Accurate single-molecule spot detection for image-based spatial transcriptomics with weakly supervised deep learning

Emily Laubscher[1], Xuefei (Julie) Wang[2], Nitzan Razin[2], Tom Dougherty[2], Rosalind J. Xu[3,4,5], Lincoln Ombelets[1], Edward Pao[2], William Graf[2], Jeffrey R. Moffitt[3,4,6], Yisong Yue[7], and David Van Valen[2]

[1] *Division of Chemistry and Chemical Engineering, Caltech, Pasadena, CA*

[2] *Division of Biology and Biological Engineering, Caltech, Pasadena, CA*

[3] *Program in Cellular and Molecular Medicine, Boston Children's Hospital, Boston MA*

[4] *Department of Microbiology, Blavatnik Institute, Harvard Medical School, Boston, MA*

[5] *Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA*

[6] *Broad Institute of Harvard and MIT, Cambridge, MA*

[7] *Division of Computational and Mathematical Sciences, Caltech, Pasadena, CA*

September 1, 2023

Abstract

Image-based spatial transcriptomics methods enable transcriptome-scale gene expression measurements with spatial information but require complex, manually-tuned analysis pipelines. We present Polaris, an analysis pipeline for image-based spatial transcriptomics that combines deep learning models for cell segmentation and spot detection with a probabilistic gene decoder to quantify single-cell gene expression accurately. Polaris is available through the DeepCell software library (`https://github.com/vanvalenlab/deepcell-spots`) and `https://www.deepcell.org`.

Advances in spatial transcriptomics have enabled system-level gene expression measurement while preserving spatial information, enabling new studies into the connections between gene expression, tissue organization, and disease states [1,2]. Spatial transcriptomics methods fall broadly into two categories. Sequencing-based methods leverage arrays of spatially barcoded RNA capture beads to integrate spatial information and transcriptomes [3–6]. Image-based methods, including multiplexed RNA fluorescent in situ hybridization and in situ RNA sequencing, perform sequential rounds of fluorescent staining to label transcripts to measure the expression of thousands of genes in the same sample [7–11]. Because these methods rely on imaging, the data they generate naturally contain the sample's spatial organization. While image-based spatial transcriptomics enables measurements with high transcript recall and subcellular resolution [1,2], rendering the raw imaging data interpretable remains challenging. Specifically, the computer vision pipelines for image-based spatial transcriptomics must reliably perform cell segmentation, spot detection, and gene assignment across diverse imaging data. Prior methods that sought an integrated solution to this problem rely on manually-tuned algorithms to optimize performance for a particular

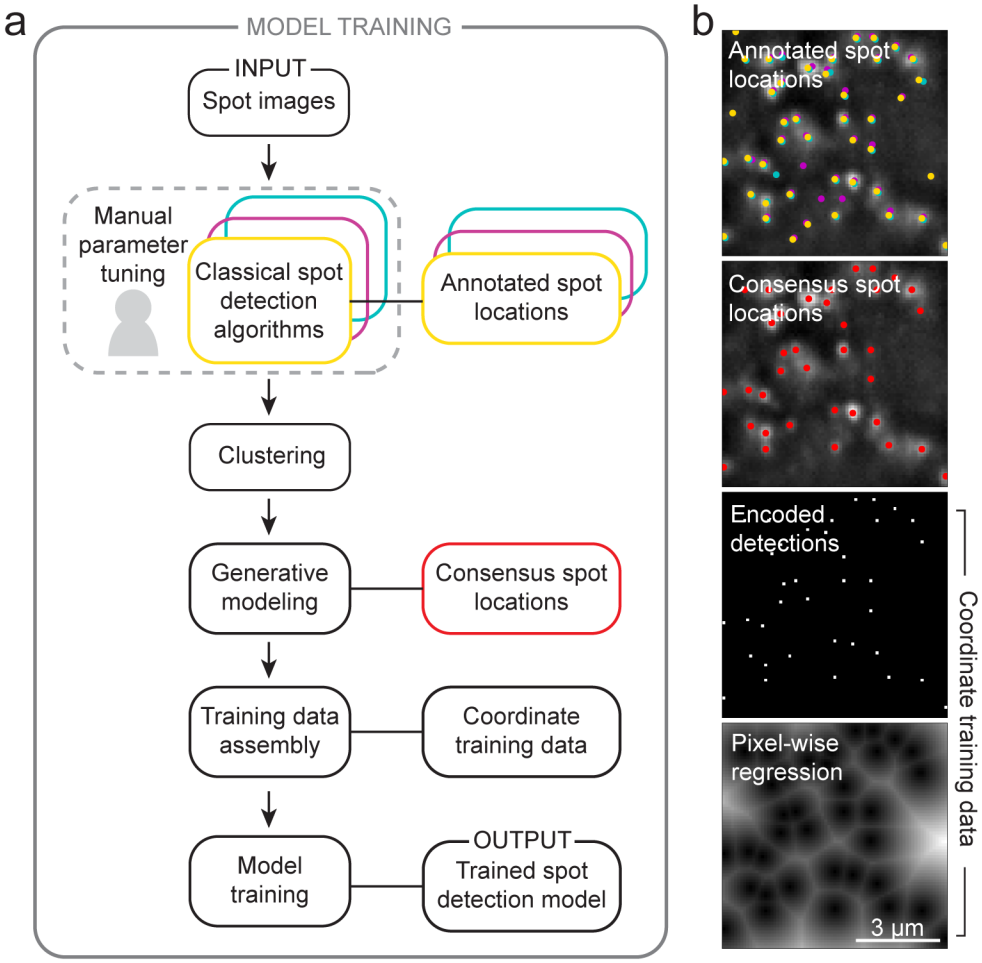

Figure 1: **A Weakly supervised deep learning framework for accurate fluorescent spot detection for spatial transcriptomics imaging data.** (a) Training data generation for spot detection. Spot labels were generated by finding consensus among a panel of commonly used classical spot detection algorithms through generative modeling. These consensus labels were then used to train Polaris' spot detection model. Sequential steps are linked with an arrow; associated methods and data types are linked with a solid line. (b) Demonstration of the training data generation for an example spot image. Spot locations are converted into detections and distance maps which guide the classification and regression tasks performed during model training. Spot colors correspond to the annotation colors in (a).

sample or spatial transcriptomics assay[12,13]. Thus, there remains a need for an integrated, open-source pipeline that can perform these steps reliably across the diverse images generated by spatial transcriptomics assays with minimal human intervention.

Deep learning methods are a natural fit for this problem. Prior work by ourselves and others has shown that deep learning methods can accurately perform cell segmentation with minimal user intervention[14–17], providing a key computational primitive for cellular image analysis. Here, we focus on the problem of spot detection for image-based spatial transcriptomics data. Existing spot detection methods fall into two categories: "classical" and "supervised"[18]. Classical methods are widely used but require manual fine-tuning on each dataset to optimize performance[19,20]. This places a fundamental limit on their scalability. Supervised methods[21–23], which often rely on deep learning methodologies, require labeled training data to learn how to detect spots. This requirement presents a major challenge, as experimentally generated data contain too many spots for manual annotation to be feasible. Training data derived from classical algorithms are limited by the characteristics of those algorithms, imposing a ceiling on model performance. Further, simulated training data lack the artifacts present in

experimentally generated data which can limit their performance on real data.

In this work, we demonstrate that deep learning can be combined with weak supervision to create a universal spot detector for image-based spatial transcriptomics data. In our approach, we first create annotations for representative fluorescent spot images by manually fine-tuning a collection of classical algorithms on each image. Inspired by prior work on programmatic labeling[24], we de-noise conflicting spot annotations with a generative model (Fig. 1a-b) and use these consensus annotations to train a deep learning model for spot detection. For each pixel in an image, our model predicts the probability that the pixel contains a spot and performs a regression to the nearest spot's center. We provide further details of dataset construction, generative modeling, and deep learning methodology in the Supplementary Information. Benchmarking on simulated data demonstrated that our method outperforms models trained with data that was either simulated or labeled with a single classical algorithm (Supplementary Fig. S6). Visual inspection showed that our model generalized to out-of-distribution, spot-like data (Supplementary Fig. S3).

Given that our benchmarking demonstrated that training deep learning models with weak supervision can yield a computational primitive for spot detection, we then constructed Polaris, an integrated deep learning pipeline for image-based spatial transcriptomics (Fig. 2a). Polaris utilizes classical computer vision methods for image alignment and deep learning models for spot detection and cell segmentation. For multiplexed spatial transcriptomics images, we infer gene identities by fitting a probabilistic graphical model to the per-spot probability "intensity" generated by the spot detection model with variational inference[25] (Fig. 2b-c). We found that modeling the distributions of spot probabilities with a mixture of relaxed Bernoulli distributions yielded the best decoding results for most datasets. We demonstrated Polaris' performance on previously published MERFISH data in a mouse ileum tissue sample[26] (Fig. 2d-f), previously published MERFISH data in a mouse kidney tissue sample[27], seqFISH data in a human macrophage cell culture sample (Supplementary Fig. **??**), and previously published in situ sequencing data of a pooled CRISPR library in HeLa cells[28] (Supplementary Fig. S10). We found that Polaris detected marker genes from expected cell types - even in the presence of packed cells with heterogeneous morphologies. We found that Polaris' gene expression counts were consistent with both bulk sequencing data and pipelines that rely on manual parameter tuning (Supplementary Fig. S9). These results demonstrate that Polaris can generalize across sample types, imaging platforms, and spatial transcriptomics assays without manual parameter tuning.

Polaris is available for academic use through the DeepCell software library `https://github.com/vanvalenlab/deepcell-spots`; a singleplex version of the pipeline is available through the DeepCell web portal `https://deepcell.org`. Polaris' outputs are compatible with downstream tools, such as squidpy[29] and Seurat[30]. This work provides a key computational primitive for spot detection and an integrated, open-source pipeline for image-based spatial transcriptomics. We believe Polaris will help make these methods more accessible and facilitate scaling analyses to larger datasets.

Code availability

Code used for cell segmentation and model development is available at `https://github.com/vanvalenlab/deepcell-tf`. The spot detection and gene assignment code is available at `https://github.com/vanvalenlab/deepcell-spots`. An example Jupyter notebook for Polaris is available at `https://github.com/vanvalenlab/deepcell-spots/blob/master/notebooks/applications/Polaris-application.ipynb`. The code used for model deployment is available at

