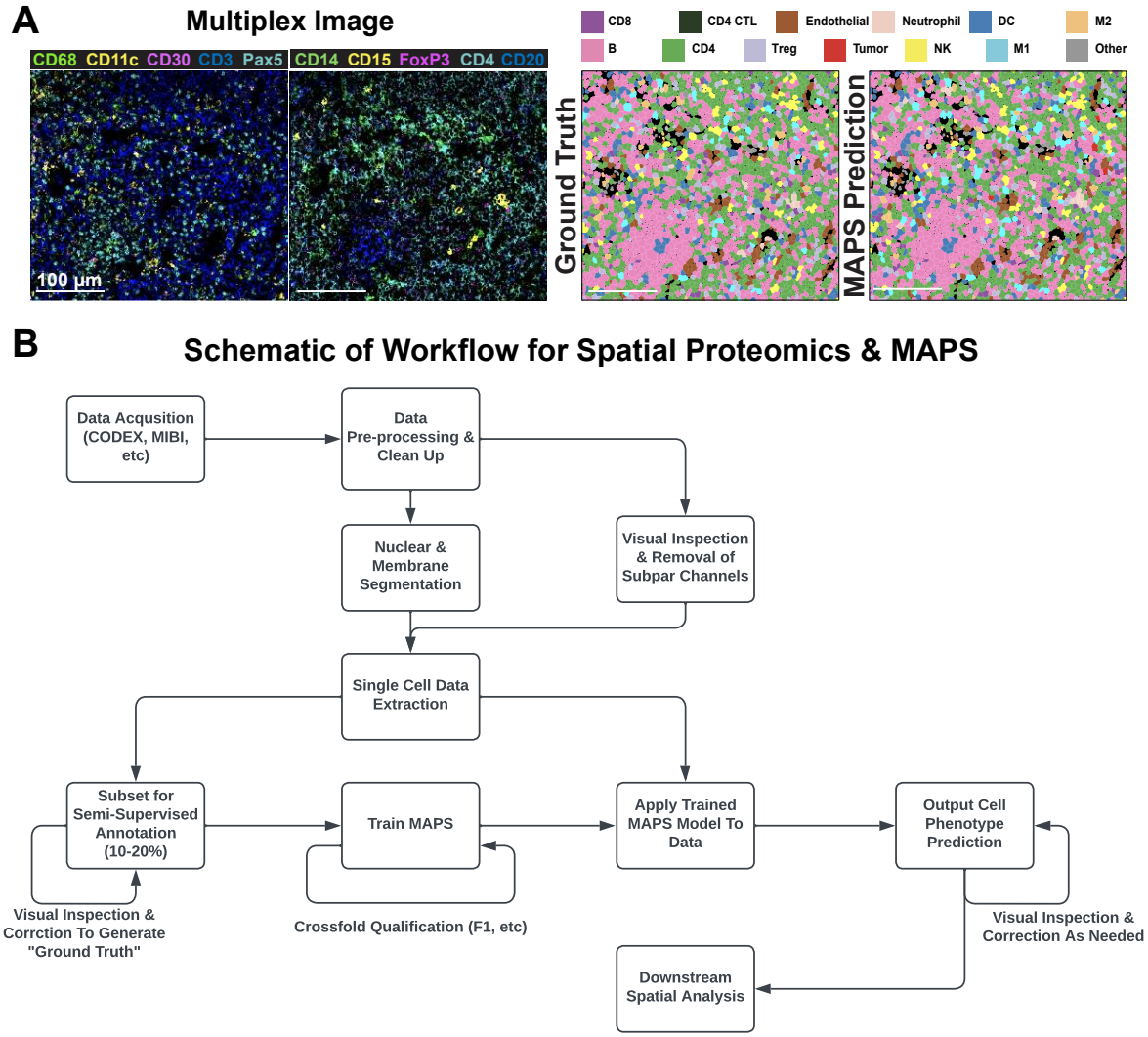

Figure S1: Performance of MAPS on rLN and scheme of the MAPS workflow, related to Figure 1. (A) Representative FOV of a multiplexed image of rLN used for cell phenotyping. Cell phenotype maps generated via manual annotation (Ground Truth) or MAPS (MAPS Prediction) are shown for visual comparison. **(B)** Schematic of the workflow for spatial proteomics cell phenotyping accelerated by MAPS.

