PNAS
www.pnas.org

Supplementary Information for

Spatial transcriptome profiling by MERFISH reveals sub-cellular RNA compartmentalization and cell-cycle dependent gene expression

Chenglong Xia[a,b,c,1], Jean Fan[a,b,c,1], George Emanuel[a,b,c,1], Junjie Hao[a,b,c], Xiaowei Zhuang[a,b,c,2]
[a] Howard Hughes Medical Institute, Harvard University, Cambridge, MA 02138;
[b] Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138;
[c] Department of Physics, Harvard University, Cambridge, MA 02138
[1] C.X., J.F. and G.E. contributed equally to this work.
[2] Corresponding author: zhuang@chemistry.harvard.edu (X.Z.)

This PDF file includes:

SI Materials and Methods
Figs. S1 to S11
References for SI reference citations

Other supplementary materials for this manuscript include the following:

Datasets S1 to S15

SI Materials and Methods

Design and construction of the MERFISH encoding probes

The MERFISH encoding probes were designed to encode each RNA species with a unique binary codeword from a set of 69-bit, Hamming distance 4 (HD 4), Hamming weight 4 (HW 4), binary barcodes (Dataset S1). We constructed a set of barcodes following this encoding scheme by converting a tabulated list of HD4 and HW4 70-bit barcodes from the La Jolla covering repository and converting to a list of 69-bit barcodes by removing all barcodes that contained a 1 in the 70^{th} bit and then removing the 70^{th} bit from all remaining barcodes, yielding 12,903 barcodes 69-bit barcodes. Every barcode in this list is separated by a Hamming distance of at least 4 from all other barcodes, hence allowing one- and two-bit errors to be detected and one-bit errors to be corrected (1, 2). In addition, each barcode has a constant Hamming weight (i.e., the number of "1" bits in each barcode) of 4 to avoid potential bias in the measurement of different barcodes due to a differential rate of "1" to "0" and "0" to "1" errors (2). To imprint each binary barcode onto a target gene isoform, we constructed a set of 30-nt target regions that could bind to the target transcript with high specificity, by using an algorithm published previously (3). We selected target regions that had a GC fraction between 33 and 73%, a melting temperature T_M range of 61-81 °C, an isoform-specificity index between 0.75 and 1, a gene-specificity index between 0.75 and 1, and no homology longer than 15 nt to rRNAs or tRNAs, as described previously (3). Briefly, to ensure that each target region specifically bound to the target gene isoform, we constructed a list of all unique 17-nt sequences within any of the isoforms of the target gene and for each 17-nt sequence, we calculated the isoform-specificity as the ratio between the abundance of the target isoform, based on RNA sequencing, and the sum of the abundance of all isoforms in which the 17-nt sequence appears. The isoform-specificity of each 30-nt target region was calculated as the average isoform-specificity of all 17-nt sequences contained within the 30-nt target region and only target regions with a specificity between 0.75 and 1 were used for encoding probes.

We selected 9,050 of the annotated gene isoforms that were long enough to construct 48 target regions without overlap and 1,000 of the annotated gene isoforms that were not long enough to construct 48 target regions without overlap, both with FPKM values between 0.01 to 500 based on matched bulk RNA-sequencing (4). Each of the 10,050 selected RNA species was randomly assigned a unique binary barcode from all 12,903 possible 69-bit barcodes. The remaining unassigned 2,853 barcodes were left as blank controls for misidentification quantification.

Then, for each RNA species, we designed up to 48 encoding probes to imprint the selected binary codeword onto the corresponding transcript, each encoding probe containing a 30-nt target sequence complementary to one of the 30-nt target regions on the RNA and three 20-nt readout sequences. First, each of the 69 bits was assigned a 20-nt three-letter readout sequence designed as previously published (3). For the set of 9,050 genes, we then identified 48, 30-nt non-overlapping target regions for each gene. For the set of 1,000 genes for which fewer than 48 non-overlapping target regions could be designed, we identified 48, 30-nt target regions with a 20-nt overlap between adjacent target regions for each gene as previously described (5, 6). We have previously shown that this overlapping design did not substantially reduce signal from individual

RNA molecules, because any given cellular RNA is typically not bound by all targeting encoding probes (5, 6). Afterwards, each encoding probe was constructed to contain a unique 30-nt target region and three of the four 20-nt readout sequences corresponding to the four '1' bits assigned to the transcript. The encoding probes additionally contained two priming regions for amplification: a 20-nt primer binding region at the 5' end and the reverse complement of the T7 promoter at the 3' end. We then removed encoding probes that had homology to human rRNA or tRNA or to mitochondrial rRNA or tRNA. The encoding probe libraries for the 9,050 longer genes and 1,000 shorter genes were purchased as two oligo pools (Twist Biosciences) and amplified as previously described (7). The encoding probe sequences for all genes are listed in Dataset S2.

The encoding probes for the 130-gene measurements were described previously (3). Briefly, a 16-bit, HD4, HW4 code was used to encode these genes and each gene was targeted with 92 non-overlapping encoding probes. 10 of the 140 codewords in the 16-bit, HD4, HW4 code were left unassigned as blank control barcodes.

Design and construction of the MERFISH readout probes

For the 10,050-gene measurements, 69 readout probes were designed (Dataset S3), each complementary to one of the 69 readout sequences, corresponding to the 69 bits in the barcodes. For the 130-gene measurements, 16 readout probes were used, as previously described (3). Readout probes were conjugated to one of the dye molecules (Alexa750, Cy5, and Atto565/Cy3B) via a disulfide linkage, as described previously (3). These readout probes were synthesized and purified by Bio-synthesis, Inc., resuspended immediately in Tris-EDTA (TE) buffer, pH 8 (Thermo Fisher) to a concentration of 100 μM and stored at -20 °C.

Sample staining and expansion

U-2 OS cells (ATCC) were cultured with Eagle's Minimum Essential Medium (EMEM) (ATCC) containing 10% (vol/vol) FBS and 50 U/mL Penicillin-Streptomycin (Thermo Fisher), plated on 18mm-diameter, no.1.5 coverslips (Neuvitro) at a density of 25K cells per coverslip and were grown in a 12-well plate at 37 °C with 5% CO2 for 48–72 hours. Then, cells were fixed in 4% paraformaldehyde (Electron Microscopy Sciences) in 1× PBS at room temperature for 15 min. After washing the cells three times with 1× PBS, 0.5% (vol/vol) Triton X-100 (Sigma) in 1× PBS was used to permeabilize the cells for 10 min at room temperature. Cells were again washed three times with 1× PBS after permeabilization.

To perform immunostaining of cellular structures (i.e. ER) with oligonucleotide-labeled antibodies, the permeabilized cells were incubated in blocking buffer (3% (wt/vol) ultra-pure BSA (Thermo Fisher), 3% (vol/vol) RNasin Ribonuclease inhibitor (Promega), 6% (vol/vol) murine RNAase inhibitor and 1 mg/ml yeast tRNA in 1xPBS) for 1 hour. Then the cells were incubated with 10 μg/ml primary antibodies (anti-KDEL, Abcam) in blocking buffer for 1 h at room temperature, and washed three times with 1× PBS for 5 min each. Cells were then stained with 5 μg/ml oligonucleotide-labeled secondary antibodies in blocking buffer for 1 h at room

temperature, then washed with 1× PBS three times for 5 min each. To prevent antibody dissociation during MERFISH encoding probe staining, samples were fixed with 4% (vol/vol) PFA in 1× PBS for 15 min and washed three times with 1× PBS for 5 min each. For the 130-gene measurements used to estimate the detection efficiency of our 10,050-gene measurements, we did not perform this immunostaining step.

To perform MERFISH encoding probe staining, cells (after ER immunolabeling for 10,050-gene measurements or after permeabilization for 130-gene measurements) were incubated for 30 min in encoding wash buffer comprising 2× SSC (Ambion), 30% (vol/vol) formamide (Ambion), and 0.1% (vol/vol) murine RNase Inhibitor (New England Biolabs). A drop of 25 µL of 10,050 gene encoding probes (~1 nM per encoding probe, up to 48 probes per gene) or 130-gene encoding probes (~1 nM per encoding probe, up to 92 probes per gene) and acrydite-modified poly(dT) locked nucleic acid (LNA) probes with a unique 20-nt readout sequence (2 µM) in encoding hybridization buffer was added to a parafilm in a 150mm-diameter dish and were covered with cell-containing coverslips. Encoding hybridization buffer is comprised of encoding wash buffer supplemented with 0.1% (wt/vol) yeast tRNA (Thermo Fisher), 0.1% (vol/vol) murine RNase Inhibitor (New England Biolabs), and 10% (wt/vol) dextran sulfate (Sigma). The samples were then incubated in a cell culture incubator at 37 °C for 36–48 h, followed by washing with encoding wash buffer twice at 47 °C for 30 min. The samples then were incubated for 10 min with a 1:500 dilution of 0.12-µm-diameter light yellow beads (Spherotech, FP-0245-2), which were used as fiducial markers to align images obtained from sequential rounds of hybridization.

To expand the samples labeled with oligonucleotide-conjugated antibody and/or MERFISH encoding probes, we adopted a previously published gel embedding and expansion protocol (5, 8). Briefly, the samples were first incubated for 10 min in degassed monomer solution consisting of 2 M NaCl, 7.7% (wt/vol) sodium acrylate (Sigma), 4% (vol/vol) of 19:1 acrylamide/bis-acrylamide, 60 mM Tris-HCl pH 8 and 0.2% (vol/vol) TEMED (for buffer exchange). The remaining monomer solution were then kept on ice and further added 0.2% (wt/vol) ammonium persulfate just before casting the expansion gel.