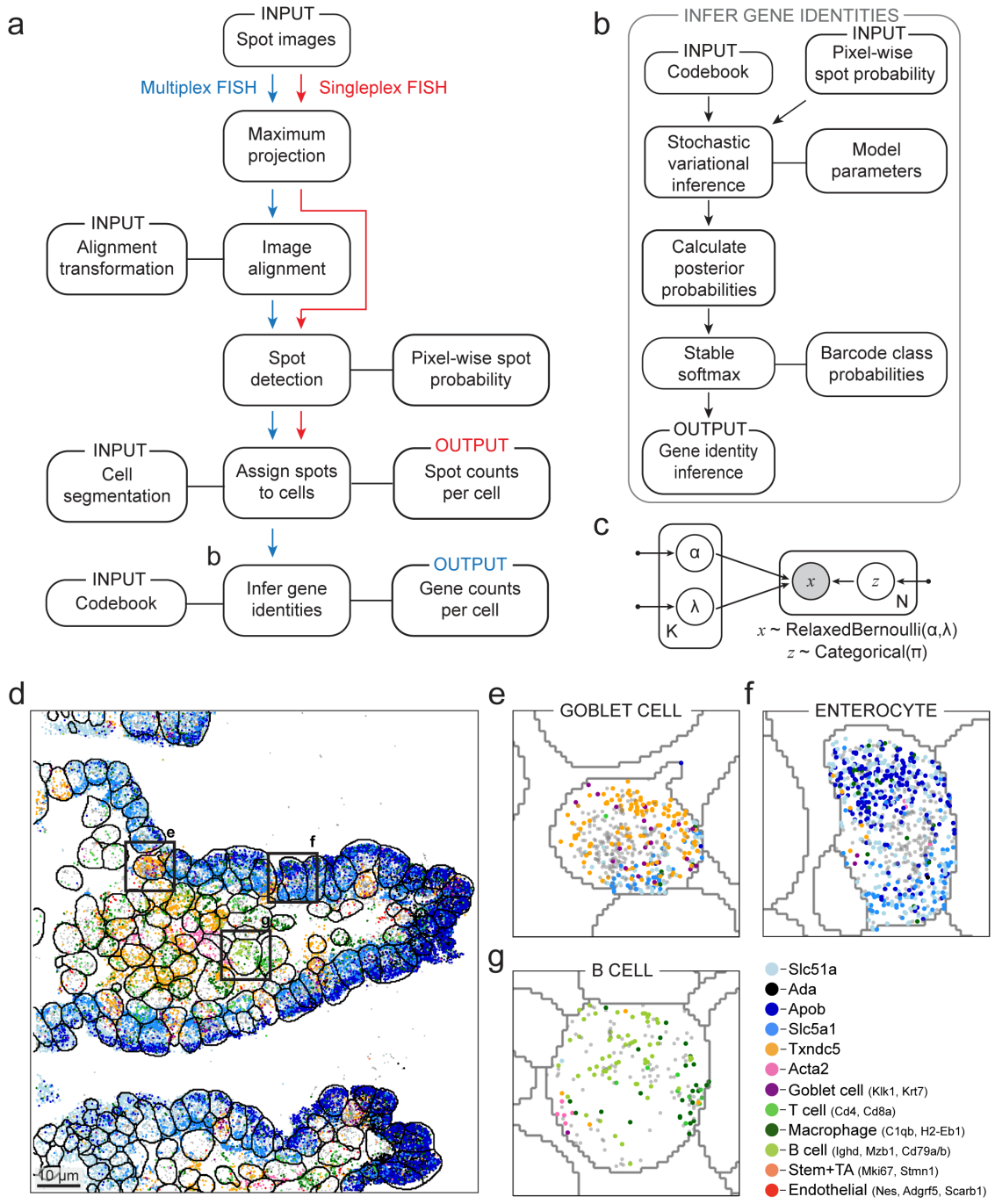

Figure 2: **Polaris produces single-cell, spatial gene expression maps for multiplex spatial transcriptomics images.** (a, b) Analysis steps for Polaris for singleplex (red) and multiplex (blue) spatial transcriptomics imaging data. Sequential steps are linked with an arrow, and associated methods and data types are linked with a solid line. Deep learning models perform spot detection and cell segmentation, while a probabilistic graphical model infers gene identities. (c) A probabilistic graphical model for inferring gene identities from spot detections. This model consists of a mixture of K relaxed Bernoulli distributions, parameterized by their probability, α, and their temperature, λ, for generating observed data, x, of size N spots. Shaded vertices represent observed random variables; empty vertices represent latent random variables; edges signify conditional dependency; rectangles ("plates") represent independent replication; and small solid dots represent deterministic parameters. (d) Spatial organization of marker gene locations in a mouse ileum tissue sample. Each spot corresponds to a decoded transcript for a cell type marker gene. Whole-cell segmentation was performed with Mesmer[16]. (e-g) Locations of decoded genes in an example Goblet cell, enterocyte, and B cell, respectively.

https://github.com/vanvalenlab/kiosk-console. Finally, code for reproducing all models and figures included in the paper is available at https://github.com/vanvalenlab/Polaris-2023_Laubscher_et_al.

Data availability

The data and annotations used to train the spot detection model are available for academic use at https://deepcell.readthedocs.io/en/master/data-gallery.

Acknowledgments

We thank Lior Pachter, Barbara Englehardt, Sami Farhi, Ross Barnowski, and the other members of the Van Valen lab for useful feedback and interesting discussions. We thank Nico Pierson and Jonathan White for contributing data and providing early annotations. The HeLa cell line was used in this research. Henrietta Lacks, and the HeLa cell line established from her tumor cells without her knowledge or consent in 1951, have made significant contributions to scientific progress and advances in human health. We are grateful to Henrietta Lacks, now deceased, and her surviving family members for their contributions to biomedical research. This work was supported by awards from the Shurl and Kay Curci Foundation (to DVV), the Rita Allen Foundation (to DVV), the Susan E Riley Foundation (to DVV), the Pew-Stewart Cancer Scholars program (to DVV), the Gordon and Betty Moore Foundation (to DVV), the Schmidt Academy for Software Engineering (to TD), the Michael J Fox Foundation through the Aligning Science Across Parkinsons consortium (to DVV), the Heritage Medical Research Institute (to DVV), and the NIH New Innovator program (DP2-GM149556) (to DVV).

Author Contributions

EL, NR, and DVV conceived the project. EL, NR, and DVV conceived the weakly supervised deep learning method for spot detection. EL and EP created the training data for the spot detection model. LO contributed to the seqFISH protocol used to create training data. EL developed software for training data annotation. EL curated and annotated the training data for the spot detection model. NR and EL developed the spot detection model training software. EL trained the models. EL and NR developed the metrics software for the spot detection model. XW, EL, YY, and DVV conceived the combinatorial barcode assignment method. EL and XW developed the barcode assignment software, with input from YY. EL developed the multiplex image analysis pipeline. EL and XW benchmarked the multiplex image analysis pipeline. EL and TD developed the cloud deployment. RJX and JRM collected and analyzed MERFISH data. EL and EP created the macrophage seqFISH dataset. WG and DVV oversaw software engineering for Polaris. YY and DVV oversaw the algorithm development for the project. EL and DVV wrote the manuscript, with input from all authors. DVV supervised the project.

Competing Interests

DVV is a co-founder and Chief Scientist of Barrier Biosciences and holds equity in the company. DVV, EL, and NR filed a patent for weakly supervised deep learning for spot detection. JRM is co-founder and scientific advisor to Vizgen and holds equity in the company. JRM is an inventor on patents related to MERFISH filed on his behalf by Harvard University and Boston Children's Hospital. All other authors declare no competing interests.

1 Methods

1.1 Generation of sequential fluorescent in situ hybridization (seqFISH) images for spot training data

1.1.1 Probe design

mRNA transcripts were targeted with single-stranded DNA probes, as previously described[10]. Primary probes were designed to target a panel of 10 genes with OligoMiner using balanced coverage settings[31]. The primary probes were designed to have secondary probe-binding sites flanking both sides of the sequence that binds to the mRNA transcript. The secondary probes were 15 bases long and consisted of nucleotide combinations that were optimized to have 40%-60% GC content and minimal genomic off-target binding.

1.1.2 Probe construction

Single-stranded DNA primary probes were obtained from Integrated DNA Technologies (IDT) as an oPools Oligo Pool. Oligos were received lyophilized and dissolved in ultrapure water at a stock concentration of 1 μM per probe. Single-stranded DNA secondary probes were also obtained from IDT and were 5'-functionalized with Alexa Fluor 647, Alexa Fluor 546, or Alexa Fluor 488. Secondary probes were received lyophilized and dissolved in ultrapure water at a concentration of 100 nM.

1.1.3 Cell culture

HeLa (CCL-2) cells were received from the American Type Culture Collection. The cells were cultured in Eagle's minimum essential medium (Cytiva #SH30024LS) supplemented with 2 mM L-glutamine (Gibco), 100 U/mL penicillin, 100 μg/mL streptomycin (Gibco or Caisson), and 10% fetal bovine serum (Omega Scientific or Thermo Fisher). Cells were incubated at 37°C in a humidified 5% CO_2 atmosphere and were passaged when they reached 70%-80% confluence.

1.1.4 Buffer preparation

The primary probe hybridization buffer consisted of 133 mg/mL high-molecular-weight dextran sulfate (Calbiochem #3710-50GM), 2X saline-sodium citrate (SSC) (IBI Scientific #IB72010), and 66% formamide (Bio Basic #FB0211-500) in ultrapure water. The 55% wash buffer was comprised of 2X SSC, 55% formamide, and 1% Triton-X (Sigma-Aldrich #10789704001) in ultrapure water. The secondary probe hybridization buffer consisted of 2X SSC, 16% ethylene carbonate (Sigma-Aldrich #E26258-500G), and 167 mg/mL of high-molecular-weight dextran sulfate in ultrapure water. The ethylene carbonate was first melted at 50°C for 30-60 minutes. The 10% wash buffer was comprised of 2X SSC, 10% formamide, and 1% Triton-X in ultrapure water. The imaging buffer base consisted of 0.072M Tris HCl (pH 8) (RPI #T60050-1000), 0.43 M NaCl (Fisher #MK-7581-500), and 3 mM Trolox (Sigma-Aldrich #238813) in ultrapure water. The anti-bleaching buffer was comprised of 70% imaging buffer base, 2X SSC, 1% catalase (10X dilution of stock) (Sigma #C3155), 0.005 mg/mL glucose oxidase (Sigma-Aldrich #G2133-10KU), and 0.08% D-glucose (Sigma #G7528) in ultrapure water.

1.1.5 seqFISH sample preparation

The cells were seeded in a fibronectin-functionalized (Fisher Scientific #33010018) glass-bottom 96-well plate (Cellvis P96-1.5H-N) at 80%-90% confluence. Cells were rinsed with warm 1X phosphate-buffered saline (PBS) (Gibco), fixed with fresh 4% formaldehyde (Thermo #28908) in 1X PBS for 10 minutes at room temperature, and then permeabilized with 70% ethanol overnight at -20°C. Prior to probe hybridization, the cells were rinsed with 2X SSC. The primary probes were diluted to 10 nM in the primary probe hybridization buffer and 100 μL of this solution was added to each well. Cells were incubated with the primary probe solution for 24 hours at 37°C. The primary probes were rinsed out twice with 55% wash buffer. Cells were incubated in 55% wash buffer for 30 minutes in the dark at room temperature and then rinsed three times with 2X SSC buffer. The secondary probes were diluted to 50 nM in the secondary probe hybridization buffer and 100 μL was added to each of the wells. The probes then incubated for 15 minutes in the dark at room temperature. The secondary probes were then washed twice in 10% wash buffer. The cells then incubated in the 10% wash buffer for 5 minutes in the dark at room temperature. Finally, the cells were washed once with 2X SSC buffer and once with imaging buffer. Before imaging, the buffer was changed to anti-bleaching buffer (100 μL).

Images of cell autofluorescence were acquired with non-specific secondary probe staining to generate simulated spot images. These samples were prepared with the same seqFISH method described above, without the addition of primary probes.

1.1.6 Imaging conditions

The seqFISH samples were imaged with a Nikon Ti2 fluorescence microscope and controlled by Nikon Elements. Images were acquired with a Nikon SOLA SE II light source, a 100X oil objective, and a Photometrics Prime 95B CMOS camera.

1.2 Creation of spot training data

1.2.1 Image annotation

Our training dataset consisted of 1000 128x128 pixel images: 400 images generated as described above by performing seqFISH on cell culture samples, 400 previously published images generated with multiplexed error-robust FISH on tissue samples[26,32], and 200 previously published images generated with SunTag labeling of nascent proteins in cell culture samples[33]. All data were scaled so that the pixels had the same physical dimension of 0.11 nm prior to training. These images contained up to approximately 200 spots and were min-max normalized to a range of [0,1] prior to annotation. We annotated each image with four classical spot detection algorithms: maximum intensity filtering (`skimage.feature.peak_local_max`), difference of Gaussians (`skimage.feature.blob_dog`), Laplacian of Gaussian (`skimage.feature.blob_log`), and the Crocker-Grier centroid-finding algorithm (`trackpy.locate`). To accelerate image labeling, we created a Tkinter graphical user interface to tune the algorithm parameters (intensity threshold, minimum distance between spots, etc.) on a per-image basis. Additional details are provided in Supplementary Note 1.

For the spot detection algorithms that return spot locations with pixel-level resolution (`peak_local_max`, `blob_dog`, and `blob_log` in `skimage.feature`), the subpixel localization was determined by fitting a 2D isotropic Gaussian to the spot intensity[34]. A 10x10 pixel portion of the image surrounding the detected spot was cropped out and used for the Gaussian

fitting. A nonlinear least squares regression was performed, with initial parameters of Gaussian mean at the pixel center, an amplitude of 1 (the image's maximum value after min-max normalization) and a standard deviation of 0.5 pixels. The spot location was constrained to be in the middle 20% of the image, and the spot standard deviation was constrained to be between 0 and 3.

1.2.2 Clustering of spot annotations

Spots detected by the four classical algorithms were clustered into groups by proximity. For clarity, we refer to these classical algorithms as "annotators". Each group of detections was presumed to be derived from the same spot in the image. To perform this clustering, we first constructed a graph with each detected spot being a node. Two detections were connected by an edge if they were within 1.5 pixels of each other. We then took the connected components of this graph to be clusters. We screened the clusters to ensure that they contained at most one detection from each algorithm. If a cluster contained more than one detection from the same algorithm, the detection closest to the cluster centroid was retained, and all other detections were separated into new clusters.

From this graph, we derived the "detection information matrix," which identifies the annotators that contribute a detection to each cluster. This matrix has dimensions of $C \cdot A$, where C is the number of connected components or clusters in the graph and A is the number of annotators. The matrix has a value of 1 or 0 when a particular annotator does or does not have a detection in a particular cluster, respectively.

