Flexible and robust cell type annotation for highly multiplexed tissue images

Huangqingbo Sun[1,2], Shiqiu Yu[1], Anna Martinez Casals[2], Anna Bäckström[2,3], Yuxin Lu[1], Cecilia Lindskog[6], Emma Lundberg[2,3,4,5]✉, and Robert F. Murphy[1,]✉

[1]Ray and Stephanie Lane Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA, USA
[2]Department of Bioengineering, Stanford University, Stanford, CA, USA
[3]Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH–Royal Institute of Technology, Stockholm, Sweden
[4]Department of Pathology, Stanford University, Stanford, CA, USA
[5]Chan Zuckerberg Biohub San Francisco, San Francisco, CA, USA
[6]Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden

Identifying cell types in highly multiplexed images is essential for understanding tissue spatial organization. Current cell type annotation methods often rely on extensive reference images and manual adjustments. In this work, we present a tool, Robust Image-Based Cell Annotator (RIBCA), that enables accurate, automated, unbiased, and fine-grained cell type annotation for images with a wide range of antibody panels, without requiring additional model training or human intervention. Our tool has successfully annotated over 1 million cells, revealing the spatial organization of various cell types across more than 40 different human tissues. It is open-source and features a modular design, allowing for easy extension to additional cell types.

Spatial proteomics | Multiplexed imaging | Machine learning | Cell type annotation

Correspondence: *emmalu@stanford.edu, murphy@cmu.edu*

Main

Cells can be characterized into different types by their molecular, morphological, physiological, and functional properties (1). Different types of cells are precisely localized to allow interactions that give rise to specific environments for physiological functions. Advances in multiplexed tissue imaging now enable the probing of over 50 proteins in a single tissue sample, potentially identifying many different cell types through their proteomic profiles. However, unlike spatial transcriptomics in which thousands of markers are measured, the lower number of markers in spatial proteomics poses challenges for devising generalizable and robust cell annotation frameworks. Traditional approaches rely on manually-specified marker ranges (gates), but they often lack reproducibility and require prior knowledge to choose cell types and ranges(2). Alternatively, supervised machine learning (ML) constructs end-to-end models, but training these ML models requires large reference datasets that are often unavailable and difficult to obtain (3, 4). Unsupervised clustering methods can also be employed for cell labeling, but their performances generally fall short of manual gating or supervised methods.

Furthermore, many methods rely solely on average intensities to annotate cell types, overlooking the spatial information and subcellular protein patterns provided by images. As Rumberger *et al.* demonstrated, the difference in average intensity between positive and negative expressions can be minimal (5). Moreover, current approaches focus more on method development than on creating *ready-to-use* software, leaving users to annotate their own data and retrain models based on published methods.

To address these challenges, we introduce RIBCA for cell type annotation in multiplexed tissue images that is generalizable to new image collections without extra fine-tuning. Instead of a single model, we constructed an ensemble of image-derived models, where each base model is explicitly designed for a subset of protein markers and a subset of cell types of interest (Figure 1**a**). The rationale behind this modular design is to make it compatible with any common antibody panel by matching their markers with one or multiple base models in this ensemble. Our approach involves cropping whole tissue images into single-cell regions, using imputation models to predict the patterns of any missing markers, predicting with any matched model, and merging the results (Figure 1**b** and **c**). The output is a cell type map, its annotation confidence, and spatial statistics of cell type distribution. Our software has a Napari plugin for interactively validating annotations.

Unlike methods that convert pixel information into average channel values, image-derived models learn subcellular patterns directly from images and use them to classify. Here we employed the Vision Transformer (ViT) (6), illustrated in Figure 1**d**. We trained individual ViT for five publicly available multiplexed image datasets, and compared results with three recent methods (see Methods). We performed 7-fold crossfold validation by randomly partitioning the whole dataset into training and test sets, and reported the average scores (Figure 1**e**). Although no single method outperformed others in all datasets, ours achieves overall better and more stable performance (Figure 1**e**).

Success in training models for each dataset motivated us to develop a generic pre-trained and ready-to-use ensemble model. We began by curating a multiplexed tissue image dataset (referred to as `SPData_1`) containing more than 15 million single-cell images based on 16 publicly available image datasets. More than half of these images have annotated cell type labels. To ensure consistent annotations across the dataset, we aligned cell type names at varying levels of granularity, making them compatible with most existing annotations. We randomly partitioned this dataset into training and