To cast the gel, 25 µL of the ice-cold monomer solution containing 0.2% (wt/vol) ammonium persulfate were added to the surface of a glass plate that had been treated with GelSlick (Lonza). Coverslips containing the antibody- and/or encoding-probe-labeled cells were dried quickly with KimWipes (Kimtech) and inverted onto the 25 µL droplet to form a uniform thin layer of gel solution between the coverslip and the glass plate. The sandwiched coverslip and glass plate with gel solution were then transferred to a nitrogen-filled chamber for 2 hours to complete gel polymerization. The coverslip and the glass plate were then separated by using a thin razor blade. The coverslips with expansion gels were transferred to a digestion buffer, which was made up of 2% (wt/vol) Sodium dodecyl sulfate (SDS) (Thermo Fisher), 0.5% (vol/vol) Triton X-100 and 1% (vol/vol) Proteinase K (New England Biolabs) in 2× SSC. The samples were digested in digestion buffer overnight in a 37 °C incubator. After digestion, samples were expanded in 0.8× SSC buffer supplemented with 0.2% (vol/vol) Proteinase K at room temperature. During expansion, the buffer was changed every 20 minute for three times.

To stabilize the expanded gel for sequential rounds of readout probe hybridization and imaging, we re-embedded the expansion gel in a none expandable polyacrylamide gel. Briefly, expanded polyacrylite gel was buffer exchanged in re-embedding solution comprised of 4% (vol/vol) of 19:1 acrylamide/bis-acrylamide ,100 mM NaCl, 60 mM Tris ·HCl pH 8 and 0.2% (vol/vol) of TEMED for 20 min. The remaining re-embedding solution was kept on ice and further supplemented with 0.2% (wt/vol) ammonium persulfate just before casting the none expandable polyacrylamide gel. The expanded polyacrylate gels were transferred to silanized coverslips prepared as previously described (3), dried with KimWipes and added 100 µl re-embedding gel solution with ammonium persulfate. The samples were then put in a nitrogen chamber, covered with a GelSlick treated glass plate to facilitate polyacrylamide gel polymerization at room temperature for 1 h. The coverslip and the glass plate were again separated by using a razor. The samples were either used for MERFISH measurement immediately or stored in 2× SSC containing 0.1% (vol/vol) murine RNase inhibitor at 4 °C for no longer than a week.

MERFISH imaging platforms

The samples were imaged on a home-built imaging platform. This microscope was constructed around a Nikon Ti2 Eclipse microscope body with a Nikon, CFI Plan Apo Lambda 60x oil objective. Illumination at 750, 647, 560, 488 and 405 nm was provided using solid-state lasers (MBP Communications, 2RU-VFL-P-500-750-B1R; MBP Communications, 2RU-VFL-P-2000-647-B1R; MBP Communications, 2RU-VFL-P-2000-560-B1R; MBP Communications, 2RU-VFL-P-500-488-B1R; Coherent, Obis 405). A DAQ card (National Instruments, PCIe-6323) was used to modulate the illumination from each laser, via digital inputs to an ATOF (Gooch & Housego, 97-03309-01) for the 750, 647, 560, and 488 nm lasers and via a direct digital input for the 405 nm laser. These laser lines were used to excite readout probes labeled with Alexa750, Cy5, and Atto565/Cy3b, a probe complementary to the poly-dT readout sequence labeled with Alexa 488, and DAPI and light yellow fiducial beads, respectively. To evenly distribute the illumination over the square-shaped area imaged by the camera, the Gaussian beam profile from the single mode lasers was flattened to a square profile by coupling the laser lines into a square code multi-mode fiber (Thorlabs, M103L05) and imaging the face of the fiber onto the sample. The laser speckle from the multimode fiber was homogenized by attaching a vibration motor (Precision Microdrives, 912-101) via a custom 3D-printed adapter to shake the fiber at a frequency much faster than the camera framerate. The fluorescence emission from the sample was separated from the laser illumination using a penta-band dichroic (Chroma, zy405/488/561/647/752RP-UF1) and imaged with a scientific CMOS camera (Hamamatsu, Orca Flash V3) after passing through a custom penta-notch filter (Chroma, ZET405/488/561/647–656/752m) to remove stray excitation light. The pixel size for the sCMOS camera corresponded to 109 nm in the sample plane. The exposure time was 500 ms for each imaging frame. The sample was translated using a motorized microscope stage (Marzhauzer, SCAN IM 130 x 85). To maintain the same focal plane over the course of each experiment, an IR laser (Thorlabs, LP980-SF15) was obliquely reflected off the sample coverslip interface onto a CMOS camera (Thorlabs, DCC1545M) to measure the distance between the sample coverslip and the objective and the sample-to-objective distance was controlled by an objective nanopositioner (MCL, Nano-F100S).

The sample coverslip was held inside a flow chamber (Bioptechs, FCS2), and buffer exchange within this chamber was directed using a custom-built automated fluidics system, composed of three, 12-port valves (IDEX, EZ1213-820-4) and a peristaltic pump (Gilson, MP3), configured as described previously (2).

Sample imaging

Immunofluorescence and MERFISH imaging were carried out on the imaging platform described above. Briefly, in the first round we imaged the cell nucleus stained with DAPI, ER marker stained with an Alexa750 labeled readout probe, and poly-A stained with Alexa488 labeled readout probe complementary to the poly-dT LNA anchor probe for the 10,050-gene measurements or DAPI and ploy-dT staining for 130-gene measurements. We then performed 23 or 6 rounds of three-color MERFISH imaging for the 10,050-gene measurements or the 130-gene measurements, respectively. DAPI and ER marker stains were imaged at 6 focal planes separated by 700 nm in z and the poly-dT stain was imaged in a single z-slice. Each MERFISH round consisted of readout probe hybridization (10 min), washing (5 min), and imaging of 256 tiled fields of view (FOV) (220 um x 220 um per FOV, 40-100 min total time), readout fluorophore cleavage by TCEP (15 min), and rinsing with 2xSSC (5 min). For each round, images were acquired with 750-nm, 650-nm, and 560-nm illumination at 6 focal planes separated by 700 nm in z to image the readout probes in addition to a single z-plane image with 405-nm illumination to image the fiducial beads on the glass surface for image registration. Specifically, the readout probes staining was done by flowing 10 nM (each) readout probes in hybridization buffer composed of 2× SSC, 10% (vol/vol) ethylene carbonate (Sigma-Aldrich, E26258), 0.1% (vol/vol) murine RNase inhibitor (NEB), 0.5% (vol/vol) Triton X-100 and 0.4% (vol/vol) TWEEN 20 in nuclease-free water, following by washing with a wash buffer containing 2× SSC and 10% (vol/vol) ethylene carbonate in nuclease-free water. TWEEN 20 can passivate coverslip surface and reduce none specific readout probe binding (9). Readout probe imaging was then performed in imaging buffer containing 5 mM 3,4-dihydroxybenzoic acid (Sigma), 2 mM trolox (Sigma), 50 μM trolox quinone, 1:500 recombinant protocatechuate 3,4-dioxygenase (rPCO; OYC Americas), 1:500 Murine RNase inhibitor, and 5 mM NaOH (to adjust pH to 7.0) in 2×SSC. By imaging 6 focal planes separated by 700 nm, we imaged a ~4.2 μm depth of the expanded sample, corresponding to a ~2 μm pre-expansion depth. This depth covered the full height of the peripheral cytoplasmic regions and more than half of the depth of the nuclei. However, our subsequent analysis will not be adversely affected by this partial coverage of the cell, as described later. After imaging, dyes on the readout probe were removed by flowing in cleavage buffer comprising 2× SSC and 50 mM of Tris (2-carboxyethyl) phosphine (TCEP; Sigma) to cleave the disulfide bond connecting dyes to the probes. The imaging buffer and hybridization buffer were changed every 12 hours to prevent the decay of imaging quality.

Image processing and decoding

To align the 3-dimensional x, y, z image stacks for all 69 bit measurements, we calculated the offset between the corresponding high-pass filtered fiducial bead images with subpixel resolution

by finding the peak of the image cross-correlation and applied these offsets to align the image stacks.

To identify the RNA transcripts in the registered image stacks, we adapted our previously published voxel-based decoding algorithm (3). In short, we first assigned barcodes to each voxel independently, then aggregated adjacent voxels with the same barcodes into putative RNA molecules, and filtered the list of putative RNA molecules to enrich for correctly identified transcripts. The details are elaborated below.

First, to assign each voxel to one of the 12,903 possible barcodes, we compared the 69-dimensional intensity vectors measured for each voxel to the set of 69-bit valid barcodes. We expected a voxel containing a given encoded RNA to have a higher fluorescence intensity in the four imaging rounds corresponding to the four "1" bits of the associated barcode than in the remaining 65 imaging rounds corresponding to the 65 "0" bits. To aid comparison, we normalized intensity variations between different spots and across the point spread function of a single spot by dividing the 69-dimensional intensity vector for each voxel by its L2 norm. We similarly normalized each of the 12,903 valid 69-bit barcodes. To assign a barcode to each voxel, we identified the normalized barcode vector that was closest to the voxel's normalized intensity vector in the 69-dimensional space.