1.2.3 Creation of consensus annotations with expectation maximization

A generative model fit with the expectation-maximization (EM) algorithm[35] was used to estimate the probability that a cluster of detections corresponds to a true spot in the image. The detection information matrix, described above, was used as the input into the generative model along with an initial guess for the true positive rate (TPR) and false positive rate (FPR) of each algorithm and a prior probability of a spot being a true detection. The initial guesses for the TPR and FPR were 0.9 and 0.1, respectively. The prior probability of a spot being a true detection was defined as 0.5. Briefly, the EM algorithm consists of two steps - an expectation step and a maximization step. The expectation step yields an estimate for the probability that each detection cluster corresponds to a true detection. The maximization step yields an updated estimate for the TPR and FPR of each annotator. The expectation and maximization steps were performed iteratively 20 times, sufficient iterations for convergence to a local maximum of the likelihood. The resulting values were used as an estimate for the Bayesian probability of each cluster corresponding to a "true" spot. We provide further details of these steps in the Supplementary Information. Clusters with a probability above 0.9 were used as spots in the training dataset, and the spot location was taken to be the centroid of the detection cluster.

1.3 Spot detection deep learning model architecture

1.3.1 Preparation of coordinate annotations for training data

The coordinate spot locations were converted into two different types of images before being used for deep learning model training. The first image type is a classification image array in which pixels corresponding to spots and background are

one-hot encoded. The second image type is a regression image array, in which pixel values correspond to the distance to the nearest spot in the x- and y-direction.

1.3.2 Image preprocessing

We performed preprocessing of images prior to model training. Pixel intensities were clipped at the 0.01 and 99.9th percentiles and then min-max normalized so that all pixel values were scaled between 0 and 1.

1.3.3 Model architecture

Our deep learning model architecture is based on FeatureNets, a previously published backbone where the receptive field is an explicit hyperparameter[14]. We attach two prediction heads to this backbone: a classification head (to predict the probability a given pixel contains a spot) and a regression head (to predict the distance to the nearest spot with sub-pixel resolution). The receptive field of the network is an explicit hyperparameter that was set to a default value of 13 pixels.

All models were trained with stochastic gradient descent with Nesterov momentum. We used a learning rate of 0.01 and momentum of 0.9. We performed image augmentation during training to increase data diversity; augmentation operations included rotating (0°-180°), flipping, and scaling (0.8X-1.2X) input images. The labeled data were split into training and validation sets, with the training set consisting of 90% and the validation set consisting of 10% of the data. The test set for benchmarking model performance consisted of simulated spot images and held-out experimentally generated spot images.

1.3.4 Image postprocessing

We processed the classification and regression predictions to produce a list of spots. The local maxima of the classification prediction output was determined using maximum intensity filtering, with a default intensity threshold of 0.95. For most datasets, we found that the spot detection results do not vary widely with changes to this threshold value. In the pixels determined to be local maxima, the sub-pixel localization is determined by adding the value of the regression prediction in the x- and y-directions. These sub-pixel locations are returned as the output of the model.

1.4 Generation of multiplexed seqFISH dataset in cultured macrophages

1.4.1 Cell culture

THP-1 (TIB-202) cells were received from the American Type Culture Collection. The cells were cultured in Roswell Park Memorial Institute (RPMI) 1640 Medium (Gibco) supplemented with 2 mM L-glutamine (Gibco), 100 U/mL penicillin, 100 μg/mL streptomycin (Gibco or Caisson), and 10% fetal bovine serum (Omega Scientific or Thermo Fisher). To make the complete medium, 2-mercaptoethanol (BME) (Sigma-Aldrich #M6250) was added to a concentration of 0.05mM before every use. Cells were incubated at 37°C in a humidified 5% CO_2 atmosphere and were passaged to maintain a concentration of 0.3-1 x 10^6 cells/mL.

1.4.2 seqFISH sample preparation and imaging

The THP-1 monocyte cells were seeded on a fibronectin-functionalized glass slide (Corning #2980-246) at 80%-90% confluence contained by a rubber gasket (Grace Bio-Labs #JTR8R-2.5). To differentiate the THP-1 monocytes into macrophages, the cells were incubated with 10ng/mL phorbol 12-myristate 13-acetate (PMA) (Sigma-Aldrich #P8139) in RPMI with BME for 24 hours. The media was then replaced with fresh RPMI with BME and incubated for an additional 24 hours. Differentiation was confirmed visually based on changes in adherence and morphology.

The macrophages were dosed with 1ug/mL LPS (Sigma Aldrich #L4524) in RPMI with BME for 3 hours. Cells were rinsed with warm 1X PBS, fixed with fresh 4% formaldehyde (Thermo #28908) in 1X PBS for 10 minutes at room temperature, and then permeabilized with 70% ethanol overnight at -20°C. The primary probe library (Spatial Genomics) was added to the sample in a flow chamber provided by Spatial Genomics and incubated overnight at 37°C. The sample was washed several times with primary wash buffer (Spatial Genomics). The nuclei of the sample were stained with staining solution (Spatial Genomics).

The macrophage sample was imaged with the Spatial Genomics Gene Positioning System (GenePS). Image tiling and secondary probe staining were performed programmatically by this instrument.

1.4.3 Spatial Genomics image analysis

To generate a point of comparison for Polaris' output, we analyzed the seqFISH dataset with the Spatial Genomics software. The spot detection for this analysis was performed via manual parameter tuning. For each imaging round and channel, the threshold intensity used to detect spots was defined visually. The validity of this threshold value was confirmed across several randomly selected fields of view. The DAPI channel was used as the input for their supervised nuclear segmentation method; nuclear masks were dilated to create whole-cell masks.

1.5 Multiplex FISH analysis pipeline

1.5.1 Cell segmentation

For cell culture samples, cell segmentation was performed with nuclear and whole-cell segmentation applications from the Deepcell software library. For tissue samples, cell segmentation was performed with Mesmer[16]. The source code for these models is available at `https://github.com/vanvalenlab/deepcell-tf`; a persistent deployment is available at `https://deepcell.org`.

1.5.2 Gene identity assignment

Existing spot decoding methods for image-based spatial transcriptomics fall into two main categories: (1) pixel-wise decoding, which attempts to decode every pixel in the input image, and (2) spot-wise decoding, which attempts to detect spots before decoding them. Polaris' spot decoding method uses elements of both methods. For spot decoding, Polaris' pixel-wise spot probability output was used to determine which pixels to decode. The maximum intensity projection of the spot probability image was performed across all rounds and channels and the set of pixels to be decoded was determined by an intensity

threshold with a default value of 0.01. For each thresholded pixel, the array of spot probability values through the rounds and channels was used as the input for gene assignment.

Gene assignment was performed by fitting a generative model to the probability intensities at identified spot locations, a similar method to previously published work.[36] Our model consists of $2 \cdot R \cdot C$ relaxed Bernoulli distributions, where R is the number of imaging rounds in the experiment and C is the number of fluorescent channels in each round. The model for pixel intensities was based on a mixture of relaxed Bernoulli distributions by default, but Polaris also has two alternative distributions for modeling pixel intensities: Bernoulli and multivariate Gaussian. Therefore, the model consists of a "spot" distribution and a "no spot" distribution for each imaging round and channel. This model requires two inputs:

1. Spot probabilities at the pixel location across all imaging rounds and channels of the detected spots, as predicted by Polaris' deep learning model

2. A codebook defining the imaging rounds in which each gene in the sample is labeled, referred to as the gene's barcode

The distributions are fit to the pixel values of the detected spots with stochastic variational inference[25]. The codebook is used to constrain the logit function of the relaxed Bernoulli distribution, and the temperature is learned. We assume independence across channels and imaging rounds, but the distribution parameters are shared across all genes. An empty barcode of zeros is added to the input codebook, corresponding to a "background" - or false positive - assignment. These distributions define the probability of each barcode assignment for a set of pixel values. For each detected spot, the probability of each barcode assignment is calculated, based on the probability of sampling out of the "spot" or "no spot" distribution for each round and channel. The gene whose barcode has the highest probability is assigned to the spot. If the prediction probability does not exceed a threshold value, set to 0.95 by default, it is instead given an "unknown" assignment.

To find the coordinate locations of decoded genes, we create a mask with the pixels successfully decoded to a gene in the codebook and apply this mask to the maximum intensity-project spot intensity image. We perform peak finding with maximum intensity filtering to yield the coordinate location of each decoded gene.

1.5.3 Gene assignment rescue methods

After prediction, "background" and "unknown" assignments can be rescued to be assigned gene identities through two methods. The first method, which we refer to as "error rescue", compares the pixel values of the spot to each barcode in the codebook. If the Hamming distance of the pixel values to a barcode is less than or equal to 1, the assignment is updated. This method catches the rare cases in which the probabilistic decoder misses an assignment. The second method, which we refer to as "mixed rescue", catches the spots that contain two mixed barcodes. This situation occurs when two RNA molecules are in close proximity in the sample and the signal from their barcode labels is mixed. Mixed barcodes often lead to a low probability assignment, so all spots below a threshold probability, set to 0.95 by default, are checked for this case. For this method, the barcode values for the original assignment are subtracted and the update pixel values are compared to each barcode in the codebook. If the Hamming distance of the pixel values to a barcode is less than or equal to 1, the assignment is updated.

1.5.4 Background masking

Bright objects in the background of smFISH images can interfere with barcode assignment, so detected spots in these regions can be masked out. This step requires a background image without FISH staining to assess the initial fluorescence intensity in the sample. The background image is min-max normalized so that pixels are scaled between 0 and 1, and a mask is created with a threshold intensity, which defaults to 0.5. Any detected spots in the masked regions are excluded from downstream analysis.

2 Supplementary Information

2.1 Supplementary Note 1: Construction of training data for Polaris' spot detection model

2.1.1 Overview

To construct a training dataset for Polaris' spot detection model, we assembled a set of representative images from sequential fluorescence in situ hybridization (seqFISH)[10], multiplexed error-robust FISH (MERFISH)[26,32], and SunTag labeled imaging data[33]. We detected the location of these spots using the four classical spot detection algorithms described in the methods section. Here, we refer to each of these classical spot detection algorithms as "annotators". The parameters of these methods were manually fine-tuned on a per-image basis to optimize the spot detection results for each image. To generate subpixel resolution, we performed Gaussian fitting on the pixel values surrounding the spot location to estimate its subpixel location, as described in the methods section.

To generate consensus spot annotations, we fit a generative model to the detections generated by the four annotators (Supplementary Fig. S1). To determine which annotators detected each spot in an image, the detections from all annotators are first grouped by proximity into clusters. We presume that the detections closer than a defined threshold of 1.5 pixels have derived from the same spot in the image. Each cluster is allowed to contain only one detection from each annotator. If a cluster contains more than one detection from an annotator, the centroid of the cluster is determined, and the detection from that annotator closest to the centroid is included in the cluster. All other detections from that annotator are then separated into new clusters.

The generative model characterizes annotators with two parameters:

1. True positive rate (TPR), which is an annotator's probability of detecting a ground-truth true spot
2. False positive rate (FPR), which is an annotator's probability of detecting a ground-truth false spot

The generative model characterizes detected spots by the probability of corresponding to a "true" spot ($p(\text{TP})$). Said another way, $p(\text{TP})$ is the probability that a spot is a true spot given that it is detected. The probability that a detection corresponds to a "false" spot ($p(\text{FP})$) has a value of $1-p(\text{TP})$. The generative model is provided an initial guess for the TPR and FPR of each classical algorithm and a matrix, which we name the "detection information matrix". This matrix of annotation data, $x = x_{ic i=1,\ldots,n}$ consists of binary variables, x_{ic}, which are equal to 1 if annotator i detected cluster c, and 0 if not. We define z_c to be a binary variable indicating if cluster c is a true spot or not (1 if true, 0 if not). We assume that each annotator i produces Bernoulli distributed annotations. Therefore, for every cluster c and annotator i, the distribution of x_{ic} given the cluster assignment z_c and annotator characteristics $\theta_i(z_c)$ is a Bernoulli distribution:

$$P(x_{ic} = 1|z_c, \theta_i(z_c) = \theta_i(z_c)^{x_{ic}}(1-\theta_i(z_c))^{1-x_{ic}} \tag{1}$$

We assume the variables x_{ic} are independent to obtain that the probability to observe the data $x = x_{ic}$, given θ_i and z_c is

$$P(\{x_{ic}\}|\{z_c\},\{\theta_i\}) = \prod_i \prod_c \theta_i(z_c)^{x_{ic}}(1-\theta_i(z_c))^{1-x_{ic}} \tag{2}$$

We fit our generative model to our spot detection data using the expectation-maximization (EM) algorithm[35]. The EM algorithm iteratively performs three steps. First, the $p(\text{TP})$ and $p(\text{FP})$ values of each detection cluster are estimated given the TPR and FPR for each spot detection algorithm. Second, these probabilities are used to calculate the expectation value for the number of true positives ($\mathbb{E}(\text{TP})$), false positives ($\mathbb{E}(\text{FP})$), true negatives ($\mathbb{E}(\text{TN})$), and false negatives ($\mathbb{E}(\text{FN})$) from each algorithm. These two steps together comprise the expectation step. Third, these expectation values are used to update the estimated TPR and FPR for each algorithm. This step comprises the maximization step. These steps are performed iteratively until the estimate for the $p(\text{TP})$ value of the cluster converges. If the resulting $p(\text{TP})$ of a cluster exceeds a probability threshold of 0.9, the cluster is included in the final consensus set of spot annotations. The location of the consensus spot is taken to be at the centroid of the cluster, because each annotator is assumed to have some random localization error.