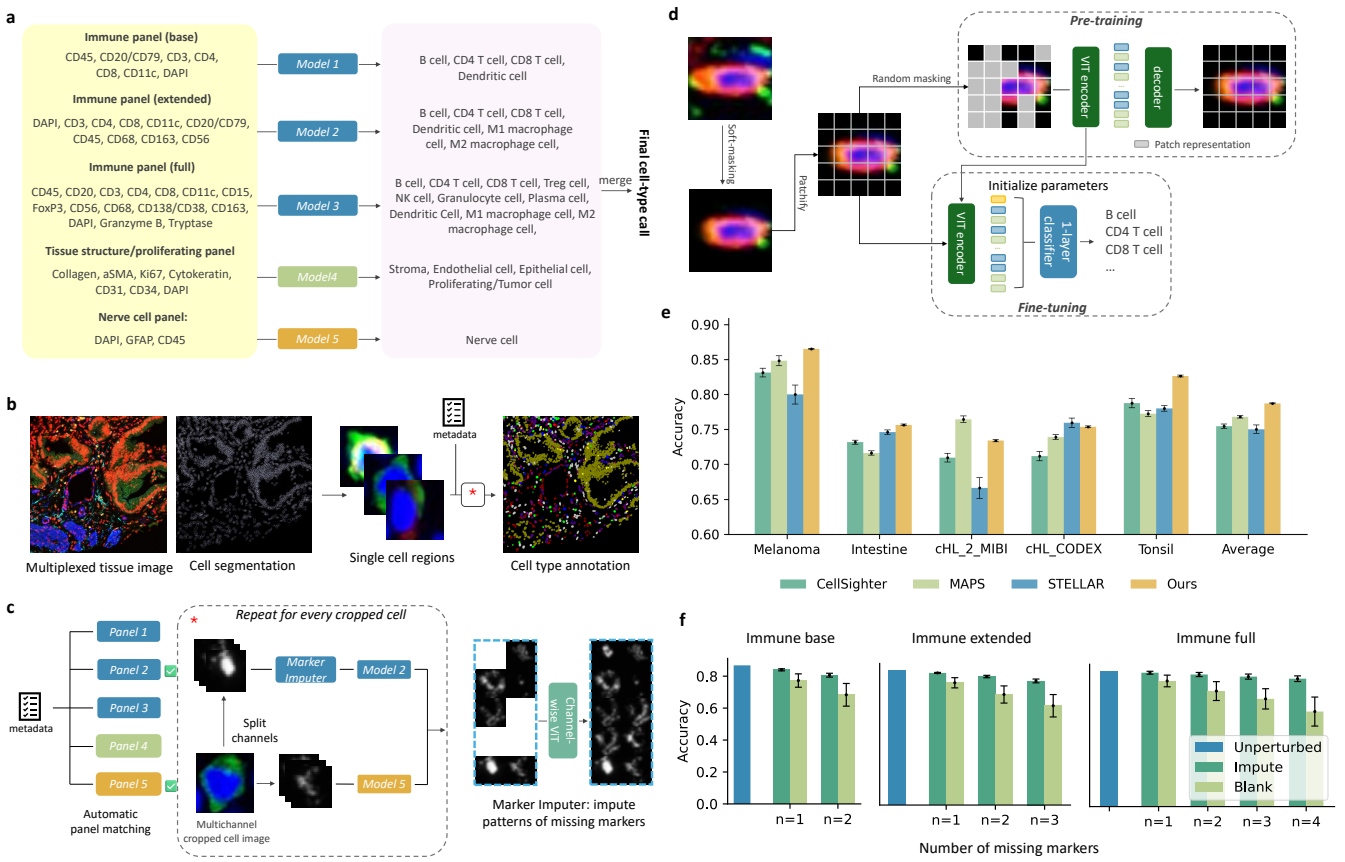

Fig. 1. Overview of RIBCA. **a**, Ensemble design for base models to map a subset of image channels to particular cell types. **b**, A multiplexed tissue image and its cell segmentation is cropped into single cell regions to annotate individual cells. **c**, Available channels are checked to determine if they include enough to be acceptable to a given panel (indicated by check marks). In this case, Panels 2 and 5 are selected, Marker Imputers fill in missing markers, cell type models classify the imputed images, and the classification results are merged. **d**, Diagrams of pre-training and fine-tuning Vision Transformer models for cropped cell images. **e**, Average cross-validation accuracies for different methods on five test sets. Error bars indicate ± 1 standard deviation from the mean. **f**, Accuracies of immune panel models (trained using `SPData_1`) when evaluated with full images or averaged over removing one or more channels and either replacing it with a blank image or by imputation.

validation sets (25:1) to train the ensemble model.

We trained five models using those training images that contained the full set of markers for a given panel (see Figure 1**c**. Each model was a customized ViT pre-trained and fine-tuned on `SPData_1` (model hyperparameters are listed in Table S2). Three are used to annotate immune cells, one for tissue structures and proliferating cells, and one for nerve cells. Note that CD45 in the tissue structure and nerve cell panels serves as a negatively expressed marker to distinguish immune cells. Final predictions from the ensemble are made by voting applicable base models. Figure 1**f** shows their performances on our held-out validation set, demonstrating the effectiveness of these models.

To enhance robustness, we addressed scenarios where a query image might be missing only one or a few markers needed for a panel by developing a missing marker imputation strategy that works for a specific panel. This uses ViTs trained using a channel-wise masked auto-encoding (MAE) strategy, which takes the image channels of observed markers as input and predicts images for any missing markers (Figure 1**c**). Imputing missing markers consistently improved cell type classification accuracy compared to using blank image channels (Figure 1**f**), with greater benefits as more markers were missing. The differences in accuracy in terms of each cell type are shown in Figure S3. For example, when imputing 4 missing markers of the immune full panel, the difference between using imputed images and real images is only 4.79%; however, this difference is 25.32% without imputation. The success in imputing missing markers makes RIBCA adaptable to various commonly used marker panels.

To assess the generalizability of the ensemble on images from completely new experiments (and tissue types not present in `SPData_1`), we generated a dataset using the Human Protein Atlas (HPA) tissue microarrays (TMA) (referred as `SPHPAData_1`). In total, we generated 148 images of TMA cores from 46 different human tissues, acquired using a 46-plex panel (Table S3). This dataset will be released in v24 of the HPA, and will be useful to the community for benchmarking purposes.

Examples of TMA multiplexed images and their annotations from RIBCA (without any fine-tuning) are shown in Figure 2**a** and **b**. The results align with biologists' expectations in two key ways: (1) the spatial arrangement of cells, such as epithelial cells along intestinal crypts and B cells clustering in the spleen's germinal center; (2) marker expression, evidenced by clear signals in the cell membranes of CD4 and CD8 T cells. To quantitatively assess performance, we manually validated selected images, achieving a high accuracy