Since our encoding scheme enforced a minimum Hamming distance of 4 between all pairs of barcodes, which allows correction of single-bit errors and detection but not correction of two-bit errors, we excluded barcodes with two-bit or greater errors using the following two-step approach. We recognize that because of the normalization strategy described above, barcodes with a single-bit error can have different distances to the correct barcodes depending on the type of single-bit errors. While the correct barcodes have four '1' bits and an L2 norm of 2, a '1' to '0' bit error would result in barcode with only three '1' bits and an L2 norm of $\sqrt{3}$ and a '0' to '1' bit error would result in a barcode with five '1' bits and an L2 norm of $\sqrt{5}$. Consequently, after dividing each by the L2 norm, error barcodes resulting from '1' to '0' errors are at a larger distance from the true barcodes than errors barcodes resulting from '0' to '1' errors. The barcodes with two-bit errors will thus have a maximum distance to the correct barcodes if both errors are of the '1'-to-'0' type and a minimum distance to the correct barcodes if both errors are of the '0'-to-'1' type. We thus excluded voxels that were farther than the maximum two-bit-error distance away from any valid barcode in the first step. For the remaining voxels, we then grouped adjacent voxels that were assigned to the same barcode to create a list of putative transcripts. Then in the second step, we removed putative transcripts whose minimum distance to a valid barcode is larger than the minimum two-bit-error distance.

To further reduce the chance of transcript misidentification, for each remaining putative RNA molecule, we calculated three properties: the mean intensity of its corresponding voxels, referred to as "voxel intensity", the number of voxels contained in the transcript, referred to as "voxel number", and the minimum distance between the normalized intensity vector of the corresponding voxels to the closest valid barcode, referred to as "vector distance" hereafter. We

anticipated that incorrectly identified barcodes would likely have a dimmer mean voxel intensity, a smaller voxel number, and a larger vector distance as compared to correctly identified barcodes. Then we constructed two 3D histograms of the three parameters: one containing all remaining putative transcripts and the other containing the subset of the remaining putative transcripts that were assigned to blank control barcodes (Fig. S2A, B). As expected, we observed that the putative transcripts assigned to blank barcodes, which are known to be misidentified, appeared concentrated in a region that is dimmer in mean voxel intensity, smaller in the voxel number, and greater in vector distance (Fig. S2B). We then calculated the fraction of blank control barcodes within each 3D histogram bin and any bin with a high fraction indicates a point with a high chance of barcode misidentification in the "voxel intensity"-"voxel number"-"vector distance" 3D space. We then set a threshold on this fraction and excluded putative transcripts that fall in the "voxel intensity"-"voxel number"-"vector distance" 3D space where the fraction is higher than the threshold. As expected, the gross barcode misidentification rate, estimated as (the mean count per blank control barcode per cell) / (the mean count per barcode per cell), where the barcodes in the denominator includes both RNA-encoding and blank control barcodes, decreased as we decreased the threshold blank fraction value (Fig. S2C). In this work we chose the blank fraction threshold such that the gross barcode misidentification rate is 5%.

We note that this gross misidentification rate includes misidentifications into both RNA-encoding and blank control barcodes. An experimentally more relevant estimation is to consider only misidentifications into RNA-encoding barcodes as true misidentifications because identified blank control barcodes are typically not used to draw biological conclusion. Thus, an alternative estimation of the misidentification rate, which considers only misidentification into another RNA-encoding barcode, can be calculated as (the mean count per blank control barcodes per cell) / (the mean count per RNA-encoding barcodes per cell), yielding a misidentification rate of 3.9% for our 10,050-gene measurements. Still, both calculations assume a constant misidentification rate for all barcodes, but the misidentification rate could be different for different barcodes. We observed that not all blank control barcodes were detected at the same rate, indicating some barcodes may have been systematically misidentified more frequently than others. Furthermore, barcodes encoding high abundance genes are expected to have a lower misidentification rate since the misidentified barcodes are expected to be a smaller fraction due to the large number of correctly identified barcodes, while barcodes encoding low abundance genes are expected to have a higher misidentification rate. Additionally, these calculations could have underestimated the misidentification rates since encoding probes non-specifically bound to the sample may have a higher probability to be misidentified as an RNA-encoding barcode than as a blank control barcode because blank barcodes were not assigned encoding probes.

We observed that the decoding quality was affected by (i) intensity variations between images for different bits and (ii) small distortions between the images from different fluorescence channels caused by chromatic aberration. To correct these artifacts, we iteratively estimated (i) scale factors to normalize intensity differences between different bit measurements and (ii) transformations to correct the chromatic aberrations between each fluorescence channel. For each iterative step, we decoded a single z-slice from 50 randomly selected fields of view and used the decoded barcodes to estimate the necessary corrections. From the decoded barcodes, we

calculated a scale factor for each bit such that the resulting mean '1' bit intensity for all 69 bits would be equal for the decoded barcodes. Similarly, we refined the chromatic corrections by calculating the residual displacement between the fluorescence spots corresponding to the '1' bits for each decoded barcode in different color channels and then determined a new set of chromatic transformations for each pair of fluorescence channels that minimized the chromatic displacements. We found that both the scale factors and the chromatic corrections converge within 10 iterations.

The 130-gene measurements were decoded as previously described (3).

Image segmentation for cellular structures

All image segmentation was performed using the EBImage (v4.24.0) package (10) in R (v3.5.0). All images of cellular structures (i.e. KDEL-immunostained ER, DAPI stained nuclei, or poly-dT stained cells) were binarized by global thresholding after high-pass filtering by subtracting the convolution of the image with a disc of diameter *t* pixels from the original image. The diameter *t* was selected depending on the resolution of the feature to be segmented. As mentioned earlier, we imaged 6-slices separated by 700 nm each. For each cellular structure, except for an out-of-focus effect, we did not observation substantial difference in the cellular-structure stain in different z-slices. However, since the out-of-focus blurring effect negatively impact segmentation accuracy, we used the z-slice containing the most in-focus features for segmentation.

We first segmented cell nuclei using the DAPI image. With our epi-fluorescence imaging, the nuclear images appeared nearly identical for all imaged z-slices and the slice corresponding to z = 2.1 μm was selected for segmentation. We high-pass filtered using a disc of diameter *t* = 1001 pixels and binarized the staining image using an offset of 0.01 in intensity on the normalized image. To remove holes that may have been introduced by subnuclear imaging details, we dilated using a 51-pixel diameter disc and fill all convex hulls. To remove small artifacts that may have been segmented, we restricted to segmented entities with contiguous pixels greater than a 201-pixel diameter disc. We subsequently erode by a 51-pixel diameter disc to compensate for the dilation for the final nuclear segmentation. Alternatively, we erode by a 101-pixel diameter to achieve a conservative nuclear segmentation.

We next segmented cell bodies by combining a poly-dT-staining image with decoded RNAs. We again first applied a high-pass filtered using a disk of diameter *t* = 11 pixels. We then binarized the poly-dT-staining images using an offset of 0.01 in intensity on the normalized image and added pixels corresponding to positions of decoded mRNAs. We applied a 3-pixel median filter in order to remove stray pixels.

We then used the segmented nuclei as seeds and segmented cell bodies as a mask in a Voronoi-based segmentation on image manifolds (11) established by the poly-dT-staining image. We assigned unique IDs for each cell by labeling each contiguous group of pixels in the nuclear mask with a unique integer and labeling the corresponding filled watershed from the nuclear seed with

the same integer. In this manner, nuclei and cell bodies are assigned consistent cell IDs for downstream analysis.

We next segmented the ER using the KDEL staining image. The ER in the cell peripheral region appeared in focus or largely in focus in the first two z-slices, and out of focus in the remaining z-slices that were focused above the peripheral region of the cell. The ER structures appeared essentially identical between the two first z slices. The ER in the perinuclear region could be detected over more z-slices, and also appeared similar across z-slices. We selected one of the first two z-slices where ER images were most in focus for segmentation. We used a two-pass binarization approach in order to capture both ER structures in the perinuclear region as well as finer ER structure in peripheral region of the cell. We first applied a high-pass filter using a disc of diameter t = 501 pixels to segment large, dense ER structures in the perinuclear region and subsequently binarized the image as described previously. We then applied a high-pass filter with a disc of diameter t = 51 pixels on the original staining image to capture sparser and finer ER structures in the peripheral region of the cell and subsequently binarized. We then combined the two binarized images to create a final binary mask. We applied a 7-pixel median filter and further restricted to segmented bodies with contiguous pixels greater than a 3-pixel radius disc to minimize stray pixels and small artifacts.

Since some pixels from the ER segmentation overlapped with the nuclear segmentation, we created a combined ER-nuclear segmentation mask that contained all pixels that were segmented in either the ER mask or the nuclear mask. We assigned each pixel of this combined ER-nuclear segmentation mask the ID of the cell/nucleus mask at the same coordinates.

ER enrichment, nuclear enrichment, and comparison to previously published approaches

To quantify the expression levels of RNAs in certain cellular structures, we assessed the pixel-level overlap between RNA transcripts and segmented cellular structures for each cell individually, resulting in gene by cell count matrices for each structure. Specifically, for each unique cell ID assigned during segmentation, we counted the number of detected transcripts of each RNA species that fell within the pixels of the nuclear mask and the cell body mask assigned to the ID. This directly yielded the nuclear counts as the counts within the nuclear mask, but since the cell body mask contained both the cytoplasm and the nucleus, we calculated the cytoplasm counts by subtracting the counts within the nuclear mask from the counts within the cell body mask. Similarly, to calculate ER counts, we counted the number of each RNA species within the combined ER-nuclear mask and subtracted the counts from the nuclear mask.