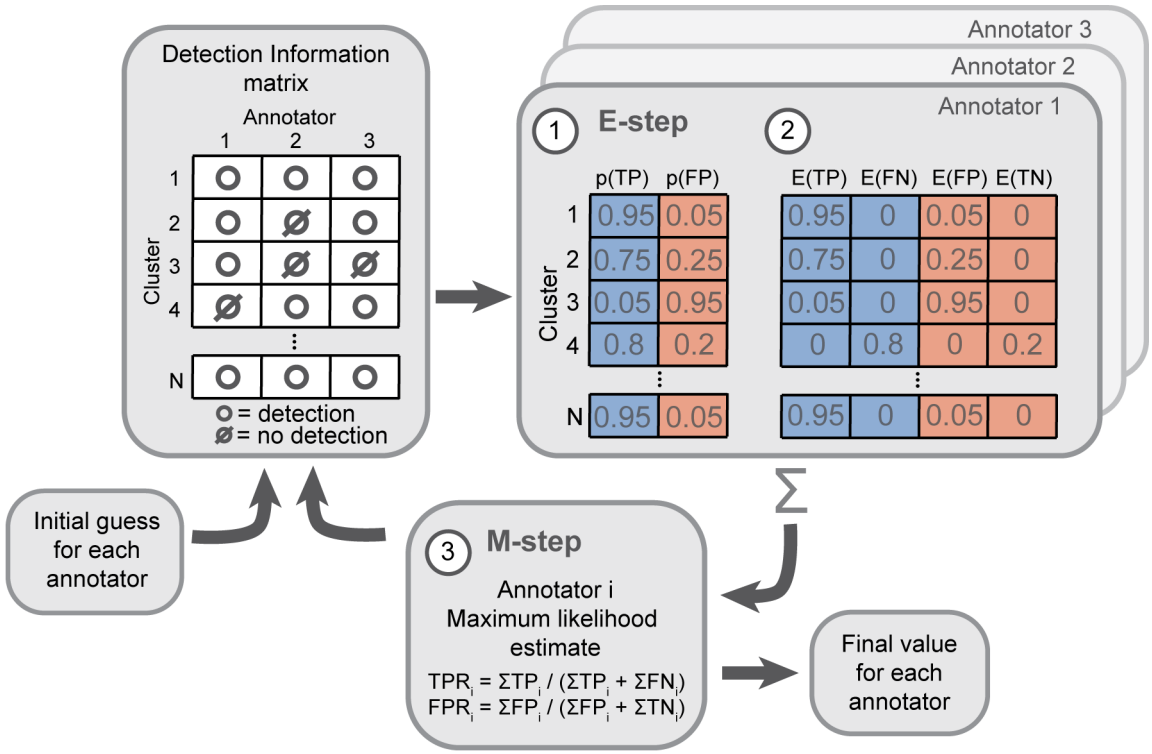

Figure S1: **Schematic diagram of the EM method for consensus spot annotation creation.** The method consists of three steps: (1) Bayesian estimation of $p(\text{TP})$ and $p(\text{FP})$. (2) Calculation of expectation values for the number of true positives (TPs), false negatives (FNs), false positives (FPs), and true negatives (TNs). (3) Update of the maximum likelihood estimate for the TPR and FPR values of each spot detection method.

2.1.2 Computation for the expectation step

We define the probability of a detection being a true or false detection with Bayes' theorem:

$$p(Z|\text{data},\theta) = \frac{p(\text{data}|Z,\theta)p(Z)}{p(\text{data}|\theta)} \tag{3}$$

where Z is the identity of the spot as a true or false detection, *data* indicates which annotators detected which clusters, and θ is the TPR or FPR of each method. The term $p(Z)$ is the prior probability of a spot being a true or false detection. We

can use the least informative value for the prior by setting $p(Z) = 1/2$, indicating an equal probability that a spot is a true or false detection. The term $p(\text{data}|\theta)$ can be expressed as:

$$p(\text{data}|\theta) = \sum_Z p(\text{data}|Z, \theta)p(Z) \tag{4}$$

Therefore, the probability $p(Z|\text{data}, \theta)$ can be written as:

$$\begin{aligned} p(Z|\text{data}, \theta) &= \frac{p(\text{data}|Z, \theta)p(Z)}{p(\text{data}|\theta)} && (5) \\ &= \frac{p(\text{data}|Z, \theta)p(Z)}{\sum_Z p(\text{data}|Z, \theta)p(Z)} && (6) \\ &= \frac{p(\text{data}|Z, \theta) \cdot \frac{1}{2}}{\sum_Z p(\text{data}|Z, \theta) \cdot \frac{1}{2}} && (7) \\ &= \frac{p(\text{data}|Z, \theta)}{\sum_Z p(\text{data}|Z, \theta)} && (8) \end{aligned}$$

This equation states that the likelihood of each possible label (e.g., true or false) is normalized by the total likelihood of both labels to calculate the probability that a cluster is a true or false detection.

We can use Bayes' theorem to calculate the probability of observing *data* given the TPR and FPR of each method. To offer a concrete example, consider the following three situations for a hypothetical set of three annotators. For a "true detection," the probability that all three annotators detect the spot is given by the product of the TPRs of each method:

$$p(\text{data}|Z, \theta) = \prod_{i=1}^{3} \text{TPR}_i \tag{9}$$

Alternatively, the probability that the first two annotators detect a ground-truth true spot while the third annotator (incorrectly) does not is given by the following equation:

$$p(\text{data}|Z, \theta) = \text{TPR}_1 \cdot \text{TPR}_2 \cdot (1 - \text{TPR}_3) \tag{10}$$

Similarly, the probability that the first two annotators incorrectly detect a ground-truth false spot while the last annotator (correctly) does not is given by the following equation:

$$p(\text{data}|Z, \theta) = \text{FPR}_1 \cdot \text{FPR}_2 \cdot (1 - \text{FPR}_3) \tag{11}$$

We can then calculate $\mathbb{E}(\text{TP})$, $\mathbb{E}(\text{FN})$, $\mathbb{E}(\text{FP})$, and $\mathbb{E}(\text{TN})$ for each annotator. Two scenarios can arise in calculating these values.

1. If an annotator detects a spot in a particular cluster, $\mathbb{E}(\text{TP})$ for that method is equal to $p(\text{TP})$ for that cluster, and $\mathbb{E}(\text{FP})$ for that annotator is equal to $p(\text{FP})$ for that cluster. $\mathbb{E}(\text{TN})$ and $\mathbb{E}(\text{FN})$ are set to zero.

2. If an annotator does not detect a spot in a particular cluster, $\mathbb{E}(\text{TN})$ for that annotator is equal to $p(\text{FP})$ for that cluster, and $\mathbb{E}(\text{FN})$ for that annotator is equal to $p(\text{TP})$ for that cluster. $\mathbb{E}(\text{TP})$ and $\mathbb{E}(\text{FP})$ are set to zero.

2.1.3 Computation for the maximization step

We sum $\mathbb{E}(\text{TP})$, $\mathbb{E}(\text{FN})$, $\mathbb{E}(\text{FP})$, and $\mathbb{E}(\text{TN})$ across all clusters to calculate an updated maximum likelihood estimate for TPR_i and FPR_i for method i with equations of the following form:

$$\text{TPR}_i = \frac{\sum_i \mathbb{E}(\text{TP}_i)}{\sum_i \mathbb{E}(\text{TP}_i) + \sum_i \mathbb{E}(\text{FN}_i)} \tag{12}$$

$$\text{FPR}_i = \frac{\sum_i \mathbb{E}(\text{FP}_i)}{\sum_i \mathbb{E}(\text{FP}_i) + \sum_i \mathbb{E}(\text{TN}_i)} \tag{13}$$

2.1.4 Evaluation of generative model performance for the creation of training data

The absence of ground-truth annotations for experimental data presents challenges for benchmarking. To evaluate the accuracy of our approach, we used simulated data. Concretely, we performed simulated experiments to demonstrate the accuracy of our generative model output by simulating spot detection methods with varying TPR and FPR values and simulating spot identities as true or false spots. These simulated experiments do not involve simulated images or application of classical spot detection methods, but instead involve simulating annotations in the form of a detection information matrix. These experiments demonstrate that the EM algorithm can be applied to accurately find consensus among noisy detections. First, we simulated a set of spot identities. The ratio of ground-truth true and false spots was determined by a pre-defined prior probability of true spots. We then used the defined TPR and FPR values to simulate detections. The probability that true spots are detected by the simulated spot detection methods is defined by the simulated TPR, and the probability that false spots are detected is defined by the simulated FPR. As with detections from experimental images, the simulated detections are stored in a detection information matrix.

We demonstrate that the generative model outputs TPR and FPR estimates for classical methods that are close to their actual values. Using the simulated detection information matrix for simulated spot detection methods with different TPR and FPR values, we determined the error distribution for the TPR and FPR estimates of the simulated spot detection methods based on 200 trials of EM parameter estimation. For the simulated spot detection methods, the TPR and FPR values were sampled from uniform distributions centered at 0.9 and 0.1, respectively (Supplementary Fig. S2a). Using four simulated spot detection methods, we investigated the relationship between the accuracy of the generative model output and the number of spots in the input data set (Supplementary Fig. S2b). The fraction of correctly labeled detections increases with the number of spot detections in the dataset, with the accuracy exceeding 95% over the typical range of dataset sizes for training Polaris' deep learning model. Additionally, using a simulated dataset size of 1000 spots, we found that the accuracy of the generative model output improves with the number of spot detection methods used to create the consensus spot locations, approaching 100% (Supplementary Fig. S2c). In these experiments, we ran 50 independent simulations per condition in which TPRs and FPRs were sampled from uniform distributions with a range of 0.2, centered at 0.9 and 0.1, respectively.

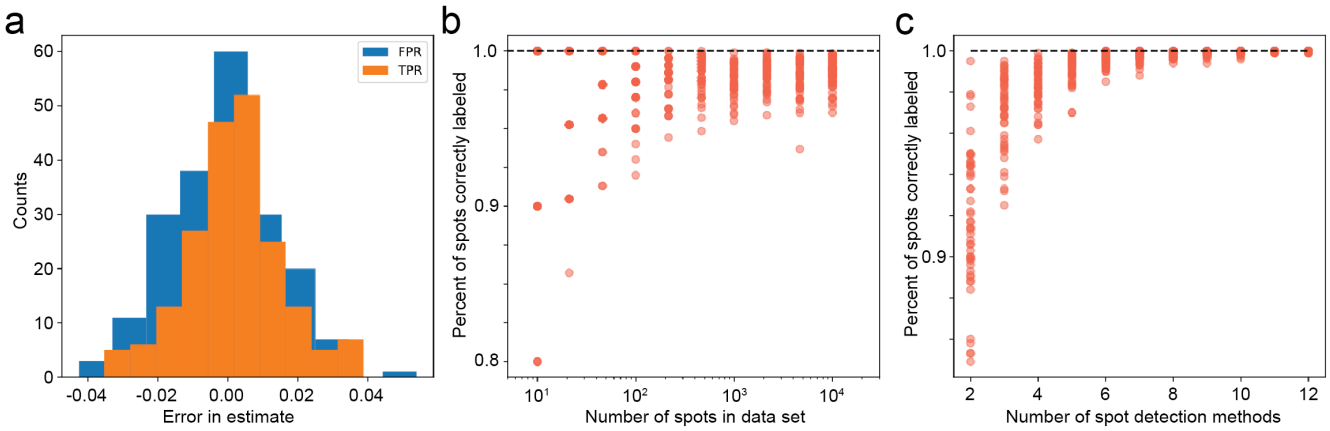

Figure S2: **Benchmarking consensus annotation output of the generative model.** (a) Error distribution for EM estimates of TPR and FPR values for 100 trials with three simulated classical methods. (b) Fraction of simulated detections correctly classified with increasing dataset size (number of spots in the dataset). (c) Fraction of simulated detections correctly classified as a true or false detection by EM for an increasing number of classical spot detection methods used in the EM method.