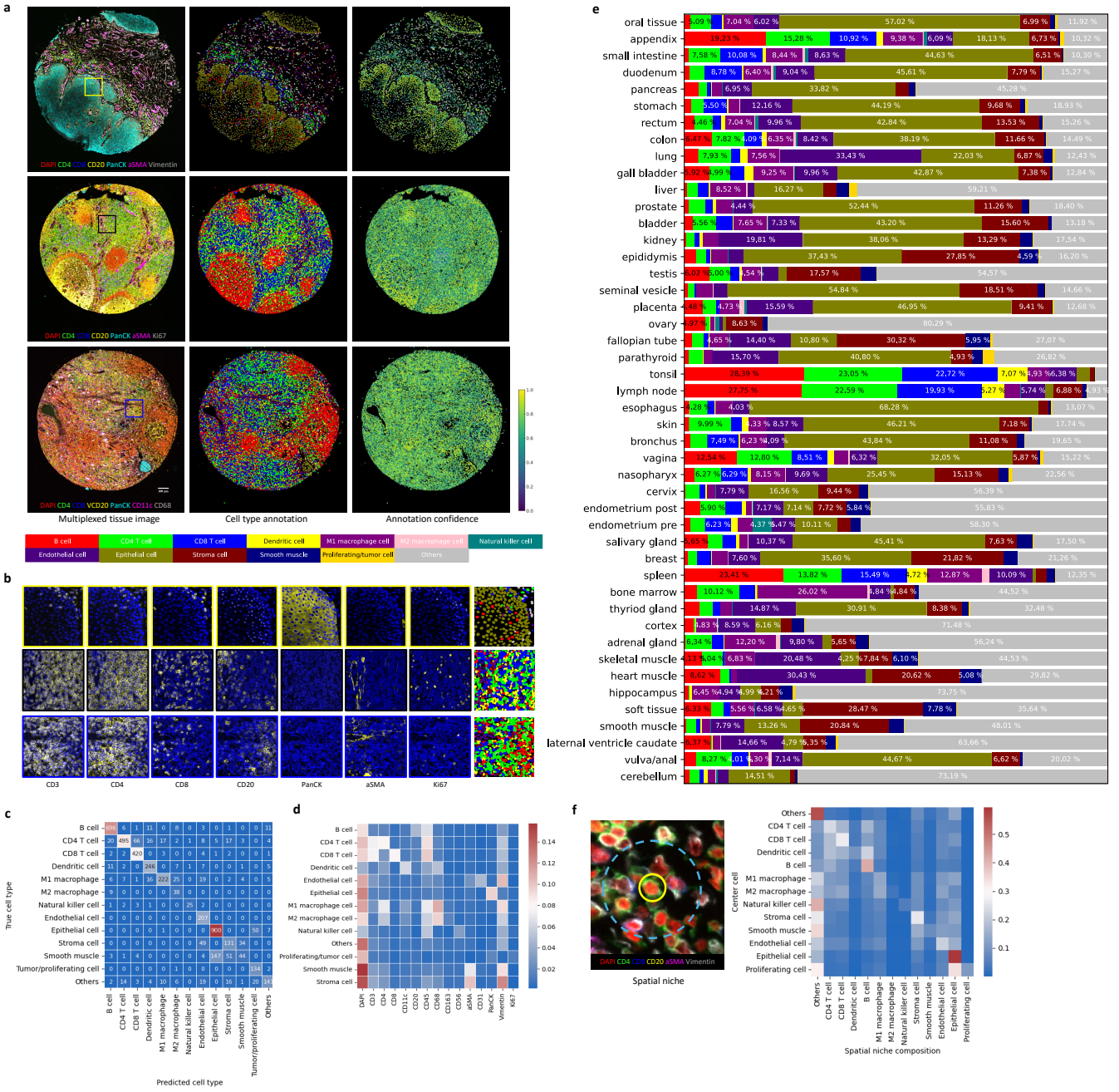

Fig. 2. Results for diverse tissues. **a**, Images of DAPI and six selected markers or TMA cores (left) from the stomach, lymph node, and seminal vesicle (top to bottom). along with cell-level annotations (middle), and associated annotation confidence (right). **b**, Magnified regions (boxes in left column above) highlight cell nuclei (blue) and seven selected markers (yellow). The rightmost column shows the matching cell type annotations for this region using the same color scheme as above. **c**, Confusion matrix comparing cell type annotations by our model with manual annotation for selected TMA cores. **d**, Average protein expression for cell types annotated by the ensemble model across all TMA images. **e**, Cell type distributions derived from our annotations across different tissue types. **f**: Neighboring cell composition for different cell types, with each row showing the normalized average fraction of neighboring cell types for the associated cell type.

of approximately 82.2% in fine-grained cell type annotation (2**c**). RIBCA generally predicts most cell types accurately, with the exception of stromal cells, which are often misclassified as epithelial cells. Additionally, the average marker expression across all TMA images (Figure 2**d**) and the correspondence between cell types and their signature markers meet expectations. We used RIBCA to efficiently annotate over 1 million cells in `SPHPAData_1`, revealing the spatially resolved cell type composition across different tissue samples, as shown in Figure 2**e**. The relative ratio of cells in the different tissues matched their previously reported and widely accepted functions. Tubular organs (e.g., GI tract) have a higher epithelial proportion than solid organs (liver and spleen). In the lymphatic system (lymph nodes, spleen, thymus), T and B cells dominate, while macrophages are more abundant in the liver, where Kupffer cells play a key role in maintaining homeostasis (7, 8).

RIBCA also enhances downstream analyses, such as tumor microenvironment and cell interaction studies. Figure 2**f** shows the average compositions of cell types within spatial

niches (15 neighboring cells), facilitating the investigation of cell interaction heterogeneity. Specifically, we found that (1) dendritic cells have the most diverse neighbor composition, typically including various immune cells, reflecting their role as antigen-presenting cells crucial for immune cell differentiation and proliferation. B cells tend to cluster and form germinal centers in secondary lymphoid tissues, where they proliferate and activate. (2) Epithelial and stromal cells, as supporting tissues, are naturally located near each other, providing the structural foundation of the body. Highly proliferative epithelial cells are often interspersed among other epithelial cells or located in the epithelium's uppermost layer. These neighborhood relationships highlight inherent cell functions. Exploring the frontier of spatial biology involves identifying cell types at the single-cell level in multiplexed proteomic images. Our model enables fast, unbiased, and automated cell type annotation for any imaging-based tissue proteomic data. It is particularly useful for exploration of complex spatial organization of different types of cells *in situ*. The models were empirically validated to be robust and generalizable to new images acquired by different assays. In general, RIBCA unlocks a major bottleneck in spatial proteomics, and the modular design allows the community to expand it to cover more cell types and states.

Availability. RIBCA is available at `https://github.com/sun-huangqingbo/multiplexed-image-annotator`. Since the size of the full `SPData_1` dataset prevents its inclusion in the repository, we include scripts (`https://github.com/sun-huangqingbo/SPData_1-curation` to reproduce the processing of the original dataset images to our cropped cell images. The raw image sources are listed in Table S1.

The HPA tissue microarray multiplexed images will be available in the Human Protein Atlas (v24).

ACKNOWLEDGEMENTS

This work was supported in part by the Wallenberg Foundation (2021.0346) and as part of the PROMINENT team supported by the Cancer Grand Challenges partnership funded by Cancer Research UK (CGCATF-2021/100010), the National Cancer Institute (OT2CA278713), and the Scientific Foundation of the Spanish Association Against Cancer, AECC, and by grant OT2 OD033761 from the National Institutes of Health Common Fund.