To identify genes enriched at the ER, we compared expression within the ER-positive region versus the ER-negative region of the cytoplasm. For each region we calculated the CPM-normalized expression of each gene in each cell: the abundance of each RNA species divided by the abundance of all RNA species in the corresponding cellular compartment and multiplied by a million. These normalized counts are proportional to the percent of total RNA counts that belongs to each RNA species within the specified cellular compartment. Therefore, although our measurements did not cover the entire cell volume (by imaging 6 focal planes, we covered most

of the cytoplasm, and more than half of the thickness of the nucleus), our CPM normalization normalized out the effect of the volume fraction that we measured for each compartment.

Differentially expressed genes between ER-positive and ER-negative regions of the cytoplasm were then identified using a two-sided pair-wise Wilcoxon-rank-sum test. Resulting p-values were corrected following a strict Bonferroni correction for multiple hypothesis testing. For the ER, we identified genes as highly significantly enriched using the criteria log2(fold-change) > 0 and adjusted-p-value < 1e-10.

To identify genes retained in the nucleus, we compared CPM-normalized expression within the nucleus versus cytoplasm, using the two-sided pair-wise Wilcoxon-rank-sum test noted previously. We identified genes as highly significantly enriched in the nucleus using stringent criteria of log2(fold-change) > 2 and adjusted-p-value < 1e-10. As in the ER enrichment analysis, our CPM normalization also normalized out the effect of the partial nucleus and cell coverage in the nuclear-enrichment analysis.

To compare with previously published approaches, we downloaded previous ER enrichment results by APEX-RIP and proximity-associated ribosomal profiling from the supplemental tables of Kaewsapsak et.al. (12). For comparative purposes, we matched genes by Ensembl identifiers and restricted comparison to genes that are commonly expressed at FPKM > 1 in both U2-OS and H3K293T cells based on bulk RNA-sequencing.

As mentioned earlier, from the 6 imaged z-slices, we used the z-slice containing the most in-focus cellular structure (ER or nucleus) image for segmentation because except for an out-of-focus effect, we did not observation substantial difference in the cellular-structure stain in different z-slices. Although this could lead to some errors in the identification of colocalized transcripts in other z-slices, we expect this problem to be largely mitigated by our statistical enrichment analysis and the strict enrichment criteria used, as described above. In this approach, a gene would only be detected as being statistically significantly enriched in the ER (or nucleus) based on the CPM-normalized gene counts if its transcripts have a higher tendency to be associated with these structures, not if some of its transcripts randomly moved close to, and thus appeared colocalized with the ER (or nucleus). Indeed, the fact that only ~1% of the blank control barcodes showed ER (or nucleus) enrichment and the agreement between our results and previous results obtained from different experimental methods support the validity of our analyses. However, we do expect that the accuracy of our analyses could be further improved by using confocal imaging or other optical segmentation methods to improve our imaging resolution in the z direction.

Gene set enrichment analysis

To identify gene sets over-represented among genes enriched at the ER, we performed rank-based gene set enrichment analysis (13) using the LIGER (v1.0) package (14) in R (v3.5.0). Significantly differentially expressed genes were ranked by fold change. Gene sets were restricted to the Cellular Component (CC) terms in Gene Ontology (GO) and further limited to gene sets for which more than 5 genes were included in our MERFISH gene library. To control for multiple-

hypothesis testing, we apply a false discovery rate (FDR) correction (15). We identify significantly positively enriched gene sets, in contrast to potential significantly depleted gene sets, as those with enrichment scores > 0, edge-values > 0, and FDR < 0.05.

To identify gene sets over-represented among genes that exhibited spatial heterogeneity (or lacked spatial heterogeneity) in expression across cells, where gene ranks were not available, we perform gene set enrichment analysis using a hypergeometric test. Gene sets were restricted to those under Gene Ontology terms "cellular communication" (GO:0007154) and "cellular response to stimulus" (GO:0051716) and further limited to gene sets for which more than 5 genes were included in our gene library.

Cell clustering and differential expression analysis

To perform cell clustering analysis, we added nuclear and cytoplasmic RNA counts to obtain single-cell expression counts. We restrict analysis to the 9050 genes measured using non-overlapping encoding probes. We further remove lowly expressed genes, which we define as genes with fewer than 1 count per cell on average, which corresponds to the 95^{th}-percentile average expression of blank barcodes, resulting in 6109 remaining genes.

To minimize clustering by experimental batches (replicates) (16), we performed batch correction using ComBat (17), which uses a parametric empirical Bayes framework for adjusting data for batch effects. We then performed CPM normalization on the batch-corrected counts to control for differences in cell size or partial coverage of cells. We normalized variance as a function of expression magnitude to identify 1598 overdispersed genes (generalized additive model fit with k = 5, alpha significance threshold = 0.05). We perform dimensionality reduction using principal component analysis (PCA), restricted to the 30 principal components (PCs) with the highest eigenvalues, and finally visualize using a 2D tSNE embedding (perplexity=10) (18). To identify transcriptionally distinct cell clusters, we perform graph-based Louvain community detection (19) (k=50) in the 30 PCs-space.

To identify differentially expressed genes across the identified cell cluster, we perform a one-sided Wilcoxon-rank based test. We defined genes as significantly differentially upregulated in one cell cluster if the gene had a Bonferroni-corrected p-value < 1e-10. We visualized these differentially upregulated genes using a heatmap by averaging the CPM-normalized gene expression magnitudes within identified cell clusters.

RNA velocity analysis based on nuclear and cytoplasmic RNA counts

In the original RNA velocity analysis based on single-cell RNA-sequencing (20), La Manno et al estimated the first time derivative of mRNA abundance by distinguishing between unspliced and spliced mRNAs, based on the underlying principal that mRNA are first transcribed in an unspliced form and subsequently spliced for translation. La Manno at al model cellular transcriptional dynamics using rate equations for the expected number of unspliced mRNA molecules u and spliced mRNA molecules s for a single gene as follows:

$$\frac{du}{dt} = \alpha(t) - \beta(t) * u(t)$$
$$\frac{ds}{dt} = \beta(t) * u(t) - \gamma(t) * s(t)$$

Where:
$\alpha(t)$ is the time-dependent rate of transcription
$\beta(t)$ is the time-dependent rate of splicing
$\gamma(t)$ is the time-dependent rate of degradation

Analogously, here, we defined RNA velocity based on spatially-resolved single-cell transcriptome imaging data from MERFISH measurements by distinguishing between mRNAs localized in nucleus versus the cytoplasm, based on the underlying principal that mRNAs are transcribed in the nucleus and subsequently exported into the cytoplasm for translation. Thus, we model cellular transcriptional dynamics using rate equations for the expected number of nuclear mRNA molecules n and cytoplasmic mRNA molecules c for a single gene as follows:

$$\frac{dn}{dt} = \alpha(t) - \lambda(t) * n(t)$$
$$\frac{dc}{dt} = \lambda(t) * n(t) - \gamma(t) * c(t)$$

Where:
$\alpha(t)$ is the time-dependent rate of transcription
$\lambda(t)$ is the time-dependent rate of nuclear export
$\gamma(t)$ is the time-dependent rate of degradation

In their original RNA velocity analysis, La Manno et al assume that the rate of transcription and rate of degradation are time-independent and constant, and set $\beta(t) = 1$, thus measuring all rates in units of the splicing rate (20). Likewise, we applied the same simplifying assumptions here for transcription and degradation rates, and set $\lambda(t) = 1$, thus measuring all rates in units of nuclear export. Here, we assumed that the values of λ are positive. We believe this to be a reasonable assumption for λ since the RNAs of nearly all cellular genes travel unidirectionally from the nucleus to the cytoplasm, rather than being reimported back from the cytoplasm. Still, we anticipate RNA velocity estimates to improve with orthogonal means of determining the nuclear export rate $\lambda(t)$ for each gene.

The above assumptions result in the following set of simplified rate equations:

$$\frac{dn}{dt} = \alpha - n(t)$$
$$\frac{dc}{dt} = n(t) - \gamma * c(t)$$

These rate equations provide the expected number of mRNAs observed at each timepoint and can be solved to extrapolate the expected mRNA abundances at future timepoints based on the current transcriptional state of the cell.

La Manno et al previously noted that γ can be estimated as the ratio between unspliced and spliced mRNAs,

$$\gamma = \frac{u}{s}$$

for steady state cells. Even if none of the cells were in steady state, they noted that γ could still be reasonably approximated using cells in extreme expression quantiles such as based on the top/bottom 5% of cells in terms of expression magnitude. Furthermore, pooling the values of u and s from across multiple cells results in more robust estimates of γ (20).