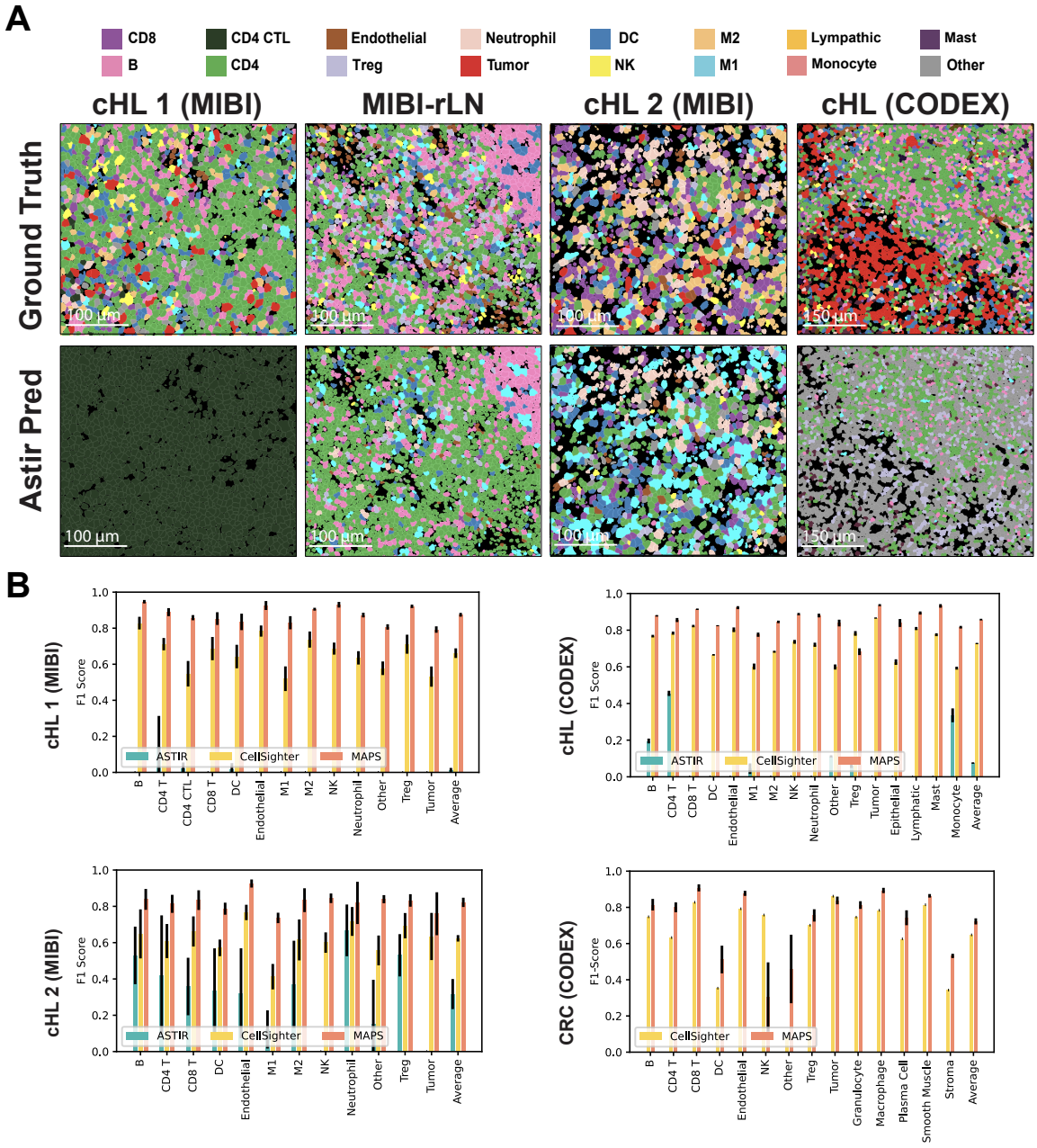

Figure S2: Visual and quantitative comparison of MAPS performance with its counterparts, related to Figure 2. (A) Comparison of ground truth and Astir performances across four multiplex image datasets. **(B)** Performance comparison at class level F1-score of three cell phenotyping methods across all four datasets with average F1-score across 5-folds. Error bars represent ± 1 standard deviation.