Bibliography

1. Hongkui Zeng. What is a cell type and how to define it? *Cell*, 185(15):2739–2755, 2022.
2. Weiruo Zhang, Irene Li, Nathan E Reticker-Flynn, Zinaida Good, Serena Chang, Nikolay Samusik, Saumyaa Saumyaa, Yuanyuan Li, Xin Zhou, Rachel Liang, et al. 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.
3. Yael Amitay, Yuval Bussi, Ben Feinstein, Shai Bagon, Idan Milo, and Leeat Keren. Cellsighter: a neural network to classify cells in highly multiplexed images. *Nature communications*, 14(1):4302, 2023.
4. Ajit Johnson Nirmal, Clarence Yapp, Sandro Santagata, and Peter Sorger. Cell spotter (cspot): A machine-learning approach to automated cell spotting and quantification of highly multiplexed tissue images. *bioRxiv*, pages 2023–11, 2023.
5. Lorenz Rumberger, Noah F Greenwald, Jolene Ranek, Potchara Boonrat, Cameron Walker, Jannik Franzen, Sricharan Varra, Alex Kong, Cameron Sowers, Candace C Liu, et al. Automated classification of cellular expression in multiplexed imaging data with nimbus. *bioRxiv*, pages 2024–06, 2024.
6. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
7. Oliver Krenkel and Frank Tacke. Liver macrophages in tissue homeostasis and disease. *Nature Reviews Immunology*, 17(5):306–321, 2017.
8. Ron Sender, Yarden Weiss, Yoav Navon, Idan Milo, Nofar Azulay, Leeat Keren, Shai Fuchs, Danny Ben-Zvi, Elad Noor, and Ron Milo. The total mass, number, and distribution of immune cells in the human body. *Proceedings of the National Academy of Sciences*, 120 (44):e2308511120, 2023.
9. Muhammad Shaban, Yunhao Bai, Huaying Qiu, Shulin Mao, Jason Yeung, Yao Yu Yeo, Vignesh Shanmugam, Han Chen, Bokai Zhu, Jason L Weirather, et al. Maps: Pathologist-level cell type annotation from tissue images through machine learning. *Nature Communications*, 15(1):28, 2024.
10. 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, et al. Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. *Cell*, 182(5):1341–1359, 2020.
11. John W Hickey, Winston R Becker, Stephanie A Nevins, Aaron Horning, Almudena Espin Perez, Chenchen Zhu, Bokai Zhu, Bei Wei, Roxanne Chiu, Derek C Chen, et al. Organization of the human intestine at single-cell resolution. *Nature*, 619(7970):572–584, 2023.
12. Tobias Roider, Marc A Baertsch, Donnacha Fitzgerald, Harald Voehringer, Berit J Brinkmann, Felix Czernilofsky, Mareike Knoll, Laura Llaó-Cid, Anna Mathioudaki, Bianca Faßbender, et al. Multimodal and spatially resolved profiling identifies distinct patterns of t cell infiltration in nodal b cell lymphoma entities. *Nature Cell Biology*, pages 1–12, 2024.
13. Shovik Bandyopadhyay, Michael P Duffy, Kyung Jin Ahn, Jonathan H Sussman, Minxing Pang, David Smith, Gwendolyn Duncan, Iris Zhang, Jeffrey Huang, Yulieh Lin, et al. Mapping the cellular biogeography of human bone marrow niches using single-cell transcriptomics and proteomic imaging. *Cell*, 187(12):3120–3140, 2024.
14. Leeat Keren, Marc Bosse, Diana Marquez, Roshan Angoshtari, Samir Jain, Sushama Varma, Soo-Ryum Yang, Allison Kurian, David Van Valen, Robert West, et al. A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging. *Cell*, 174(6):1373–1387, 2018.
15. Tyler Risom, David R Glass, Inna Averbukh, Candace C Liu, Alex Baranski, Adam Kagel, Erin F McCaffrey, Noah F Greenwald, Belén Rivero-Gutiérrez, Siri H Strand, et al. Transition to invasive breast cancer is associated with progressive changes in the structure and composition of tumor stroma. *Cell*, 185(2):299–310, 2022.
16. Erin F McCaffrey, Michele Donato, Leeat Keren, Zhenghao Chen, Alea Delmastro, Megan B Fitzpatrick, Sanjana Gupta, Noah F Greenwald, Alex Baranski, William Graf, et al. The immunoregulatory landscape of human tuberculosis granulomas. *Nature immunology*, 23 (2):318–329, 2022.
17. Shirley Greenbaum, Inna Averbukh, Erin Soon, Gabrielle Rizzuto, Alex Baranski, Noah F Greenwald, Adam Kagel, Marc Bosse, Eleni G Jaswa, Zumana Khair, et al. A spatially resolved timeline of the human maternal–fetal interface. *Nature*, 619(7970):595–605, 2023.
18. Jens Hansen, Rachel Sealfon, Rajasree Menon, Michael T Eadon, Blue B Lake, Becky Steck, Kavya Anjani, Samir Parikh, Tara K Sigdel, Guanshi Zhang, et al. A reference tissue atlas for the human kidney. *Science advances*, 8(23):eabn4965, 2022.
19. Rumana Rashid, Giorgio Gaglia, Yu-An Chen, Jia-Ren Lin, Ziming Du, Zoltan Maliga, Denis Schapiro, Clarence Yapp, Jeremy Muhlich, Artem Sokolov, et al. Highly multiplexed immunofluorescence images and single-cell data of immune markers in tonsil and lung cancer. *Scientific data*, 6(1):323, 2019.
20. Jacob H Levine, Erin F Simonds, Sean C Bendall, Kara L Davis, D Amir El-ad, Michelle D Tadmor, Oren Litvin, Harris G Fienberg, Astraea Jager, Eli R Zunder, et al. Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1):184–197, 2015.
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.
22. 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.
23. 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.
24. Stanley R Sternberg. Biomedical image processing. *Computer*, 16(01):22–34, 1983.
25. Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nature methods*, 18(1):100–106, 2021.
26. Marius Pachitariu and Carsen Stringer. Cellpose 2.0: how to train your own model. *Nature methods*, 19(12):1634–1641, 2022.
27. 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, et al. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nature biotechnology*, 40(4):555–565, 2022.
28. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
29. Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.
30. Angelika Danielsson, Fredrik Pontén, Linn Fagerberg, Björn M Hallström, Jochen M Schwenk, Mathias Uhlén, Olle Korsgren, and Cecilia Lindskog. The human pancreas proteome defined by transcriptomics and antibody-based profiling. *PLoS One*, 9(12):e115421, 2014.
31. Mathias Uhlen, Per Oksvold, Linn Fagerberg, Emma Lundberg, Kalle Jonasson, Mattias Forsberg, Martin Zwahlen, Caroline Kampf, Kenneth Wester, Sophia Hober, et al. Towards a knowledge-based human protein atlas. *Nature biotechnology*, 28(12):1248–1250, 2010.
32. Caroline Kampf, IngMarie Olsson, Urban Ryberg, Evelina Sjöstedt, and Fredrik Pontén. Production of tissue microarrays, immunohistochemistry staining and digitalization within

the human protein atlas. *JoVE (Journal of Visualized Experiments)*, (63):e3620, 2012.