2.2 Supplementary Note 2: Custom loss function for Polaris' deep learning model for spot detection

We trained the network using a custom loss function composed of a classification loss and a regression loss, which considers the outputs of both of the model's prediction heads. The loss function has the following form:

$$L(y, \hat{y}) = L_{\text{cla}}(C, \hat{C}) + L_{\text{reg}}(R, \hat{R}) \tag{14}$$

where C is the classification head output, R is the regression head output, and $y = (C, R)$. The classification loss is the weighted cross-entropy with inverse class frequency-based weights. The regression loss is given by:

$$L_{\text{reg}}(R, \hat{R}) = \frac{1}{|G_d|} \sum_{i \in G_d} \ell((dy_i, dx_i), (\hat{dy}_i, \hat{dx}_i)) \tag{15}$$

where $R_i = (dy_i, dx_i)$ (i denotes a single pixel), $G_d = \{\text{pixels } i = (i_y, i_x) |$ a spot-containing pixel j exists with $L_\infty(i, j) = \max_{k \in x,y} |i_k - j_k| \leq d\}$, and ℓ is the smooth L_1 function. dx_i is the x-coordinate of the position difference between the nearest spot to pixel i, and pixel i's center. Similarly, dy_i is the y-coordinate of this position difference. d is a configurable parameter which determines the threshold distance from the nearest spot under which the estimated nearest spot's position for that pixel is taken into account in the loss function.

2.3 Supplementary Note 3: Generalization of Polaris' spot detection model to a variety of spot images

Visually, we found that our deep learning model for spot detection trained with consensus spot locations generalizes well to images generated by several different assays (Supplementary Fig. S3). These images have a variety of spot and background characteristics and include images created with assays not represented in our training dataset. To generate the set of predicted spot locations, Polaris utilizes its pixel-wise spot probability and regression images. This result demonstrates that Polaris is a universal deep learning model trained with weak supervision that can generalize to perform spot detection for a wide variety of single-molecule imaging data.

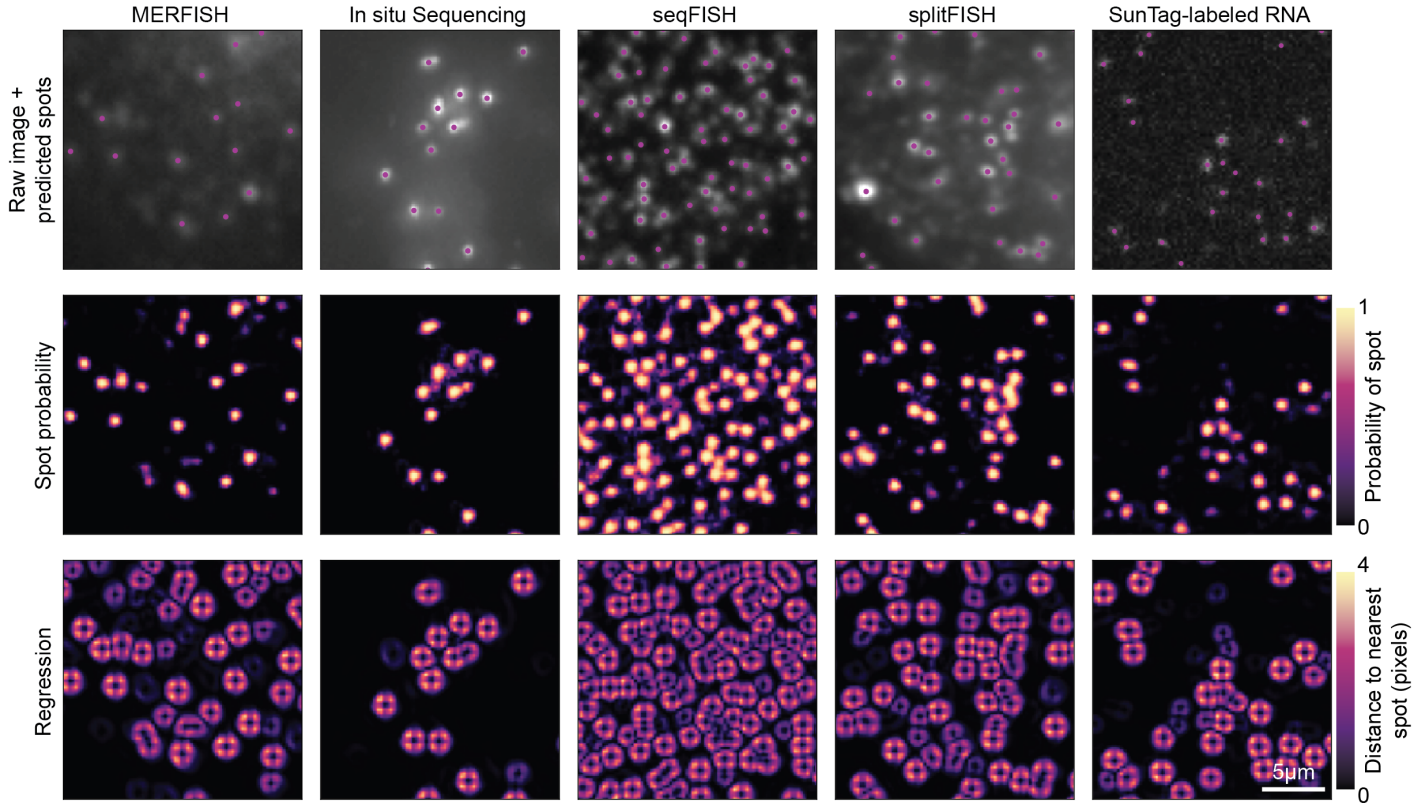

Figure S3: **Polaris' spot detection model generalizes to spot images generated with a variety of single-molecule FISH assays.** The spot probability prediction images encode the pixel-wise spot probability, and the regression prediction images. The regression image is the sum of the square of the subpixel distances to the nearest spot in the x- and y-dimensions. Pixels beyond a threshold value are set to zero. These outputs are used together to generate a set of predicted spot locations with subpixel resolution, plotted over the raw image.

2.4 Supplementary Note 4: Mutual nearest-neighbor method for matching sets of spots

To more quantitatively benchmark the performance of our deep learning models, we need a method for comparing sets of coordinate spot locations. Our method finds sets of mutual nearest neighbors to compare sets of ground-truth and predicted spot locations. Locations that can be matched between ground-truth spots and detected spots are considered true positive detections. Detected spots that cannot be matched to ground-truth spots are false positive detections, and ground-truth spots that cannot be matched to detected spots are false negative detections. These values can be used to calculate metrics such as precision and recall to quantify the performance of a spot detection method.

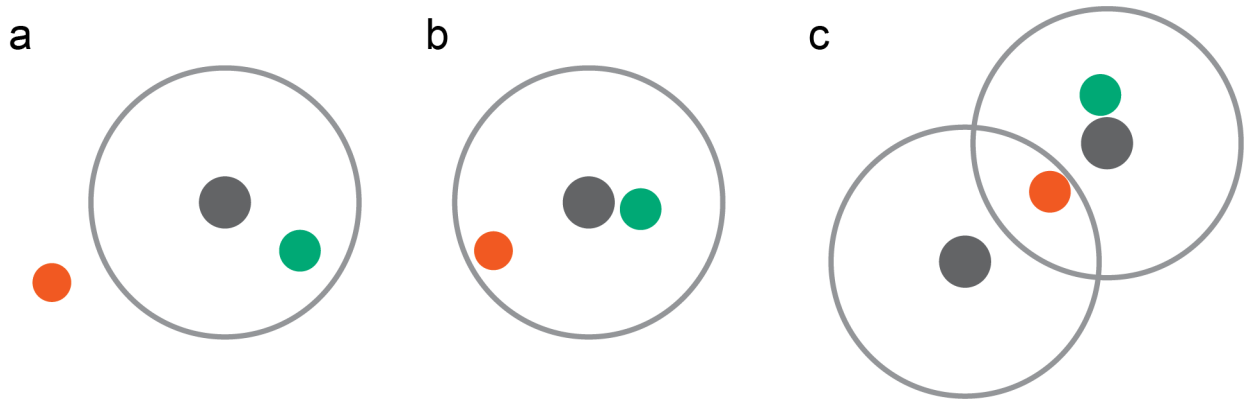

Figure S4: **Example cases handled by a mutual nearest-neighbor matching algorithm.** (a) Example with spots inside and outside the threshold distance to a ground-truth spot. Ground truth spots and their threshold distance are shown in grey. True positive detections are shown in green and false positive detections are shown in orange. (b) Example with two spots inside the threshold distance to a ground-truth spot. (c) Example with two spots within the threshold distance of two ground-truth spots.

To be considered a true detection, a detection must be within some threshold distance of a ground-truth detection. All spots outside this threshold distance are considered false detections (Supplementary Fig. S4a). If more than one detection is within the threshold distance of a ground-truth spot, the detection that is the closest to the ground-truth spot is considered a true detection, and all others are considered false detections (Supplementary Fig. S4b). If more than one detection is within the threshold distance for more than one ground-truth spot, edge cases may arise. For a detection to be considered a true detection, that detection and the corresponding ground-truth spot must be mutual nearest neighbors. Therefore, if a detection is within the threshold distance for two ground-truth spots, it is only paired with a ground-truth spot if they are each others' mutual nearest neighbors. Otherwise, the detection is considered a false detection, even if the detection is within the threshold distance of a ground-truth spot (Supplementary Fig. S4c).

2.5 Supplementary Note 5: Inter-algorithm agreement of spot detection results

We sought to explore the degree to which the detection sets from classical spot detection methods and our deep learning model agree, denoted as inter-algorithm agreement. For this analysis, we utilized held-out images. We used our mutual nearest-neighbor method to match the sets of detected spots from different spot detection algorithms (Supplementary Fig. S5). The inter-algorithm agreement between the different classical methods, quantified by the F1 score, varies more widely than the agreement between the deep learning model and each classical method. The lack of agreement between the detection sets from each classical spot detection method demonstrates the need for a method that can provide a consensus between detection sets. Furthermore, the low variance in the agreement between the deep learning detection sets and the detection sets of the classical methods suggests that the model yields detection sets that resemble a consensus between outputs of the classical methods. This analysis demonstrates Polaris' learning of consensus labels generalizes to images held-out from the training data set.

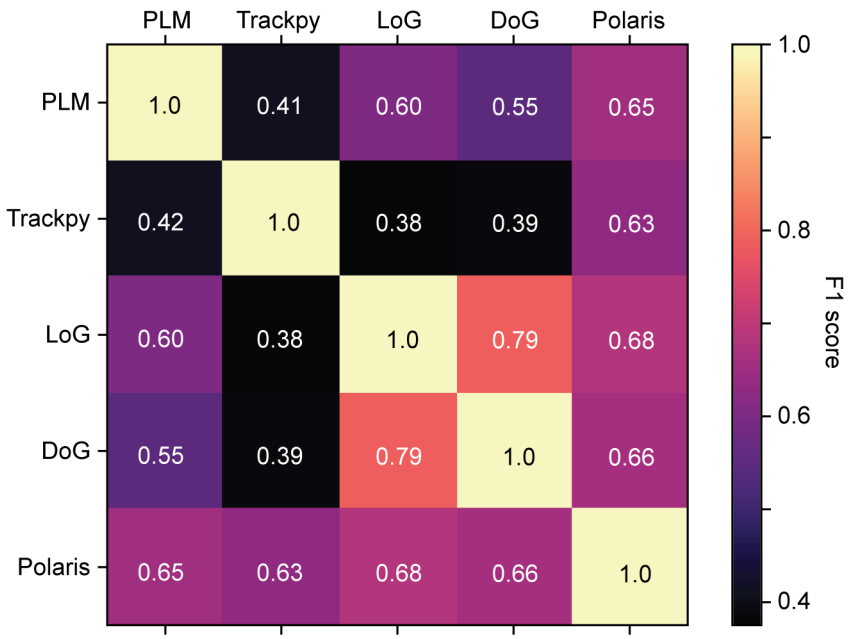

Figure S5: **Quantification of agreement between Polaris' deep learning model and different classical spot detection methods.** The benchmarked methods include maximum-intensity filtering (PLM), the Crocker-Grier centroid-finding algorithm (trackpy), Laplacian of Gaussian (LoG), difference of Gaussians (DoG), and Polaris (DL model).

2.6 Supplementary Note 6: Benchmarking the receptive field of Polaris' spot detection model

To quantify the effect of the receptive field parameter of Polaris' spot detection model on its performance, we followed prior work[37,38] and used simulated spot images, which is a common practice in the field. We created simulated images by adding Gaussian spots at random locations to cellular autofluorescence images. This approach to benchmarking has the advantage of having an unambiguous ground truth. We find that the receptive field parameter has a modest effect on the model's performance. The model's precision decreases with increasing receptive field, a trend that was also reflected in the model's F1 score across different values for the receptive field (Supplementary Fig. S6a,c). Intermediate values for the receptive field had the lowest recall, but the best values for the validation loss during training (Supplementary Fig. S6b,d). Because the effect of the receptive field parameter on the average value for all metrics is relatively modest, the intermediate value of 13 was selected as the default value for the spot detection model.