Likewise, here we note that at steady state,

$$\frac{dn}{dt} = \alpha - n = 0$$
$$\frac{dc}{dt} = n - \gamma c = 0$$
$$\alpha = n$$
$$\gamma = \frac{n}{c}$$

Thus, we also estimated the degradation rate γ for each gene as the ratio between nuclear and cytoplasmic mRNAs using pooled cells in the lower and upper 5th extreme expression quantiles. With γ estimated for each gene, we then calculated the RNA velocity $\frac{dc}{dt} = v = n - \gamma c$ based on the nuclear and cytoplasmic expression levels n and c, which also allowed us to predict the cytoplasmic expression at a future time $c(t + dt) = c(t) + vdt$. As noted by La Manno et al (20), this estimation does not necessarily require the knowledge of $\alpha(t)$ or the assumption of a constant $\alpha(t)$.

Because this adapted RNA velocity theory relies on the assumption that RNA is exported from the nucleus to the cytoplasm, alternative biological processes such as nuclear retention may violate this underlying assumption and therefore making some genes not appropriate for RNA velocity inference. For example, we note that nuclear-retained genes often exhibit negative correlations between nuclear and cytoplasmic RNA counts. Thus, for the RNA velocity analysis, we restricted to genes that exhibit positive correlation between nuclear and cytoplasmic counts thereby removing nuclear-retained genes.

To estimate RNA velocity, we applied the RNA velocity estimation framework as implemented in the velocyto.R package (v0.6) via the gene.relative.velocity.estimates() function (20). We used our nuclear counts matrix and cytoplasmic counts matrix as analogues to the original unspliced and spliced counts matrix in La Manno et al (20). We applied our analysis to the 9050 genes measured using the non-overlapping encoding probe design, and removed 2941 lowly expressed genes detected at fewer than 1 count per cell on average, resulting in 6109 genes. We further removed 1213 genes with a nuclear to cytoplasmic expression correlation of less than 0.05 as well as a minimum fitted slope of less than 0.05, as recommended by the original analysis (20) and reasoned above. This thresholding not only removed nuclear-retained RNAs with a negative correlation but also removed genes that exhibit little correlation between nuclear and

cytoplasmic counts from the RNA velocity analysis. After this thresholding, 4896 positively correlated genes remained to include in our RNA velocity analysis. Although we found the number of genes with little or negative correlation between nuclear and cytoplasmic counts to be relatively small compared to the number of genes exhibiting positive correlation, we do not require this fraction to be small in order to perform RNA velocity analysis. For example, having a number of RNA species permanently retained in the nucleus and not exported to the cytoplasm would not affect our ability to use the genes that do get exported to the cytoplasm to carry out RNA velocity analysis based on the kinetic model for nuclear export. We also note that some nuclear-enriched genes still exhibited significant export into the cytoplasm, and showed positive correlation between nuclear and cytoplasmic counts, and these RNAs were not excluded from the RNA velocity analysis.

To minimize effect due to batch variation, we restricted RNA velocity analyses to the largest batch of cells, i.e. the replicate that contained ~650 cells. For estimating γ, we used cells in the 5th extreme expression percentile and performed k-nearest neighbor pooling with K=30 on cell distances estimated on 30 PCs as recommended by the original RNA velocity analysis (20) and reasoned above.

To visualize RNA velocity on our lower dimensional tSNE embedding as well as PC embedding with velocity arrows (projecting from current states of cells to the predicted future states based on the calculated RNA velocity), we likewise applied the approach previously described and implemented by La Manno et al in the velocyto.R package via the show.velocity.on.embedding.cor() function (20). Given our large number of cells, we visualized a vector field showing the local velocity for a group of cells on a grid rather than each cell individually. We used a square-root scale for projected arrows on a 15x15 grid with otherwise default parameters recommended by the original RNA velocity analysis. To visualize phase portraits for select genes, for each cell among all batches, we plotted the mean nuclear and cytoplasmic expression based on its 50 nearest neighbors estimated on 30 PCs using the gene.relative.velocity.estimates() function provided in the velocyto.R package passing the appropriate show.gene parameter (20).

For these analyses, it is worth keeping in mind that the 6 z-slices that we measured covered most of the cytoplasm, and more than half, but not the full thickness, of the nucleus. However, we do not anticipate this partial coverage to affect the relative RNA velocity or subsequent pseudotime ordering. Consider the true RNA velocity following our model described above is:

$$\frac{dc}{dt} = n - \gamma c$$

With partial coverage of the nucleus or the cytoplasm such that the measured counts are $n_m = An$ and $c_m = Bc$, respectively, we measured the RNA velocity as:

$$\frac{dc_m}{dt} = n_m - \gamma_m c_m$$

Here γ_m is our estimated RNA degradation rate using the steady state of the nuclear and cytoplasmic counts as described above, i.e. $\gamma_m = \frac{n_m}{c_m} = \frac{A}{B}\gamma$. This measured RNA velocity is then:

$$\frac{dc_m}{dt} = An - \frac{A}{B}\gamma Bc = A\frac{dc}{dt}$$

Therefore, our measured RNA velocity is proportional to the true RNA velocity with the same scaling factor *A* for all genes, and therefore the directions of the velocity vectors in the gene space or PC space would not change by the partial coverage of the cell. Given that we measured more than half of nucleus ($0.5 < A < 1$), this common scaling factor for all genes will not have a significant effect on pseudotime ordering of cells, as the prediction of the future cell state from the current cell state based on RNA velocity relies on an arbitrary choice of time delay (*dt*) anyway. However, our analysis does lead to a moderate underestimate of the RNA degradation rate, γ, by a constant *A*/B factor for all genes.

Pseudotime regression analysis

To order cells in pseudotime, we note that the projected velocity arrows fall approximately along a circle in tSNE space. Therefore, we estimated the pseudotime ordering of all cells from all three batches as the $arctan(y, x)/\pi * 180$ where x and y are the zero-centered tSNE coordinates, and performed all pseudotime-related analyses on cells from all three batches.

To identify potentially novel cell-cycle related genes, we fitted a linear regression model of batch-corrected, CPM-normalized gene expression magnitude versus pseudotime to determine the amount of variance that is captured by pseudotime. If this amount of variance is significant (Bonferroni corrected p-value < 0.05), then we consider the gene to have a putative cell-cycle-dependent expression. In this manner, we identified 1654 genes with putative cell-cycle-dependent expression. We further classified genes as known cell-cycle-related genes if they are annotated in the Gene Ontology gene set for mitotic cell cycle (GO:0000278).

To order the 1654 putative cell-cycle-related genes based on their pattern of expression in pseudotime, we fitted a smooth spline curve for batch-corrected, CPM-normalized gene expression magnitude versus pseudotime with a penalty parameter of 5. In addition, we winsorized the top and bottom 0.1% of expression magnitude values in order to remove outliers. We then predicted the point in pseudotime at which the fitted curve is at the maximum. We used this point to sort genes in the order of peak expression in pseudotime. We z-scored the batch corrected, CPM-normalized expression of each gene and visualized the resulting z-score matrix of all 1654 putative cell-cycle-related genes using a heatmap with the cells ordered by pseudotime and genes ordered as described above. As a randomization control, we randomly shuffled the cells in pseudotime, or randomly shuffled cells within cell clusters but kept the order of the clusters, and repeated the process of ordering genes for the same 1654 genes using the above-described randomized pseudotime.

To characterize patterns of nuclear enrichment as a function of pseudotime, we focused on the 1488 genes that we identified as highly significantly nuclear-enriched. We used a smooth-spline to fit the log2(fold-change) of CPM-normalized nuclear versus cytoplasmic expression for post-mitotic cells within G1 to G1/S phase to quantify the minimum and maximum degree of nuclear enrichment post mitosis. We winsorized the top and bottom 0.1% of log2(fold-change) values in order to remove outliers. We defined instantaneous re-establishment of nuclear enrichment as having a minimum fitted log2(fold-change) > 1.5 in G1 phase. We define gradual re-establishment of nuclear enrichment as having a minimum fitted log2(fold-change) < 1.5 and a difference between maximum and minimum degree of nuclear enrichment > 1.5 from post-mitotic to G1/S.

Analysis of spatial heterogeneity of transcriptionally distinct cells

To identify spatial heterogeneity of cells, we focused on the largest batch of cells, i.e. the replicate that contained ~650 cells, and first identified the location of each cell as the average of all pixels corresponding to the cell's body, which we term cell centroid. Then, for each cell, we identify its K = 3 nearest neighbors in space by Euclidean distance. For any give cell cluster, all of these nearest neighbor cells to all cells in this cluster is considered the neighbor cell set, and all other cells are considered the non-neighbor set. To determine if cells of the same transcriptional cell cluster tend to be spatially close to cells of the same cluster, we quantified the cell-cluster identity fraction in the neighbor set and compare to that of the non-neighbor set for each cluster type. For example, for cluster C1, we determined the fraction of C1 cells in the neighbor set for C1 and the fraction of C1 cells in the non-neighbor set for C1, and defined a ratio between the two numbers as the enrichment of C1 cells in C1 neighbor set. We also calculated the enrichment of C2, C3, C4, or C5 cells in C1 neighbor set. We apply a Fisher's exact test to determine if the cell-identity distributions for the neighbor set and non-neighbor set are significantly different. Similar analyses were also performed for C2 – C5 clusters.

To identify genes with significant spatial heterogeneity in expression across cells, we focus on the set of over-dispersed genes that were not significantly associated with pseudotime as determined by our pseudotime regression analysis. Then for these genes, we compute the Moran's I statistic for spatial autocorrelation (21) for their expression in individual cells, using the K = 3 nearest neighbors in space as the binary weight matrix for the spatial autocorrelation. For interpretability of spatial relationships and visualization, we again restricted analyses to the largest batch of cells, which contained ~650 cells. We designate genes as having highly significant spatial heterogeneity in expression if they have Bonferroni-corrected p-values < 1e-10 and as not spatially heterogeneous if they have Bonferroni-corrected p-values > 0.05.