Materials and Methods

Dataset curation for ensemble model. We collected 15 datasets from various sources, as listed in Table S1. For each dataset, we first performed sanity checks by randomly picking a few images and visually inspecting them to verify that they corresponded to the descriptions in the original papers. For published datasets without cell segmentation, we performed cell segmentation using CellPose 3.0. For datasets with annotated cell types but no cell segmentation masks, we did an additional step that maps our cell segmentation with the cell coordinates provided to ensure the correspondence between the images and cell type labels. For provided cell coordinates that did not match any of our segmented cells, we generated a circular pseudo-mask centered at that coordinate as a replacement for cell mask.

We also created a cell type marker list for each individual model (which we refer to as a panel) as shown in Table 1, and matched markers associated with each dataset to our panels accordingly. In cases where a marker was absent, we substituted it with a blank image, under the assumption that the absence of a specific marker indicates the corresponding cell type is not present in that dataset. To minimize these substitutions, in rare cases, we replaced some channels with closely related markers; for example, we used CD21 instead of CD20 and occasionally CD38 instead of CD138 in lymphoma datasets (9) for plasma cells. Although these markers serve distinct functional roles, they are considered interchangeable to some extent as cell type markers. Due to differences in granularity of cell type annotations and markers across different dataset, we aimed to maximize the use of the original labels. We manually assigned cell types for panels at different levels of granularity. The correspondence (or assignments) of original cell type annotations to our consolidated cell types (and panels) can be found in Table2 S5-S9.

We used the unlabeled dataset for pretraining and the labeled dataset for fine-tuning. The datasets are summarized below. Scripts to regenerate the collection of datasets are provided (see Availability).

1. The Human colorectal cancer CODEX dataset consists of 140 tissue CODEX images from 35 advanced-stage colorectal cancer patients, using a panel of 56 proteins. This dataset consists of over 220,000 cells with 29 cell types. Cell types were identified through a process of manual gating, clustering, and assigning cluster groups to specific cell types. (10)

2. The Human Intestine CODEX dataset comprises 64 CODEX images of human colon tissue, with a panel of 47 proteins. It includes over 210,000 cells, with 21 distinct cell types. Cell type annotations were performed through clustering, followed by manual labeling of the clusters. (11)

3. The Lymphoma CODEX dataset includes 35 tumor-free lymphoid tissue cores from 19 patients, analyzed using a panel of 53 proteins. It comprises over 5.3 million cells, classified into 19 distinct cell types. Cell

type annotation was performed using Leiden-based clustering on phenotypic markers, followed by merging of related clusters. (12)

4. The bone marrow and Acute Myeloid Leukemia (AML) CODEX dataset contains 7 AML and 12 bone marrow CODEX images, with a panel of 54 proteins. It includes over 1 million cells, with 37 distinct cell types. Cell type annotations were performed through Louvain clustering, manual annotation of clusters, and lateral spillover correction. (13)

5. The triple negative breast cancer MIBI-TOF dataset include 64 MIBI data from triple negative breast cancer, with a panel of 36 proteins. It includes over 60,000 cells, with 17 distinct cell types. Cell type annotations were performed by clustering in a hierarchical scheme, including FlowSOM of immune and non-immune cells and merging similar clusters by hierarchical clustering. (14)

6. The breast cancer progression MIBI dataset contains 79 MIBI samples from patients with ductal carcinoma in situ and invasive breast cancer, using a panel of 37 proteins. It includes over 200,000 cells. This dataset does not include annotations. (15)

7. The human tuberculosis MIBI-TOF dataset contains 17 MIBI samples from tuberculosis patients, using a panel of 29 proteins. It includes over 22,000 cells, representing 20 distinct cell types. Cell type annotations were performed using FlowSOM clustering. (16)

8. The fetally derived extravillous trophoblast MIBI dataset contains 211 images from the human maternal-fetal interface, utilizing a panel of 32 proteins. It includes over 400,000 cells, representing 25 distinct cell types. Cell type annotations were performed using a deep learning pipeline, supported by 93,000 manual annotations. (17)

9. The human kidney CODEX image dataset consists of 25 images from 56 healthy adult kidney tissues, using a panel of 21 proteins. It includes over 860,000 cells. This dataset does not have annotations. (18)

10. The HuBMAP 29-marker CODEX dataset contains 10 images from healthy adult spleen, thymus, and lymph node tissues. It includes over 940,000 cells. This dataset does not have annotations.

11. The classic Hodgkin Lymphoma CODEX dataset consists of one image from cHL patients, utilizing a panel of 49 proteins. It includes over 130,000 cells, representing 17 distinct cell types. Cell type annotations were conducted using traditional iterative clustering and visual inspection. (9)

12. The Tonsil CODEX Image I dataset consists of 45 images from healthy tonsil tissue, using a panel of 88 proteins. It includes over 110,000 cells. This dataset does not contain annotations.

13. The multi-tumor CODEX image dataset comprises 61 images from various tumor tissues, using a panel of 88 proteins. It includes over 240,000 cells. This dataset does not contain annotations.

14. The lung dataset comprises 3 images from human lung cancer tissue, using a panel of 27 proteins. It includes over 297,000 cells. This dataset does not contain annotations. (19)

15. The cHL CODEX dataset contains one FOV obtained using a 50-marker CODEX panel, comprising approximately 145,000 cells. An iterative approach with Rphenoannoy and FlowSOM (20, 21) categorized the majority of these cells into 16 distinct phenotypes, each averaging over 8,000 cells.

16. The Tonsil CODEX dataset II consists of one FOV image of human tonsil tissue, detected using a panel of 56 antibodies, as presented in (22). Over 120,000 cells were identified and grouped into 10 common cell types.

Cell Type Annotation Benchmarking Comparison. We compared results from the ViT model with those from three other recently published representative methods, CellSighter (3) and MAPS (9), and one classic method, STELLAR (23), by testing them on 5 public datasets with their provided software. The five data sets are the cHL MIBI, 4 FOVs of human small intestine (including over 150,000 cells, as used for benchmarking in (23)), cHL CODEX, tonsil CODEX datasets, and one additional Melanoma Lymph Node dataset that was not included in `SPData_1` (published in (3)). This dataset includes 16 FOVs of MIBI images from lymph nodes affected by melanoma metastases across multiple patients, using a panel of 25 protein markers. Around 115,000 cells were identified through a process involving FlowSOM, gating, and visual inspection, followed by manual annotation. This approach led to the identification of 15 distinct cell types.
For each dataset, we randomly partitioned it into a small training dataset (10% of total data) and a validation set (90%), where each cell image/profile in the training set is accompanied by their associated cell type annotation. We independently repeated all tests 7 times and reported the average scores.