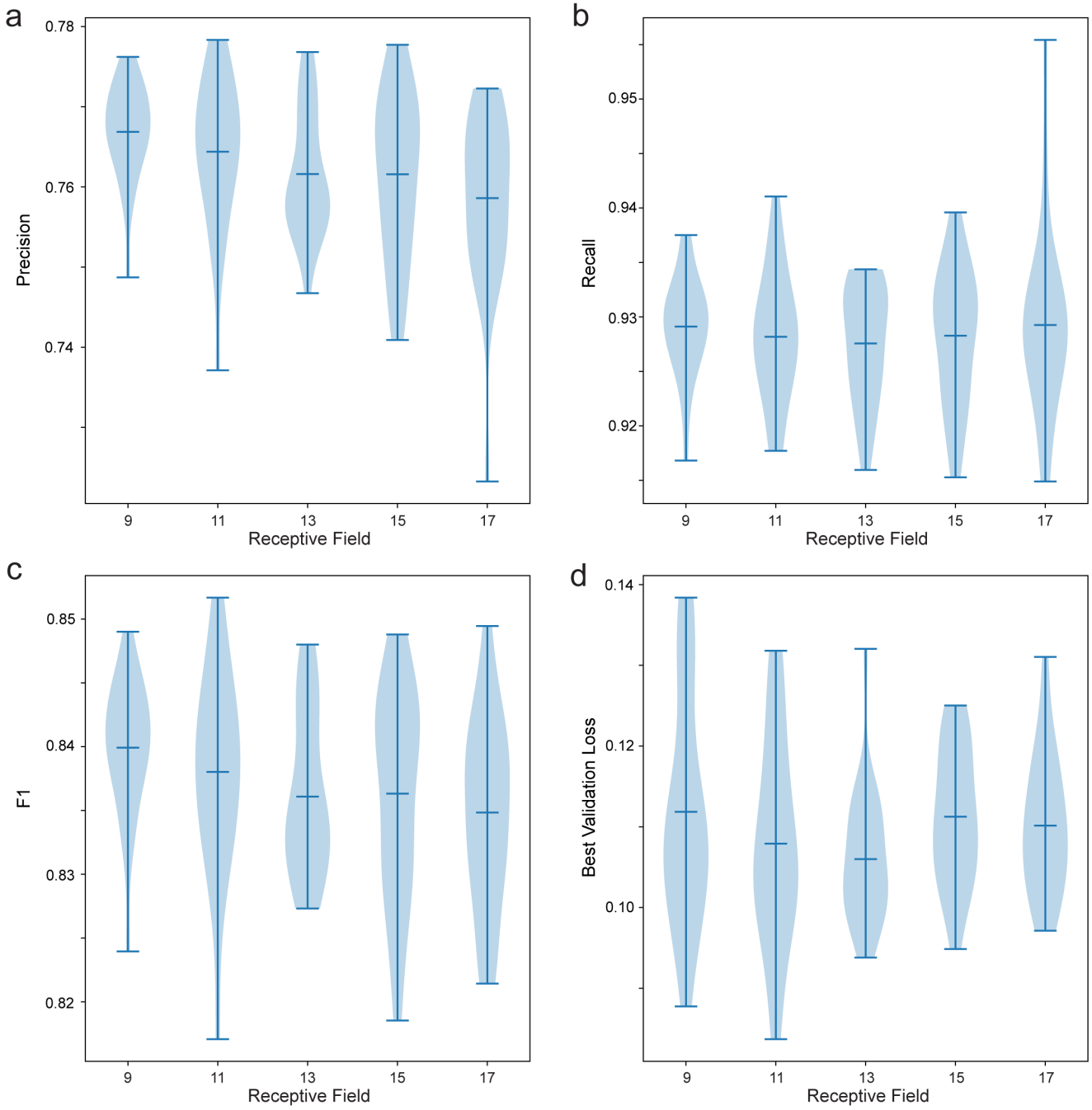

Figure S6: **Benchmarking the receptive field parameter of Polaris' spot detection model.** (a-d) Violin plot quantifying the performance metrics ((a) precision, (b) recall, (c) F1, (d) best validation loss during training) for models trained with different values for receptive field of Polaris' spot detection model. n=24 trained models per receptive field condition.

2.7 Supplementary Note 7: Benchmarking Polaris' spot detection model on simulated images with ranging spot characteristics

To quantitatively benchmark the performance of our trained models, we used simulate spot images, similarly to the previous figure. Because we control the image generation, we can explore model performance as a function of image difficulty by tuning parameters such as the spot density and signal-to-noise ratio. The location and number of spots in an image were sampled from uniform distributions. The intensity and width of the simulated Gaussian were sampled from normal distributions reflecting a spot distribution that is characteristic of experimental images.

We used the above-mentioned mutual nearest-neighbor method to link predicted spots with ground-truth spots and used the F1 score as a measure of performance. We observed that our deep learning model trained with consensus annotations outperformed models trained with either single classical algorithm annotations or simulated images (Supplementary Fig. S7).

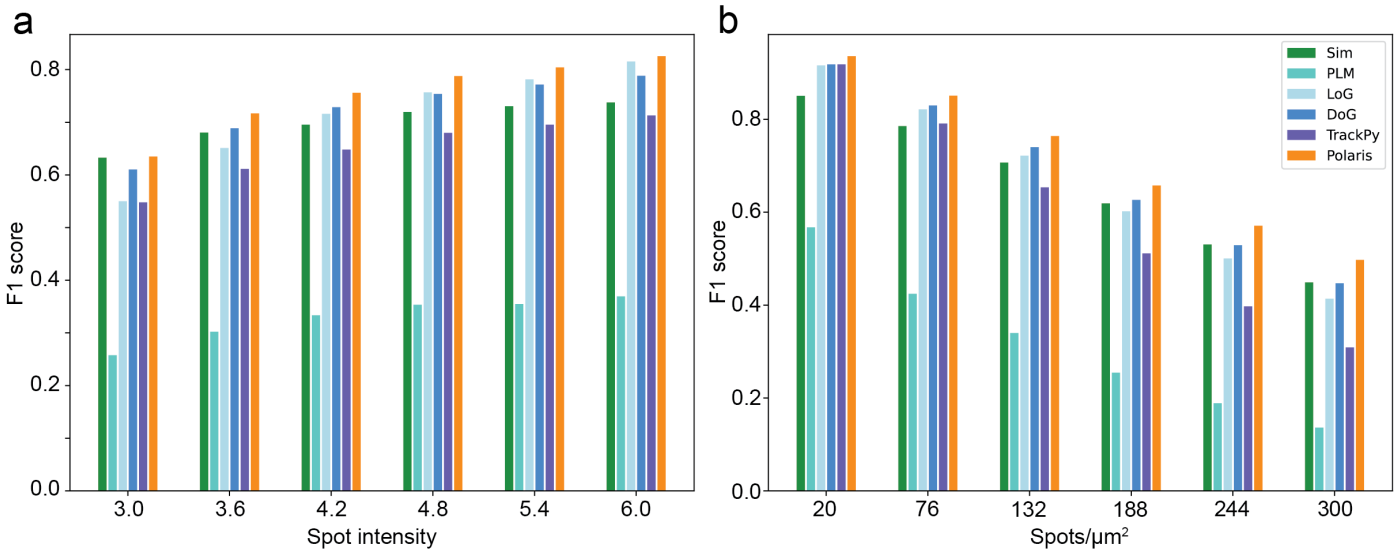

Figure S7: **Benchmarking model performance on simulated spot images with a range of spot intensities and densities.** (a) Performance quantification for models with various training datasets applied to images with varying spot intensity. (b) Performance quantification for models with various training datasets applied to images with varying spot density. Units of spot intensity are arbitrary.

Because the model trained with the consensus annotations outperforms all models trained with annotations from a single classical algorithm, we conclude that the consensus annotations more accurately capture the ground-truth locations of spots in the training images. The training dataset of simulated images was created in an identical fashion to the test set used to benchmark the models. Therefore, because the model trained with the consensus annotations outperforms the model trained with simulated images, there is some "intangible" value to training models with experimentally generated images.

2.8 Supplementary Note 8: Robustness of Polaris' barcode decoding method to dropout

To quantitatively benchmark the performance of our barcode assignment method against other methods, we used simulated data. We generated simulated barcode pixel values by sampling from distributions of spot and background pixel values. Similar to our benchmarking analysis of spot detection models, this method benefits from an unambiguous ground truth, unlike experimental data. Furthermore, we can control the dropout rate in the simulated barcodes. A dropout event occurs when a transcript is not detected in a round due to labeling failure, image quality, or failure of our spot detection pipeline. For example, we observed that Polaris' spot detector would fail to detect some low-intensity spots in images with high variability in spot intensity. This specific failure mode may be due to Polaris' image normalization method, which can further reduce a spot's intensity if it is near a very bright spot. Regardless of origin, the presence of dropout imposes a robustness constraint on our barcode decoding method. Our expectation is that a decoding method robust to dropout would better tolerate labeling and spot detection model failures.

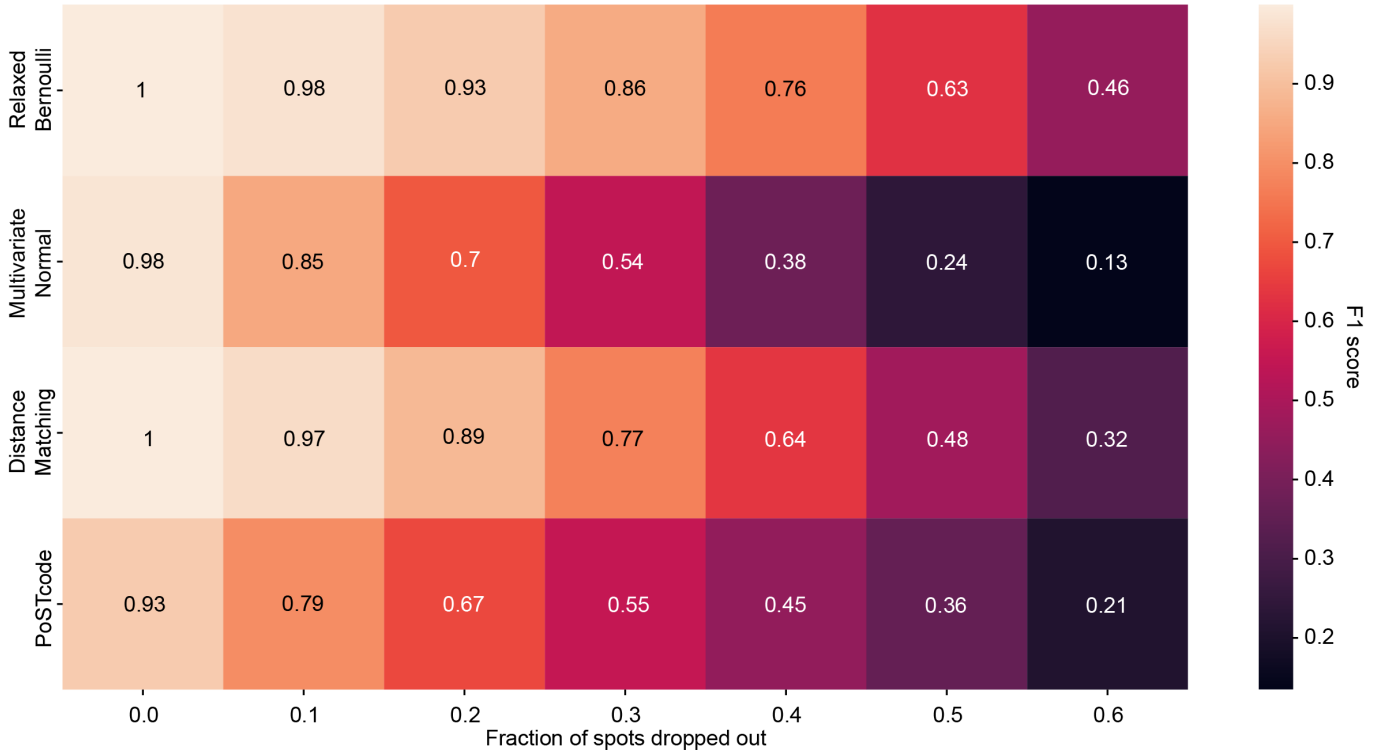

Figure S8: **Benchmarking of robustness of gene decoding methods to dropout.**

This benchmarking experiment demonstrates that our relaxed Bernoulli method is more robust to dropout than any other benchmarked method (Supplementary Fig. S8). The other benchmarked methods include an analogous graphical model of multivariate normal distributions, barcode matching by Hamming distance, and PoSTcode [36]. The relaxed Bernoulli graphical model, the multivariate normal graphical model, and the distance matching method take input pixel values sampled from a distribution simulating spot probability values output by Polaris. Alternatively, PoSTcode takes input pixel values from a distribution simulating pixel values of the raw imaging data, consistent with its original methodology. Our method for simulating barcode values does not consider correlations in pixel values between images acquired in the same fluorescence channel or imaging round. PoSTcode relies on this nuanced characteristic of experimental data; thus, its performance on the simulated data used in this benchmarking analysis may have been negatively impacted.