Data and software availability

The data that support the findings of this study are included in the paper and the MERFISH decoding software used in this study is available at https://github.com/zhuanglab/merlin.

SI figures

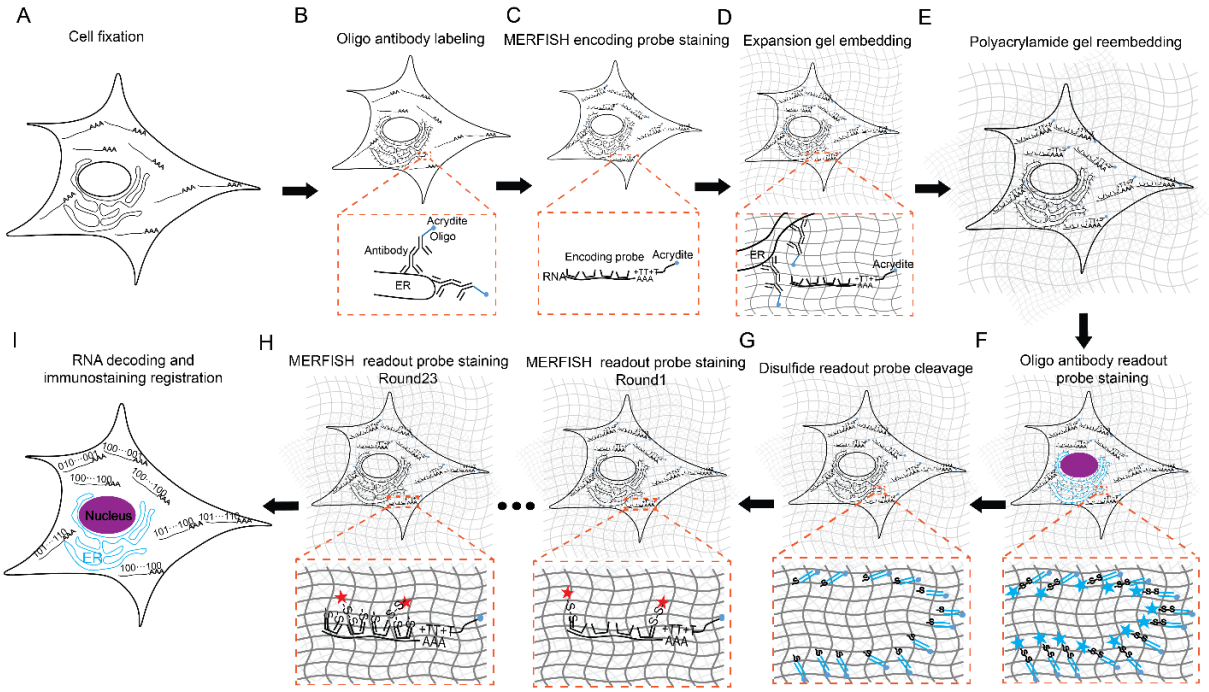

Fig. S1. Schematic illustration of 10,050-gene MERFISH imaging with cellular-structure markers. (A) Cell fixation by 4% PFA. **(B)** Immunostaining of cellular structures with oligonucleotide-conjugated antibodies. The oligonucleotides are modified by acrydite, which can be linked to a polyacrylate gel matrix. **(C)** Staining of cellular RNAs with encoding probes for imprint the barcodes onto the RNAs and acrydite-modified poly(dT) locked nucleic acid (LNA) probe for linking the RNAs to a polyacrylate gel matrix. **(D)** Embedding of the sample in an expandable polyacrylate gel. After gel embedding (i.e. linking cellular RNAs and oligonucleotides on the antibody to the gel), the other cellular components (e.g. proteins and lipids) are digested and extracted. **(E)** Expansion of the gel-embedded sample, and re-embedding the expanded samples in a non-expandable, polyacrylamide gel. **(F)** Hybridization of the sample with readout probes complementary to the oligonucleotides originally linked to the antibody for imaging the cellular structure. **(G)** Cleavage of the disulfide linkage between fluorescent dyes and the readout probes to remove the readout signal. **(H)** Hybridization of the sample with readout probes complementary to the encoding-probe readout sequences that correspond to the first three bits of the MERFISH barcodes, and imaging of these readout probes in three distinct color channels, cleaving the dye off the readout probe as described in (G), hybridization and imaging of the next set of readout probes corresponding to the next three bits, and iterate this process until all bits are measured. **(I)** Decoding the RNA signals and registration of decoded RNAs and the immunofluorescence image of the cellular structure.

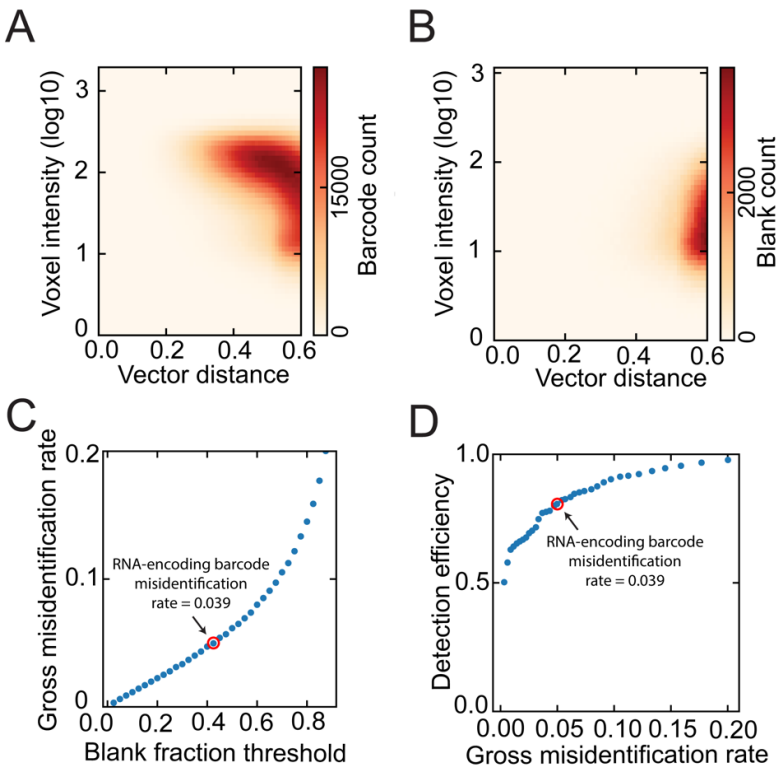

Fig. S2. Adaptive thresholding strategy to remove incorrectly identified barcodes. To identify transcripts in the 69 acquired image stacks corresponding to the 69 bits, each individual voxel is assigned one of the 12,903 valid barcodes that is closest to the normalized 69-dimensional intensity vector for the voxel. Then, adjacent voxels assigned the same barcode are aggregated to create a list of putative transcripts. Barcodes with two-bit errors were removed as described in SI Materials and Methods, "Image processing and decoding" section. Then, for each putative transcript, the mean intensity of the corresponding voxels, referred to as "voxel intensity"; the minimum of the distances of the normalized intensity vectors of the corresponding voxels to the closest valid barcode, referred to as "vector distance"; and the number of voxels, i.e. "voxel number", are calculated. A 3D histogram is constructed for the number of decoded transcripts in the "voxel intensity"-"voxel number"-"vector distance" space. **(A)** A constant "voxel number" slice of the 3D histogram depicting the number of putative transcripts decoded within each "voxel intensity" and "vector distance" bin for all putative transcripts. **(B)** Same as (A) but for the number of putative transcripts that correspond to blank barcodes. We then calculated the fraction of blank barcodes within each 3D histogram bin, and any bin with a high fraction indicates a point with a high chance of barcode misidentification in the "voxel intensity"-"voxel number"-"vector distance" 3D space. We then set a threshold on this fraction and excluded putative transcripts that fall in the "voxel intensity"-"voxel number"-"vector distance" 3D space where this fraction is higher than the threshold. **(C)** The gross barcode misidentification rate as a function of the blank fraction thresholds as described above (filled points). The gross misidentification rate for all barcodes, considering misidentifications into both RNA-encoding and blank control barcodes, is estimated as (the mean count per blank control barcode per cell) / (the mean count per barcode per cell), where the barcodes in the denominator includes both RNA-encoding and blank control barcodes. The blank fraction threshold that achieves a gross

misidentification rate of 5% is selected to construct the final list of detected transcripts in this work (red open circle). We note that an alternative estimation of the misidentification rate is to consider only misidentification into another RNA-encoding barcode as a true mis-identification. This rate can be estimated as (the mean count per blank control barcode per cell) / (the mean count per RNA-encoding barcodes per cell), yielding a misidentification rate of 3.9% for our 10,050-gene measurements. This latter misidentification rate estimate could be considered an experimentally more relevant number as barcodes misidentified as blank control barcodes are typically not used to draw biological conclusion. Additional caveats in misidentification rates are discussed in SI Materials and Methods, "Image processing and decoding" section. **(D)** The detection efficiency as a function of the gross misidentification rate (filled points). The detection efficiency is determined by comparing the number of transcripts detected per cell for each of the 128 common genes in the 10,050-gene experiments versus the 130-gene MERFISH experiments, taking into consideration that the 130-gene MERFISH experiment has a 96% detection efficiency by comparison with smFISH measurements, as we have shown previously. The red circle indicates the point corresponding to 5% gross misidentification rate and 3.9% RNA-encoding barcode misidentification rate.