Full Field of View Image Preprocessing. Our method applies to registered and stitched 2-dimensional multiplexed images. (it can be extended for 3-dimensional images in general.) First, we subtracted the background signal using the rolling-ball method for each image channel individually (24). Next, for each image channel, the single-channel intensity was clipped at the 98th percentile of nonzero pixels, i.e.,

$$\mathbf{x}_{\text{clipped}} = \min\left(\mathbf{x}, q_{0.98}(\mathbf{x}^{+})\right)$$

given $\mathbf{x}$ denotes the (flattened) image vector, $\mathbf{x}^{+}$ denotes the set of nonzero elements in $\mathbf{x}$ and $q_{0.98}(\cdot)$ as the 98th percentile. Then, the clipped single-channel image was linearly

Panel Name	Marker List	Cell Type names
Basic Panel (Immune)	CD45, CD79A, CD4, CD8, DAPI, CD11c, CD3	B cells, CD4 T cells, CD8 T cells, Dendritic Cell
Extended Panel (Immune)	DAPI, CD3, CD4, CD8, CD11c, CD20, CD45, CD68, CD163, CD56	CD4 T cell, CD8 T cell, Dendritic cell,B cell, Macrophage, M2 Macrophage, NK cell
Full Panel (Immune)	DAPI, CD3, CD4, CD8, CD11c, CD15, CD20, CD45, CD56, CD68, CD138, CD163, FoxP3, Granzyme B, Tryptase	CD4 T cell, CD8 T cell, Dendritic cell,B cell, Macrophage, M2 Macrophage, Treg cell, Granulocyte (Neutrophil), Plasma cell, NK cell, Mast cell
Structure Panel	DAPI, aSMA, CD31, CK, Vimentin, Ki67, CD45	Stroma, Smooth Muscle, Endothelial, Epithelial, Proliferating/Tumor
Nerve Panel	DAPI, CD45, GFAP/ Chromogranin A	Nerve cell

Table 1. Panel names, marker lists and corresponding cell types

scaled into the range of -1 to 1, i.e.

$$\mathbf{x}_{\text{rescaled}} = \frac{2\mathbf{x}_{\text{clipped}}}{\max(\mathbf{x}_{\text{clipped}})} - 1.$$

Finally, for images acquired using multiplexed ion beam imaging (MIBI) method, we applied Gaussian blurring with $\sigma = 0.5$ to each channel of the image.
Our approach assumes cell-level segmentation associated with the multiplexed images in use. Segmentation can be obtained from tools like Cellpose (25, 26) and Deepcell Mesmer (27). In this work, we used the cell segmentation provided with the datasets or Cellpose 3.0 to segment cells when the segmentation was not provided.

Soft-masking of Individual Cell Images. For each segmented cell, a square image tile with a side length of 40 pixels centered on the cell was cropped out of the full image. However, since the resulting single-cell image may still contain protein patterns belonging to its neighboring cells, we applied a *soft-masking* operation; the softened cell mask is essentially an average of original cell mask being morphologically dilated and Gaussian blurred using different parameters. That is, for a cell image, we denote its associated cell mask (from the original segmentation) as $\mathbf{b} \in \{0,1\}^{40\times 40}$, and refer $\texttt{dilate}(\cdot, r)$ to the binary dilation operation of an image using a disk-shaped kernel, with r specifying the radius, and $\texttt{blur}(\cdot, r)$ as the Gaussian blur operation with kernel radius r. Then, a soft-mask is written as $\tilde{\mathbf{b}} = \frac{1}{10}\sum_{i=1}^{4}\sum_{j=1}^{i} \texttt{blur}(\texttt{dilate}(\mathbf{b}, j), i)$. To apply soft-masking to individual images, we simply perform element-wise multiplication between $\tilde{\mathbf{b}}$ and each image channel. Intuitively, soft-masking produces a smoother transition from the inside to outside of the cell region, avoiding sudden changes of intensity on the edges of the cell region.

Vision Transformer. ViT is a widely used neural network design introduced by Dosovitskiy *et al.* (6). We briefly review its mechanism here. The idea of ViT is to partition an image into a sequence of non-overlapping patches, and use a Transformer encoder neural network originally designed for 1-dimensional sequences to transform this array of image patches. In particular, given an input image $\mathbf{x} \in \mathbb{R}^{C\cdot H\cdot W}$, it will be cropped (on the plane of $H \times W$) into a grid of patches in size $C \cdot P \cdot P$, resulting in $n = \lfloor \frac{H}{P} \rfloor \cdot \lfloor \frac{W}{P} \rfloor$ patches in total denoting as $[x_{n_1,n_2}]_{n_1=1,n_2=1}^{\lfloor \frac{H}{P} \rfloor \lfloor \frac{W}{P} \rfloor}$. Each patch x_{n_1,n_2}

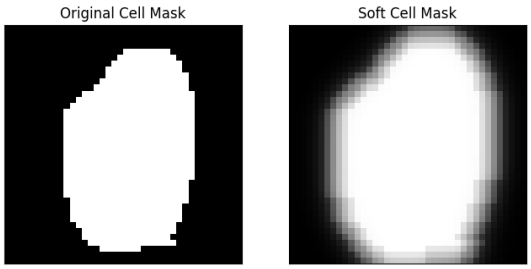

Fig. 3. An example of applying soft-masking to a cell segmentation mask.

is then flattened as a 1-dimensional vector and linearly transformed to dimensionality of d, *i.e.* $z^0_{n_1,n_2} = W^0 x^{\text{flatten}}_{n_1,n_2}$, $W \in \mathbb{R}^{d\times C\cdot P\cdot P}$. These transformed vectors are also called tokens of this image. Following the original design of ViT, a `class` token (a learnable L-dimensional vector) is prepended to this sequence of tokens, $\mathbf{z}^0 = \left[z_{\texttt{class}}, z^0_{1,1}, \ldots, z^0_{\lfloor \frac{H}{P} \rfloor, \lfloor \frac{W}{P} \rfloor}\right]$, the intuition of this operation is to use this additional vector as a summarized feature of the rest of transformed patch vectors after ViT encoding. To preserve the spatial arrangement of the patches in the original image, another sequence of d-dimensional vectors generated by periodic functions is added to $\mathbf{z}^0$ to indicate the position of each patches in the original image, realized by $\mathbf{z}^1 = \mathbf{z}^0 + \mathbf{P}$, $\mathbf{P} \in \mathbb{R}^{n+1\times d}$,

$$\mathbf{P}_{k,2i} = \sin\left(\frac{k}{10000^{2i/d}}\right) \quad \mathbf{P}_{k,2i+1} = \cos\left(\frac{k}{10000^{2i/d}}\right).$$