2.9 Supplementary Note 9: Benchmarking Polaris' performance on various multiplexed FISH image sets

We benchmarked Polaris' performance on two previously published MERFISH datasets[26,27] and one seqFISH dataset. We demonstrate that Polaris accurately detects and decodes FISH spots in cell culture and tissue samples (Supplementary Fig. S9a-b). Visual inspection of Polaris' output illustrates the value of multiplexed FISH methods, which quantify gene expression while retaining the sub-cellular location of each transcript. In the kidney tissue sample, anatomical structures create tissue-level organization of gene expression (Supplementary Fig. S9a). These structures would be lost during the tissue homogenization required for sequencing-based assays.

The cultured macrophage sample was created by stimulating immortalized macrophages with LPS. In response, the macrophages demonstrate cell-to-cell and sub-cellular heterogeneity in inflammatory gene expression (Supplementary Fig. S9b). While single-cell sequencing technologies would be able to capture the cell-to-cell heterogeneity in gene expression, the sub-cellular spatial distribution of transcripts would be lost. Furthermore, image-based spatial transcriptomics assays can be more easily paired with other imaging assays, such as dynamic observation of a live cell reporter or spatial proteomics. These paired measurements allow multiple nodes of an individual cell's signaling state to be observed.

To benchmark Polaris' performance on MERFISH data in tissue samples, we compare its output to counts measured with RNA-seq. We find that for the mouse ileum MERFISH dataset[26], Polaris' mean gene expression counts per cell correlate well with RNA-seq counts (r=0.582), which is similar to the correlation between the counts from the original analysis of this data with RNA-seq counts (r=0.683) (Supplementary Fig. S10a). The counts from these two analyses are also highly correlated (r=0.871) (Supplementary Fig. S10b). For the mouse kidney MERFISH dataset[27], we find similar results; Polaris and the original analysis yield mean gene expression counts per cell that are similarly correlated with RNA-seq counts (r=0.552 and r=0.565, respectively) (Supplementary Fig. S10c). The counts from the two MERFISH analyses are also highly correlated (r=0.954) (Supplementary Fig. S10d). These comparisons demonstrate that Polaris can output a similar result as the original analysis pipelines but without fine-tuning of the underlying spot detection method.

Furthermore, we demonstrate Polaris' performance on seqFISH data in a cell culture sample. We compare its output mean gene expression counts per cell to counts obtained with the Spatial Genomics analysis software and found the two analyses yielded counts that were similarly correlated with RNA-seq counts (r=0.802 and r=0.694, respectively) (Supplementary Fig. S10e). As with the MERFISH analyses, the counts from the two seqFISH analysis pipelines were highly correlated (r=0.909) (Supplementary Fig. S10f). This analysis demonstrates that Polaris generalizes between different FISH assays and sample types without manual parameter tuning. However, the spatial transcriptomics field acknowledges that FISH counts are prone to overdispersion, limiting the utility of comparisons made with a linear regression model[27,39].

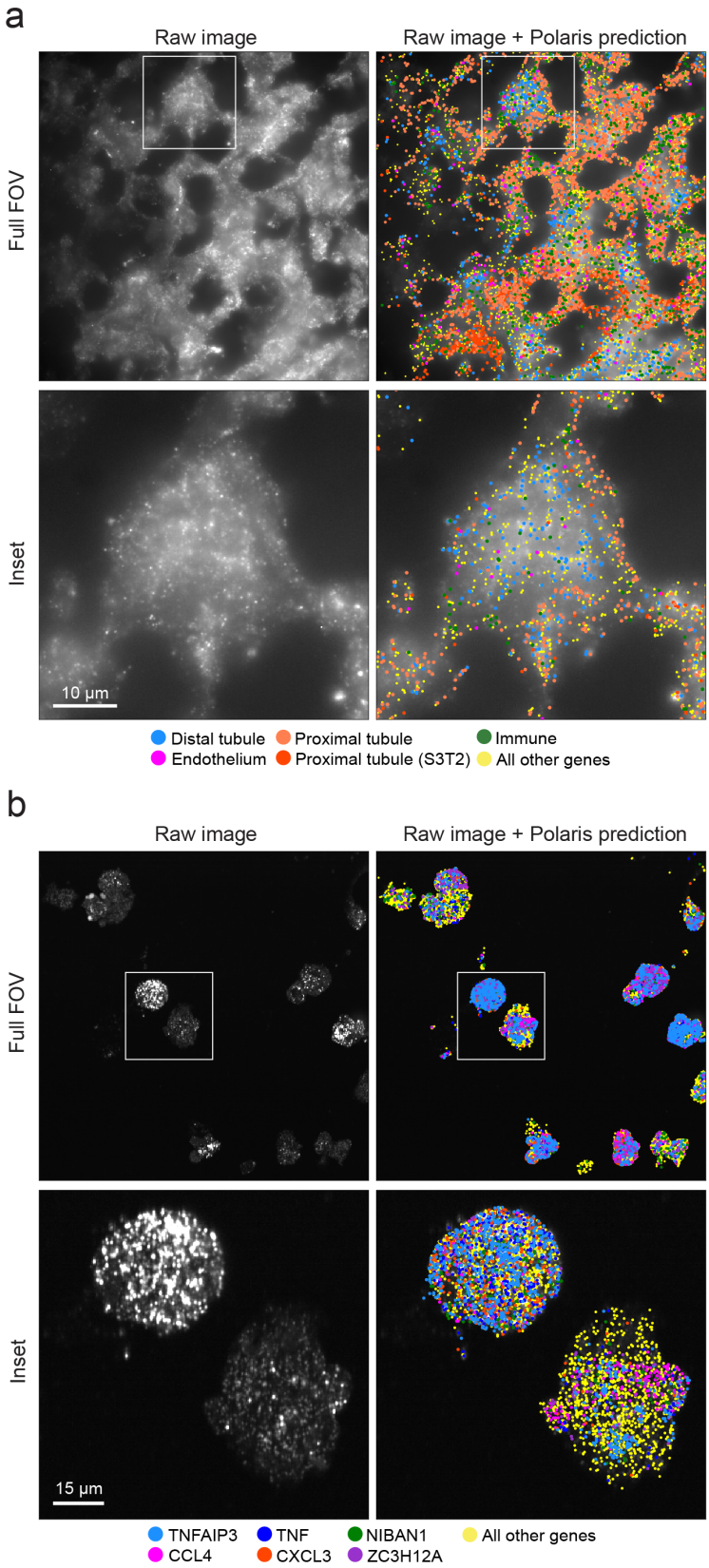

Figure S9: **Demonstration of Polaris' performance on a MERFISH and seqFISH data.** (a) Example Polaris prediction for a MERFISH experiment in a mouse kidney tissue sample.[27] (b) Example Polaris prediction for a seqFISH experiment in a macrophage cell culture sample. The spot colors of the Polaris prediction in (a,b) denote the predicted gene identities. The inset image location is defined by the white box in the full field of view (FOV).

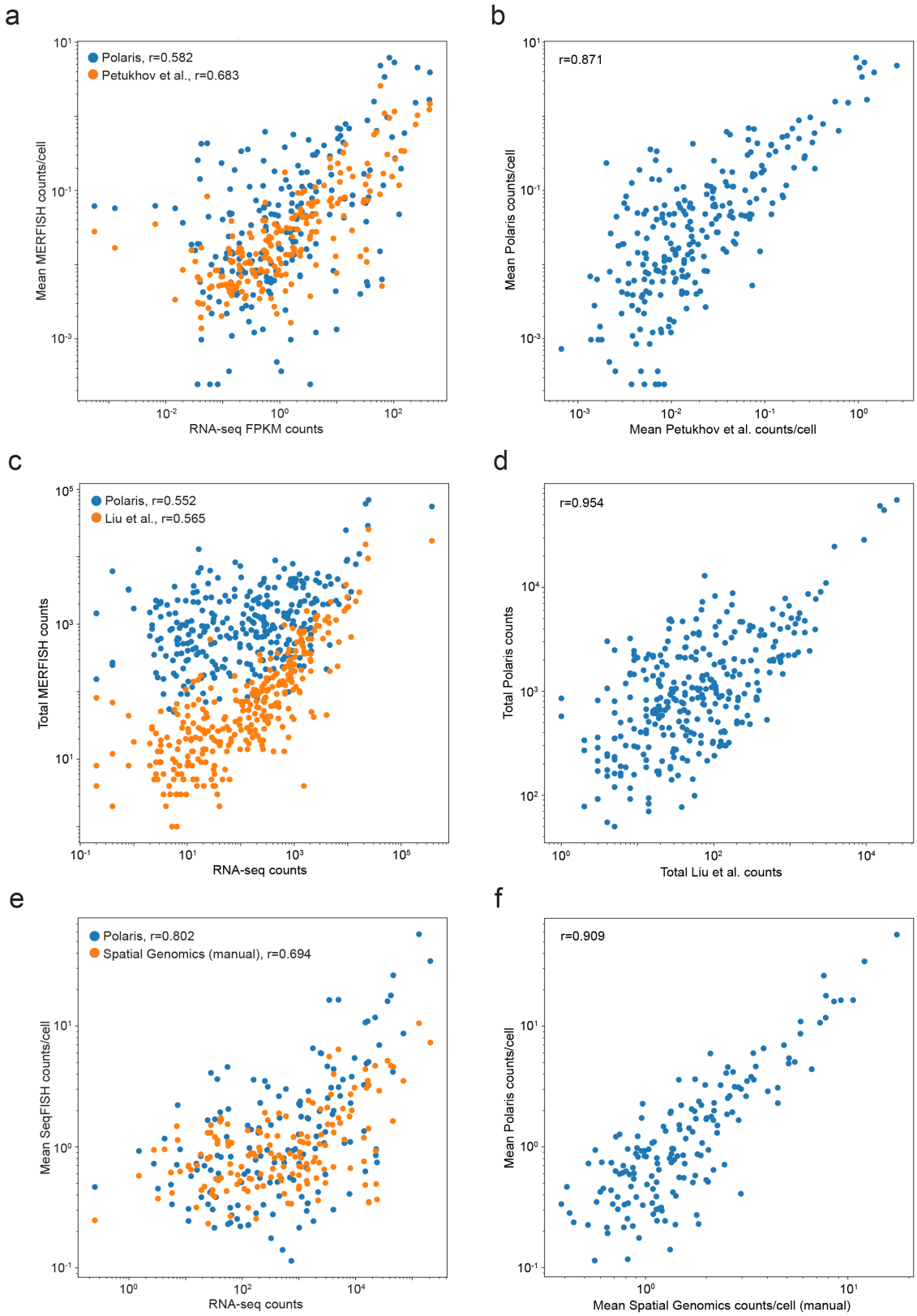

Figure S10: **Correlation of Polaris' quantification of MERFISH data with other quantification methods.** (a,c) Scatter plot plotted in logspace, comparing genes expression counts quantified with MERFISH with counts measured with RNA-seq. (b,d) Scatter plot plotted in logspace, comparing previously published MERFISH gene expression counts with counts quantified with Polaris. (e) Scatter plot plotted in logspace, comparing mean genes expression counts per cell quantified with seqFISH with counts measured with RNA-seq. (f) Scatter plot plotted in logspace, comparing mean gene counts per cell obtained by manual analysis of seqFISH data with gene counts quantified with Polaris.

2.10 Supplementary Note 10: Demonstration of Polaris' performance decoding a ISS barcode library

We benchmarked Polaris' combinatorial ISS barcode assignment performance on a previously published image set [28] in HeLa cells created with pooled optical genetic screen assay, which uses ISS to encode a genetic perturbation (Supplementary Fig. S11a). We find that the barcode counts decoded by Polaris are highly correlated with the counts quantified in the original analysis (r=0.956) (Supplementary Fig. S11b). For all barcodes, Polaris consistently yields higher counts of decoded spots, demonstrating higher sensitivity than the originally published image analysis pipeline.