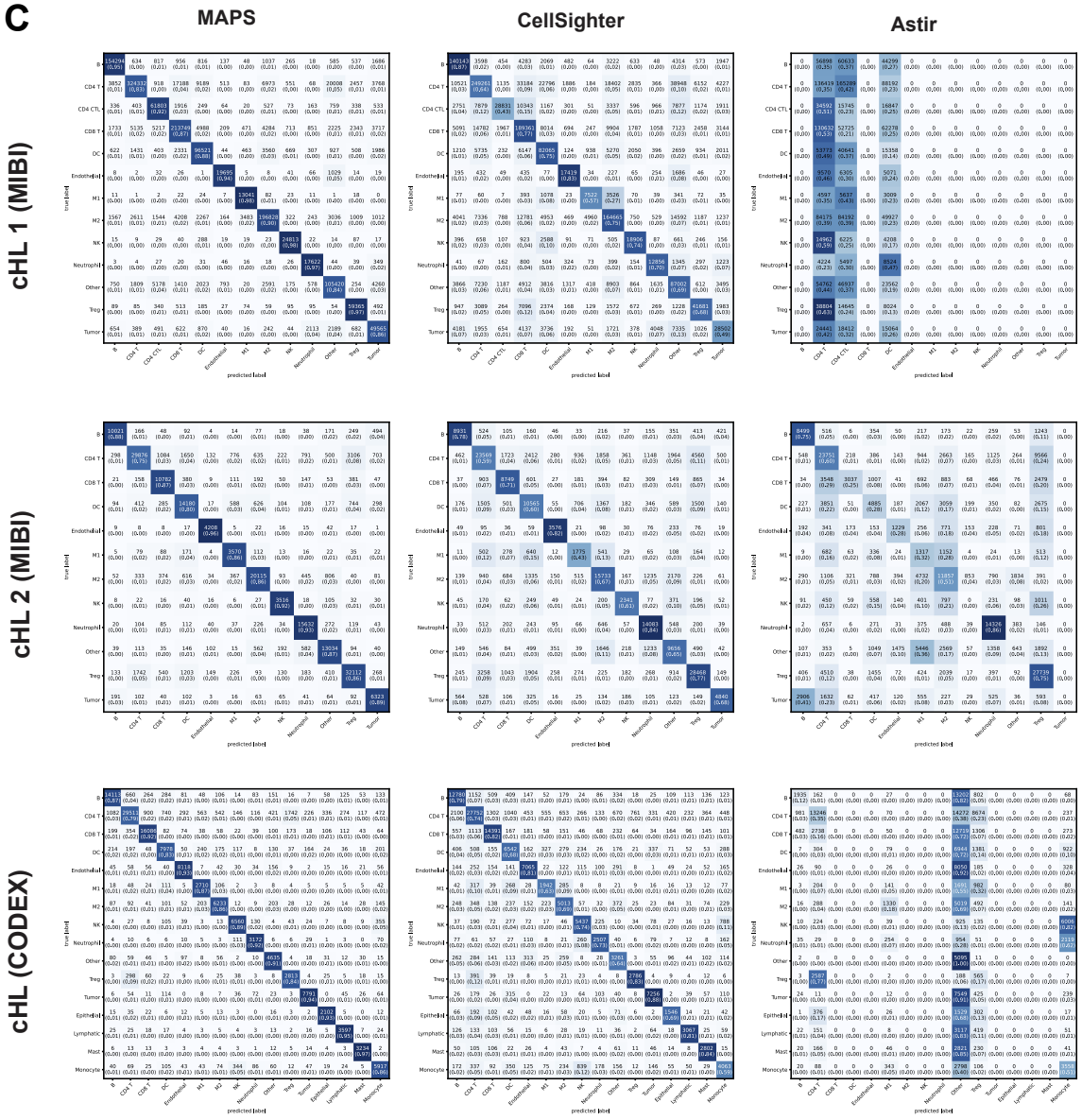

Figure S2 continued: Visual and quantitative comparison of MAPS performance with its counterparts, related to Figure 2. (C) Confusion matrix of MAPS prediction along with CellSighter and Astir on three dataset similar to **Figure 1C**. Numbers in parentheses indicate the percentage of cells that are correctly predicted by the corresponding method.