Next, these positional-aware tokens are further encoded by a multi-head self-attention based Transformer encoder neural network as described (28). We denote the encoded tokens as $\mathbf{z}^{\text{enc}} = \texttt{enc}\left(\mathbf{z}^1\right) \in \mathbb{R}^{n+1\times d}$. To map the encoded tokens onto the classification decision, a linear map is applied to the encoded `class` token, *i.e.* the first element of $\mathbf{z}^{\text{enc}}$, say $\mathbf{z}^{\text{enc}}_0$ and $\hat{y} = W_2\mathbf{z}^{\text{enc}}_0 + b_2$ where $W_2 \in \mathbb{R}^{m\times d}$ and $b \in \mathbb{R}^m$, m is the number of classes and $\hat{y}_u n$ denotes the unnormalized probabilities.

Masked Auto-encoding. Dosovitskiy *et al.* (6) showed that unsupervised pre-training of a ViT encoder can improve its performance of image classification. A popular pre-training method is masked auto-encoding MAE. MAE aims to recover the original full image from a partial observation of that image. The realization of MAE relies on a ViT as an image encoder. Instead of receiving all patches cropped from the

input image, during MAE pre-training, only a small subset of patches are *observed* and used as the input to the ViT encoder (and no `class` prepended). To decode the image from tokens, the encoded tokens associated with the observed patches are placed in their original position in the image, and the positions associated with unobserved patches are filled by a vector (the same size as each encoded token) called mask representation (a model parameter used as the representations of unobserved patches), this sequence of tokens are then decoded by a simpler decoder model. That is, only $m = \lfloor(1 - \text{mask ratio})n\rfloor$ patches are observed, denoting their indices as I. They are linearly transformed to d-dimensionality, added with their associated positional tokens, and encoded by the ViT encoder as, $[z_{n_1,n_2}]_{(n_1,n_2)\in I} = \texttt{enc}\left([z^1_{n_1,n_2}]_{(n_1,n_2)\in I}\right)$, then the mask representation will be inserted, i.e. $[\tilde{z}_{n_1,n_2}]_{n_1=1,n_2=1}^{\lfloor\frac{H}{P}\rfloor\lfloor\frac{W}{P}\rfloor}$ where $\tilde{z}_{n_1,n_2} = z_{n_1,n_2}$ if $(n_1, n_2) \in I$, otherwise, $\tilde{z}_{n_1,n_2} = \mathbf{m}$. Here, $\mathbf{m}$ denotes the token associated with any masked patches. A decoder will decode the these tokens to reconstruct the original image, as $[\hat{x}_{n_1,n_2}]_{n_1=1,n_2=1}^{\lfloor\frac{H}{P}\rfloor\lfloor\frac{W}{P}\rfloor} = \texttt{dec}\left([\tilde{z}_{n_1,n_2}]_{n_1=1,n_2=1}^{\lfloor\frac{H}{P}\rfloor\lfloor\frac{W}{P}\rfloor}\right)$, and the mean square error (MSE) is calculated between the original signal and this reconstruction. The intuition of MAE is to use a simple but effective way to learn spatial relationships between different parts of images without any references and pre-train a scalable visual representation learning model. After MAE pre-training, only the pre-trained ViT encoder is preserved.

To apply it for cell type identification tasks, we pre-trained our ViT encoder on images cropped from a full-view multiplexed tissue image, each containing only one single centered cell with very small margins around. In practice, we encoded each individual cell image (as described in Methods Section) using the ViT encoder, as illustrated in Figure 1**d**. We used a small ViT as encoder; the hyperparameters of our models are listed in Table S2. The models were trained to minimize the average mean square error of each mini-batch using an Adam optimizer with an adaptive learning rate of 10^{-3}.

To train a channel-wise ViT using the MAE approach, we first split a multi-channel image into individual image channels, and tile them into a larger single-channel image. For example, given an image with the size of $C \times H \times W$ and $C = ab$, a, b are integers, we can tile it into a new image with a size of $1 \times aH \times bW$. By choosing the patch size equal to $1 \times H \times W$ which results in C patches in total, we can train a ViT using the MAE technique described in this section by randomly masking patches in this larger image (that is, individual image channel of the original image). In particular, we did not use a fixed number of masked patches (image channels) during this training, rather we randomly masked nearly half (plus or minus two) of the patches to train this model.

Fine-tuning. After pre-training, we fine-tuned the ViT model following a standard classification task training paradigm. The weights of the ViT encoder parameter are borrowed from the pre-trained model, and the rest of parameters are initialized randomly according to a Gaussian distribution (standard deviation of 0.02). Given some reference cell type annotation, say $\{\mathbf{x}_i, y_i\}_{i=1}^n$ provided $\mathbf{x}_i, y_i, i \in [n]$ are single-cell images and their associated cell type labels, the ViT encoder encodes $\mathbf{x}_i$ into tokens. Next, the encoded `class` token is mapped onto unnormalized probabilities $\hat{y}_u n$ as described previously, which are further normalized by a Sigmoid function. The normalized predicted probability is used to calculate the label smoothing cross-entropy loss (29), as

$$l(y, \hat{y}) = -\sum_{i=1}^{m} \tilde{y}_i \log(\hat{y}_i),$$

where $\tilde{y}_i = (1 - \epsilon) \cdot y_i + \frac{\epsilon}{m}$ and m is the number of classes. Whole ViT model is trained by an AdamW optimizer with a learning rate of 2×10^{-4}.

Merging predictions. As the ensemble model possesses multiple individual models, a necessary step before final classification is to merge the decisions from applied models. For the three immune panels, the larger model is prioritized. For example, when both extended and base models can be used, only the extended model is applied. To merge the predictions across other panels, the predicted normalized probabilities (scores) from all relevant models are collected. To establish a cutoff for determining whether a specific cell type should be assigned, we take the lower of (1) the minimum of all scores for the "Others" class and (2) a user-specified hyperparameter (set to 0.3 for the analyses here). If the cell type with the highest score (other than the class of "Others") is greater than this value, it is assigned; othewise it is assigned "Others".

Runtime. For a single TMA core image with 10,000 cells, it takes about 5 to 10 minutes to generate cell type annotation maps and all downstream analyses on an ordinary desktop computer.