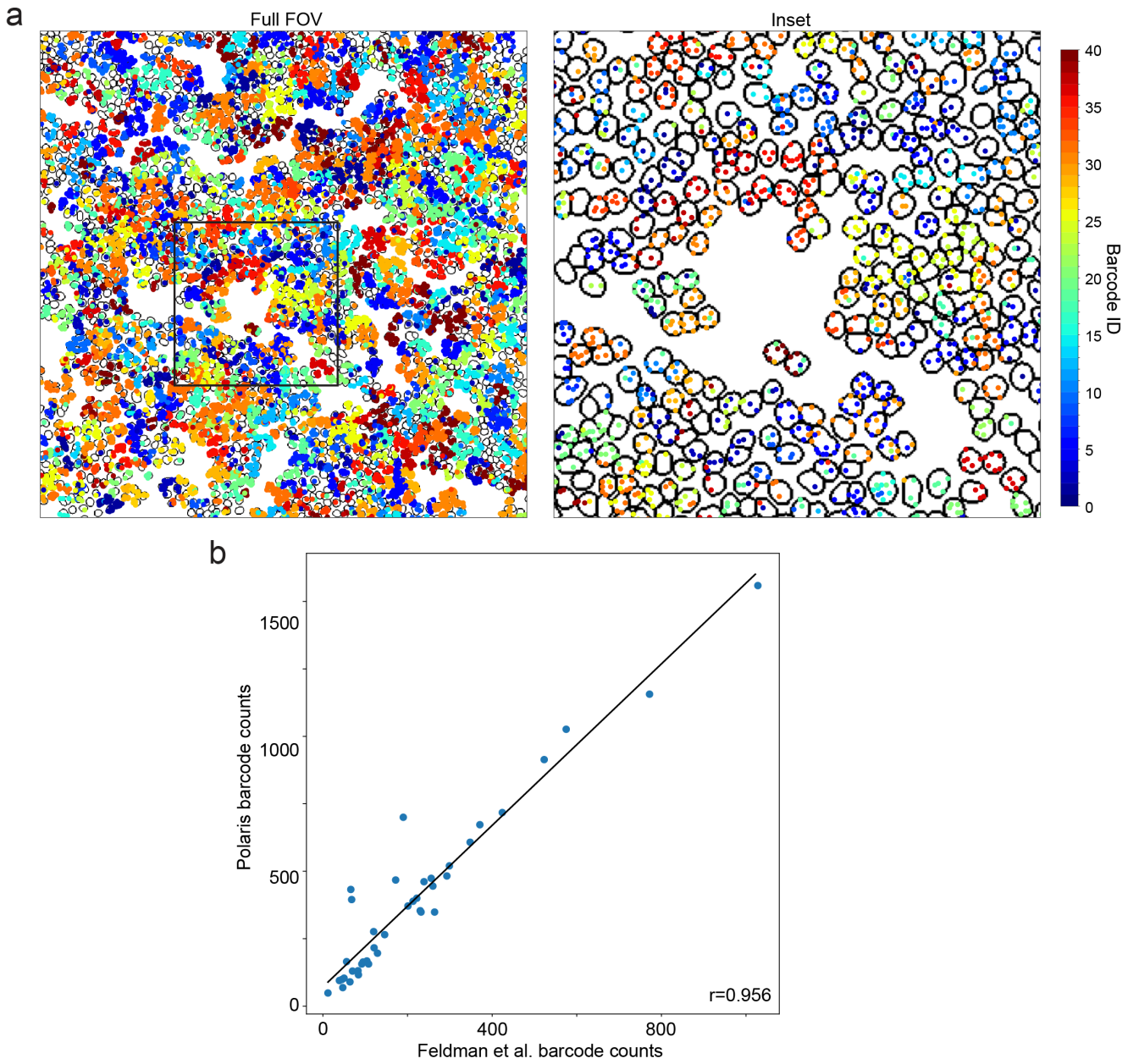

Figure S11: **Demonstration of Polaris' performance on an ISS dataset in HeLa cells.** (a) Example Polaris prediction for the ISS sample. The spot colors correspond with barcode identities. The inset location is defined by the black box in the full field of view (FOV). (b) Scatter plot correlating total counts for each barcode decoded by the original published analysis with counts quantified by Polaris (r=0.956).

References

[1] Asp, M.; Bergenstråhle, J.; Lundeberg, J. Spatially Resolved Transcriptomes—Next Generation Tools for Tissue Exploration. *BioEssays* **2020**, *42*, 1–16.

[2] Zhang, L.; Chen, D.; Song, D.; Liu, X.; Zhang, Y.; Xu, X.; Wang, X. Clinical and translational values of spatial transcriptomics. *Signal Transduction and Targeted Therapy* **2022**, *7*.

[3] Ståhl, P. L. et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **2016**, *353*, 78–82.

[4] Vickovic, S. et al. High-definition spatial transcriptomics for in situ tissue profiling. *Nature Methods* **2019**, *16*, 987–990.

[5] Rodriques, S. G.; Stickels, R. R.; Goeva, A.; Martin, C. A.; Murray, E.; Vanderburg, C. R.; Welch, J.; Chen, L. M.; Chen, F.; Macosko, E. Z. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **2019**, *363*, 1463–1467.

[6] Stickels, R. R.; Murray, E.; Kumar, P.; Li, J.; Marshall, J. L.; Bella, D. J. D.; Arlotta, P.; Macosko, E. Z.; Chen, F. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nature Biotechnology* **2020**, *39*, 313–319.

[7] Ke, R.; Mignardi, M.; Pacureanu, A.; Svedlund, J.; Botling, J.; Wählby, C.; Nilsson, M. In situ sequencing for RNA analysis in preserved tissue and cells. *Nature Methods* **2013**, *10*, 857–860.

[8] Chen, K. H.; Boettiger, A. N.; Moffitt, J. R.; Wang, S.; Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **2015**, *348*.

[9] Codeluppi, S.; Borm, L. E.; Zeisel, A.; La Manno, G.; van Lunteren, J. A.; Svensson, C. I.; Linnarsson, S. Spatial organization of the somatosensory cortex revealed by osmFISH. *Nature Methods* **2018**, *15*, 932–935.

[10] Eng, C.-H. L.; Lawson, M.; Zhu, Q.; Dries, R.; Koulena, N.; Takei, Y.; Yun, J.; Cronin, C.; Karp, C.; Yuan, G.-C.; Cai, L. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **2019**, *568*, 235–239.

[11] Goh, J. J. L.; Chou, N.; Seow, W. Y.; Ha, N.; Cheng, C. P. P.; Chang, Y. C.; Zhao, Z. W.; Chen, K. H. Highly specific multiplexed RNA imaging in tissues with split-FISH. *Nature Methods* **2020**, *17*, 689–693.

[12] Axelrod, S.; Cai, M.; Carr, A.; Freeman, J.; Ganguli, D.; Kiggins, J.; Long, B.; Tung, T.; Yamauchi, K. Starfish: Scalable Pipelines for Image-Based Transcriptomics. *Journal of Open Source Software* **2021**, *6*, 2440.

[13] Cisar, C.; Keener, N.; Ruffalo, M.; Paten, B. A unified pipeline for FISH spatial transcriptomics. *Cell Genomics* **2023**, 100384.

[14] Valen, D. A. V.; Kudo, T.; Lane, K. M.; Macklin, D. N.; Quach, N. T.; DeFelice, M. M.; Maayan, I.; Tanouchi, Y.; Ashley, E. A.; Covert, M. W. Deep Learning Automates the Quantitative Analysis of Individual Cells in Live-Cell Imaging Experiments. *PLOS Computational Biology* **2016**, *12*, e1005177.

[15] Stringer, C.; Wang, T.; Michaelos, M.; Pachitariu, M. Cellpose: A generalist algorithm for cellular segmentation. *Nature Methods* **2020**, *18*, 100–106.

[16] Greenwald, N. F. et al. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nature Biotechnology* **2021**, *40*, 555–565.

[17] Pachitariu, M.; Stringer, C. Cellpose 2.0: how to train your own model. *Nature Methods* **2022**, *19*, 1634–1641.

[18] Mabaso, M.; Withey, D.; Twala, B. Spot detection methods in fluorescence microscopy imaging: A review. *Image Analysis and Stereology* **2018**, *37*, 173–190.

[19] van der Walt, S.; Schönberger, J. L.; Nunez-Iglesias, J.; Boulogne, F.; Warner, J. D.; Yager, N.; Gouillart, E.; Yu, T. scikit-image: image processing in Python. *PeerJ* **2014**, *2*, e453.

[20] Allan, D. B.; Caswell, T.; Keim, N. C.; van der Wel, C. M.; Verweij, R. W. soft-matter/trackpy: Trackpy v0.5.0. 2021; `https://zenodo.org/record/4682814`.

[21] Gudla, P.; Nakayama, K.; Pegoraro, G.; Mistelli, T. SpotLearn: Convolutional Neural Network for Detection of Fluorescence In Situ Hybridization (FISH) Signals in High- Throughput Imaging Approaches. *Cold Spring Harbor Symposia on Quantitative Biology* **2017**, *82*, 57–70.

[22] Eichenberger, B. T.; Zhan, Y.; Rempfler, M.; Giorgetti, L.; Chao, J. A. deepBlink : threshold-independent detection and localization of diffraction-limited spots. *Nucleic Acids Research* **2021**, *49*, 7292–7297.

[23] Wollmann, T.; Rohr, K. Deep Consensus Network: Aggregating predictions to improve object detection in microscopy images. *Medical Image Analysis* **2021**, *70*, 102019.

[24] Ratner, A.; Bach, S. H.; Ehrenberg, H.; Fries, J.; Wu, S.; Ré, C. Snorkel: rapid training data creation with weak supervision. *The VLDB Journal* **2019**, *29*, 709–730.

[25] Hoffman, M. D.; Blei, D. M.; Wang, C.; Paisley, J. Stochastic Variational Inference. *Jounral of Machine Learning Research* **2013**, *14*, 1303–1347.

[26] Petukhov, V.; Xu, R. J.; Soldatov, R. A.; Cadinu, P.; Khodosevich, K.; Moffitt, J. R.; Kharchenko, P. V. Cell segmentation in imaging-based spatial transcriptomics. *Nature Biotechnology* **2022**, *40*, 345–354.

[27] Liu, J. et al. Concordance of MERFISH spatial transcriptomics with bulk and single-cell RNA sequencing. *Life Science Alliance* **2022**, *6*, e202201701.

[28] Feldman, D.; Singh, A.; Schmid-Burgk, J. L.; Carlson, R. J.; Mezger, A.; Garrity, A. J.; Zhang, F.; Blainey, P. C. Optical Pooled Screens in Human Cells. *Cell* **2019**, *179*, 787–799.e17.

[29] Palla, G.; Spitzer, H.; Klein, M.; Fischer, D.; Schaar, A. C.; Kuemmerle, L. B.; Rybakov, S.; Ibarra, I. L.; Holmberg, O.; Virshup, I.; Lotfollahi, M.; Richter, S.; Theis, F. J. Squidpy: a scalable framework for spatial omics analysis. *Nature Methods* **2022**, *19*, 171–178.

[30] Satija, R.; Farrell, J. A.; Gennert, D.; Schier, A. F.; Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* **2015**, *33*, 495–502.

[31] Beliveau, B. J.; Kishi, J. Y.; Nir, G.; Sasaki, H. M.; Saka, S. K.; Nguyen, S. C.; ting Wu, C.; Yin, P. OligoMiner provides a rapid, flexible environment for the design of genome-scale oligonucleotide in situ hybridization probes. *Proceedings of the National Academy of Sciences* **2018**, *115*, E2183–E2192.

[32] Moffitt, J. R.; Hao, J.; Wang, G.; Chen, K. H.; Babcock, H. P.; Zhuang, X. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences* **2016**, *113*, 11046–11051.

[33] Boersma, S.; Rabouw, H. H.; Bruurs, L. J.; Pavlovič, T.; van Vliet, A. L.; Beumer, J.; Clevers, H.; van Kuppeveld, F. J.; Tanenbaum, M. E. Translation and Replication Dynamics of Single RNA Viruses. *Cell* **2020**, *183*, 1930–1945.e23.

[34] Thompson, R. E.; Larson, D. R.; Webb, W. W. Precise Nanometer Localization Analysis for Individual Fluorescent Probes. *Biophysical Journal* **2002**, *82*, 2775–2783.

[35] Moon, T. The expectation-maximization algorithm. *IEEE Signal Processing Magazine* **1996**, *13*, 47–60.

[36] Gataric, M.; Park, J. S.; Li, T.; Vaskivskyi, V.; Svedlund, J.; Strell, C.; Roberts, K.; Nilsson, M.; Yates, L. R.; Bayraktar, O.; Gerstung, M. PoSTcode: Probabilistic image-based spatial transcriptomics decoder. *bioRxiv* **2021**,

[37] Smal, I.; Loog, M.; Niessen, W.; Meijering, E. Quantitative comparison of spot detection methods in fluorescence microscopy. *IEEE Transactions on Medical Imaging* **2010**, *29*, 282–301.

[38] Ruusuvuori, P.; Äijö, T.; Chowdhury, S.; Garmendia-Torres, C.; Selinummi, J.; Birbaumer, M.; Dudley, A. M.; Pelkmans, L.; Yli-Harja, O. Evaluation of methods for detection of fluorescence labeled subcellular objects in microscope images. *BMC Bioinformatics* **2010**, *11*.

[39] Zhao, P.; Zhu, J.; Ma, Y.; Zhou, X. Modeling zero inflation is not necessary for spatial transcriptomics. *Genome Biology* **2022**, *23*.