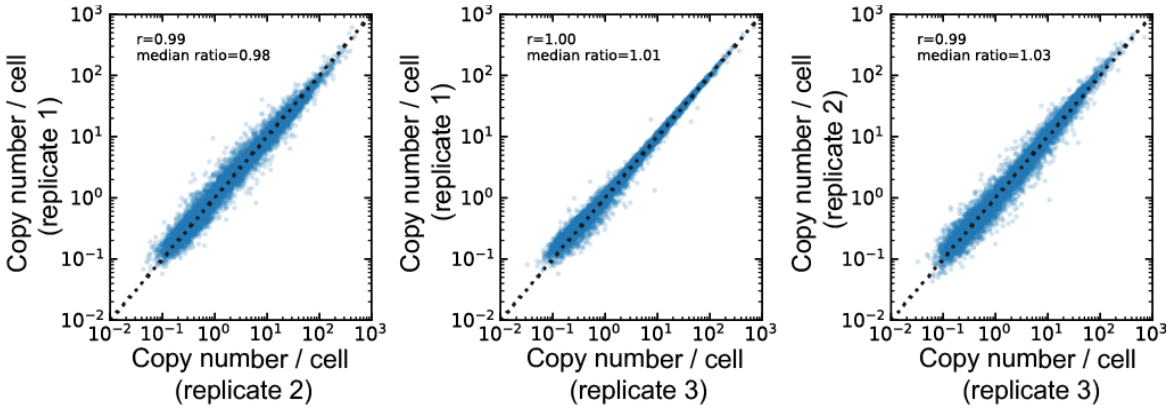

Fig. S3. Correlation between three replicate 10,050-gene MERFISH measurements. Each point depicts the measured abundance of one RNA species between two 10,050 gene experiments. The dotted line indicates equal copy number per cell between the two replicates.

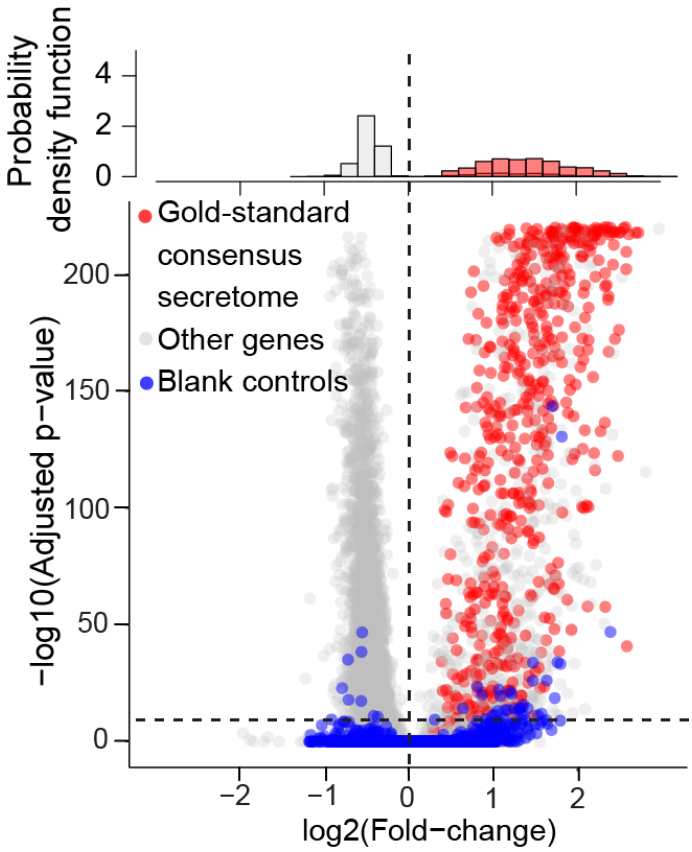

Fig. S4. Identification of RNA species enriched at the ER, including both the 9,050 genes labeled with the non-overlapping encoding probe strategy and the 1,000 genes labeled with the overlapping encoding probe strategy. The fold-change between count-per-million (CPM)-normalized transcript counts localized to the ER versus those localized in the non-ER region of the cytoplasm and the corresponding p-values were calculated for each gene. P-values are determined based on a two-sided pairwise Wilcoxon Rank Sum Test across all cells and adjusted for multiple testing using Bonferroni correction. The bottom panel shows the scatter plot of the p-value versus fold change for each gene. Gold-standard consensus secretome genes, other genes and blank control barcodes are marked in red, grey and blue, respectively. The horizontal dashed line indicates the adjusted p-value = 1e-10 significance threshold and the vertical dashed lines indicates equal transcript counts in the ER and the cytoplasm. The top panel shows the histograms of the fold-changes for the consensus secretome genes (red) and other genes (grey). For the other genes, only the genes with p-value < 1e-10 were included in the histogram. For the 1,000 genes labeled with the overlapping encoding probe strategy, the sensitivity and false discovery rate, as described in the main text for the 9,050 genes, are 81% and 2%, respectively.

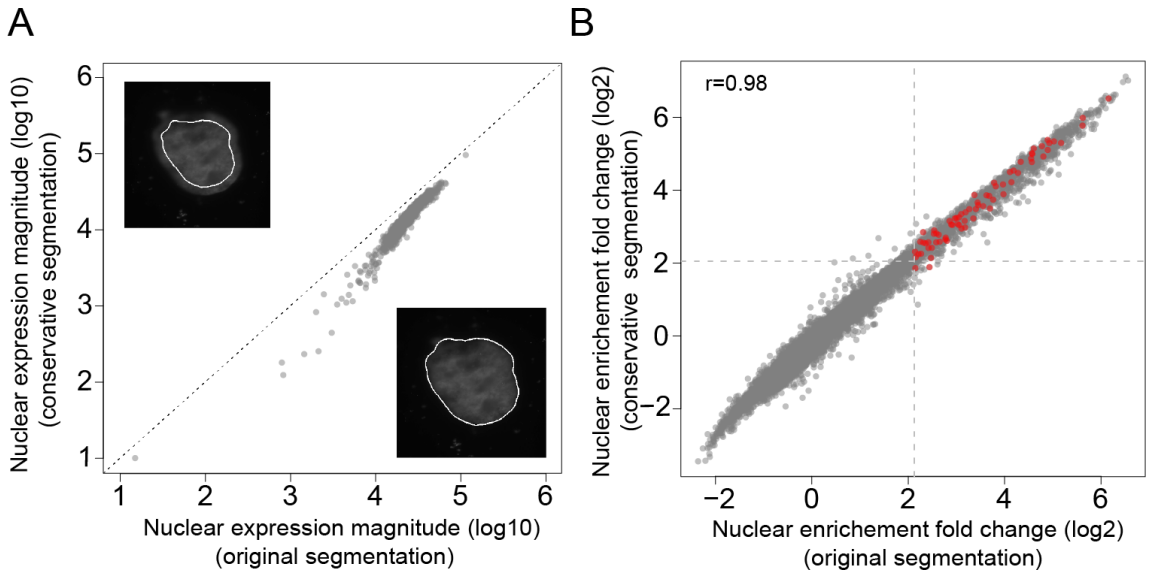

Fig. S5. Nuclear enrichment comparison between the original nuclear segmentation and a more conservative nuclear segmentation. The more conservative segmentation was generated by eroding away ~1 μm on the periphery of the original segmentation (See *SI Materials and Methods*). **(A)** Correlation between expression magnitude for the conservative nuclear segmentation versus the original nuclear segmentation for each RNA species (points). The dotted line indicates equal counts between the two segmentation approaches. As expected, the conservative nuclear segmentation results in fewer transcripts total per nuclei. Insets: DAPI image of an example cell nucleus, shown together with the original nuclear segmentation (lower right) and the more conservative nuclear segmentation (upper left). **(B)** Correlation in nuclear enrichment fold change for regular and conservative segmentation approaches for each gene (points) with significantly nuclear-enriched genes highlighted in red determined using the thresholds of log2(fold-change) > 2 (indicated by the dashed lines) and p-value < 1e-10 using the original segmentation. The Pearson correlation coefficient is 0.98.