Expert Validation of the Ensemble Model. To validate the assigned cell types, we chose a small region of interest in each TMA and checked their corresponding marker signals. To ensure the composition of different cell types, we chose the region based on the criteria that the area contains different compartments. Based on the spatial organization of markers, certain tissues appear to have a clear compartmentalization. These compartments were visually determined by the enrichment of markers. As an example, for intestine tissue, at least two layers should be included, which would include both epithelial and immune cells. Epithelial layers are enriched in cytokeratin expression and immune cells layer are enriched in CD45 signals, with a clear boundary between them.We want to ensure that the model is able to distinguish between different tissue compartments at the boundary where the enrichment of cells is different, and label assignment is not affected by signal spillover. For each selected region, we manually validated the cell assignments using the Napari plugin we developed. This plugin allows us to visualize both RIBCA's annotation and multiplexed data together. For each cell, we validated its annotation by checking if its signature

markers were fully expressed. Cells with valid marker expression were considered correctly classified. When discrepancies were observed, we manually reviewed the phenotypic markers to identify the positive ones and reassign the cell to the correct type. We repeated this procedure for each selected cell to plot the confusion matrix as shown in Figure 2**c**.

Cell Type	Phenotypic Marker
CD4 T cell	CD3, CD45, CD4
CD8 T cell	CD3, CD45, CD8
Dendritic cell	CD11c
B cell	CD20
Macrophage	CD68
M2 Macrophage	CD163
Treg cell	CD3, CD45, FoxP3
Granulocyte (Neutrophil)	CD15
Plasma cell	CD138
NK cell	CD56
Mast cell	Granzyme B, Tryptase
Stroma	Vimentin
Smooth Muscle	aSMA
Endothelial	CD31
Epithelial	CK
Tumor	Ki67
Neuron	CHGA/GFAP
Others	N/A (Defined but not included above)
Artifact/not defined	N/A(Not defined in original dataset)

Table 2. Phenotypic Markers and Their Corresponding Cell Types

Human tissue sample preparation. We used the same tissue sample preparation protocol as previously desbribed by Danielsson *et al.* (30). Tissue samples were handled in accordance with Swedish laws and regulation and obtained from the Department of Pathology, Uppsala University Hospital, Sweden, as part of the sample collection governed by the Uppsala Biobank (http://www.uppsalabiobank.uu.se/en/). All human tissue samples used were anonymized in accordance with approval and advisory report from the Uppsala Ethical Review Board (Dnr Ups 02-577). TMAs were generated in accordance with strategies used in the Human Protein Atlas and as previously described (31, 32). In brief, hematoxylin-eosin stained tissue sections from formalin-fixed paraffin-embedded donor blocks were examined in order to verify the histology and select representative regions for sampling into the TMA. Normal tissue was defined as microscopically normal and was most often selected from specimens collected from the vicinity of surgically removed tumors. Two TMA blocks were generated containing triplicate 1mm cores of 45 total types of normal tissue (tissues listed in Table S4).

Spatial targeted proteomic profiling using PhenoCycler Fusion platform. Highly multiplexed protein data were generated using the PhenoCycler Fusion dual-slide system (Akoya Biosciences, Marlborough, MA). We cut a 4 µm thick section of TMA block in a positively charged microscope slide (VWR, Epredia, Cat No. 76406-502) and baked it at 65 °C in an oven, overnight. The next day, the sections were dewaxed manually using HistoChoice (Sigma, Cat No. H2779), followed by rehydration in ethanol solutions of ethanol (100% ethanol twice, 90%, 70%, 50%, 30%, ddH2O twice; each step for 5minutes). We performed then heat-induced epitope retrieval (HIER) treatment using pH 9 EDTA buffer (Akoya Biosciences, Cat No. AR900250ML) in a pressure cooker (Bio SB TintoRetriever, Cat No. BSB-7087) at 114–121 °C during 20 minutes. Once the slides were equilibrated at room temperature, we transferred them to a ddH2O container for a total of two washes, each step of 2 minutes. To reduce the tissue autofluroescence, we performed a photobleaching treatment: we submerged the slides in a transparent slide mailer (Heathrow Scientific, Cat No. HS15982) containing 25 ml 1× PBS (Gibco, Cat No. 18912-014), 4.5 ml 30% H2O2 (Sigma, Cat No. 216763) and 0.8 ml 1M NaOH (Sigma, Cat No. S5881). The slides in solution were exposed to two LED lamps (GLIME, Light Therapy Lamp UV Free 32000), one on each side, and incubated for 45 minutes. We prepared a fresh photobleaching solution and incubated the slides for an additional 45 minutes, followed by four washes in 1× PBS, each for 2 minutes. The slides were then transferred to a Coplin jar containing Hydration Buffer (Akoya Biosciences, Cat No. 7000017) for two 2-minute incubations. Afterward, they were moved to a jar with Staining Buffer (Akoya Biosciences, Cat No. 7000017) and equilibrated for 30 minutes. Meantime, we diluted all the antibodies (antibody dilutions listed in Table S3) together in a cocktail solution containing Staining Buffer and N, J, G and S blockers (Akoya Biosciences, Cat No. 7000017) and added them on the slides for an overnight incubation at 4 °C. The day after, the primary antibodies were removed from the slides through washes on Staining Buffer and then we performed 3 different consecutive post-fixation steps with washes between them: 1) 1.6% paraformaldehyde (ThermoFisher, Cat No. 043368) diluted in Storage Buffer (Akoya Biosciences, Cat No. 7000017) for a 10 minutes incubation at room temperature; 2) ice cold methanol (Sigma, Cat No. 322415) for 5 minutes at 4 °C; 3) Fixative Solution (Akoya Biosciences, Cat No. 7000017) diluted at 1:50 in 1× PBS. The slides were stored in Storage Buffer (Akoya Biosciences, Cat No. 7000017) until they were attached to flow cells and moved to the PhenoCycler Fusion system for analysis. Two 96-well reporter plates, each containing fluorescent dyes conjugated to PhenoCycler oligonucleotide sequences, were prepared, allowing for up to three spectral different dyes plus nuclear stain in each well.

PhenoCycler image acquisition. We performed the image acquisition on the PhenoImager Fusion 2.2.0 (Akoya Biosciences, Marlborough, MA) where three fluorescent oligo reporters with spectrally distinct dyes were applied to the tissue in iterative imaging cycles. The images were captured using a widefield microscope with a Olympus UCPlanFL 20×/0.70 NA objective and a resolution of 0.5 µm/pixel. Whole-slide images were automatically aligned using DAPI as a reference, and background subtracted to be assembled into the the final QPTIFFs image files.