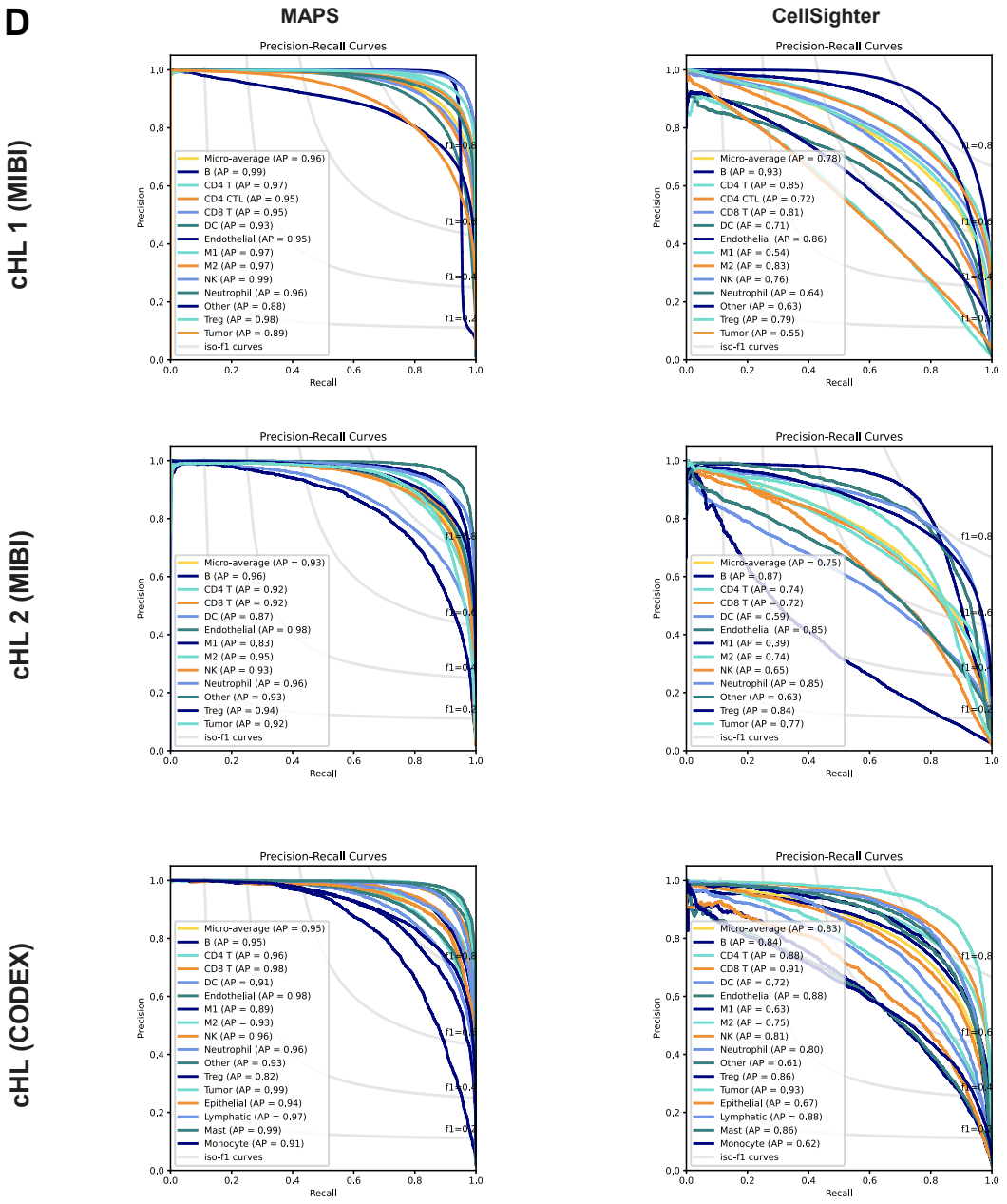

Figure S2 continued: Visual and quantitative comparison of MAPS performance with its counterparts, related to Figure 2. (D) Precision and recall curves of MAPS and CellSighter based on the prediction for each dataset. Each plot presents not only the overall precision and recall curve of the model but also the curves for each cell type. Each curve shows the precision and recall trade-off for different thresholds.