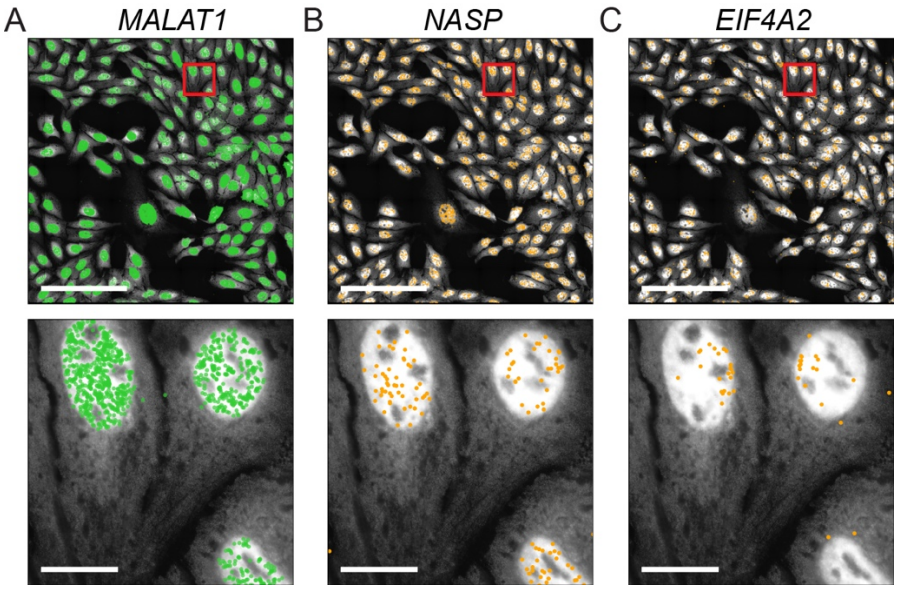

Fig. S6. Spatial distribution of select nuclear-enriched RNAs. (A-C) Spatial distribution of *MALAT1*, a lncRNA (A); *NASP*, an intron-retained RNA (B); and *EIF4A2*, an intron-retained RNA (C) overlaid onto the poly-dT-staining image. Each point in (A-C) indicates the spatial position of a detected transcript. Scale bars: 500 μm (top row), 50 μm (bottom row).

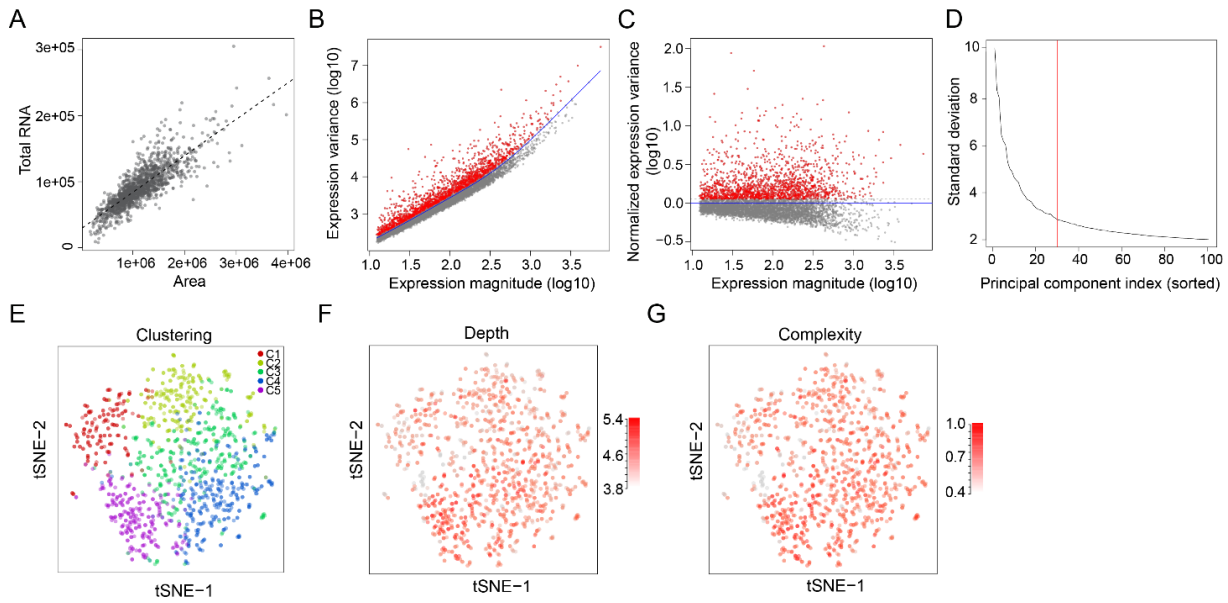

Fig. S7. Single-cell clustering quality controls. (A) Correlation between total RNA counts and cell area. A linear regression fit is plotted as a dotted line. **(B)** Scatter plot of variance of the batch-corrected and CPM-normalized expression magnitude among cells versus the batch-corrected expression magnitude for each of the 9,050 genes labeled with the non-overlapping encoding probe design. A generative additive model (GAM) fit line is noted as a blue line. The 1598 over-dispersed genes are shown in red and the other genes are shown in grey. **(C)** Scatterplot of residual variance after variance normalization by regression to the GAM fit line. Over-dispersed genes are defined as genes exhibiting more variance than expected from expression magnitude based on a χ^2 test using an adjusted p-value = 0.05 threshold (grey dotted line). **(D)** Principal component elbow plot depicting the standard deviation along each of the first 100 principal components sorted by the standard deviation. The red line indicates the cut off used for selecting the number of principal components (30) to use for clustering. **(E-G)** A two-dimensional tSNE embedding of the gene expression profiles for each cell measured by MERFISH. Each point is a cell and is colored by the Louvain cluster annotations of the cell (E), the log10 of the total number of transcripts detected (F), or the percentage of number of unique gene species detected (G).

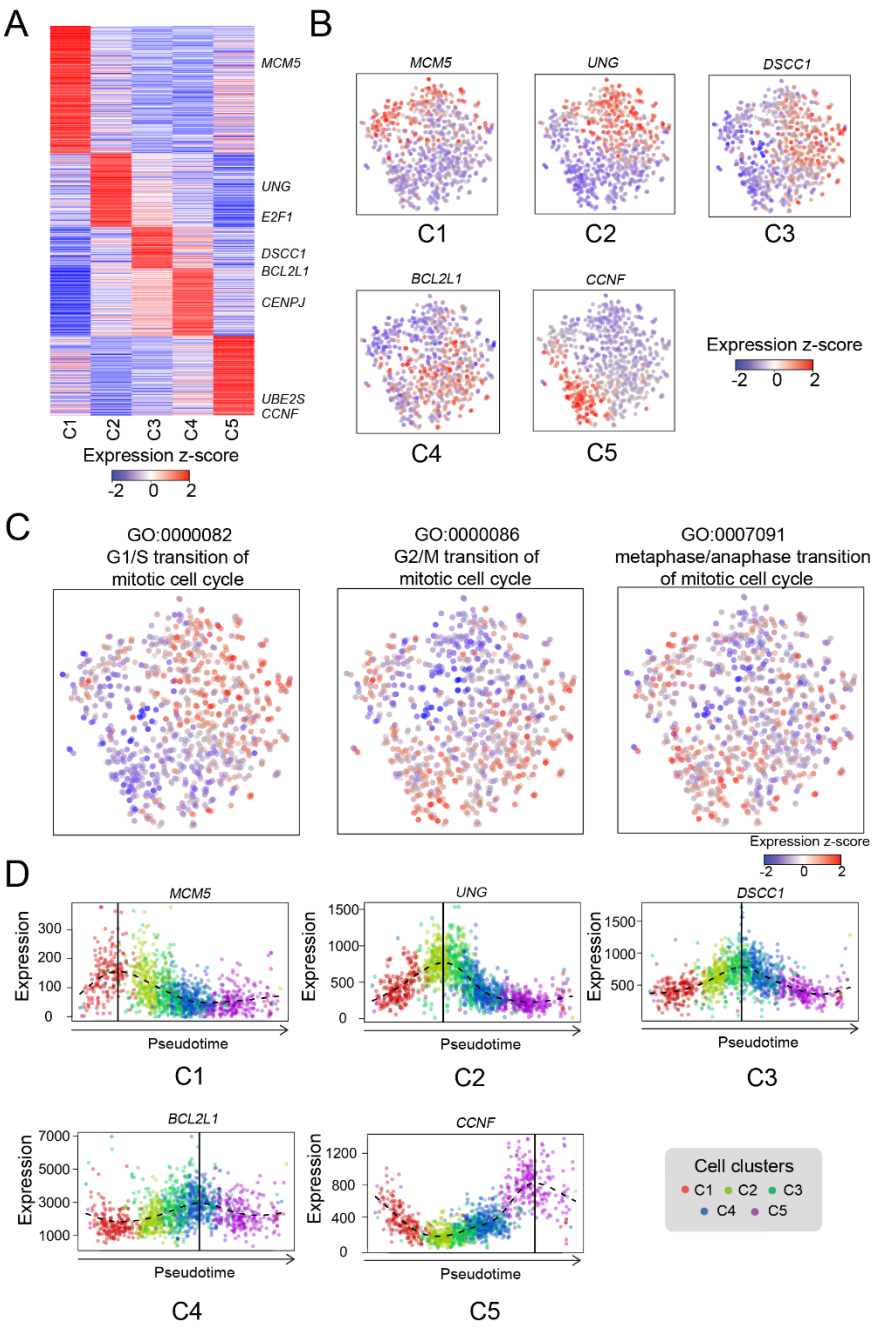

Fig. S8. Differential gene expression characterization of cell clusters and pseudotime ordering of known cell-cycle markers. (A) Heatmap of differentially upregulated genes for each cell cluster. **(B)** Two-dimensional tSNE embeddings colored by CPM-normalized, z-scored expression magnitude of select differentially upregulated genes within each cluster. Each point is a cell. **(C)** Two-dimensional tSNE embeddings colored by averaged z-scores of genes within select Gene Ontology gene sets. Each point is a cell. **(D)** CPM-normalized expression magnitude, with the top and bottom 0.1% winsorized, versus pseudotime for the genes described in (B). Each point represents a cell and is colored by the cluster annotations. A smooth-spline curve is fitted for each gene (dashed line) and the pseudotime corresponding to the maximum expression of the fitted curve is determined (vertical solid line).

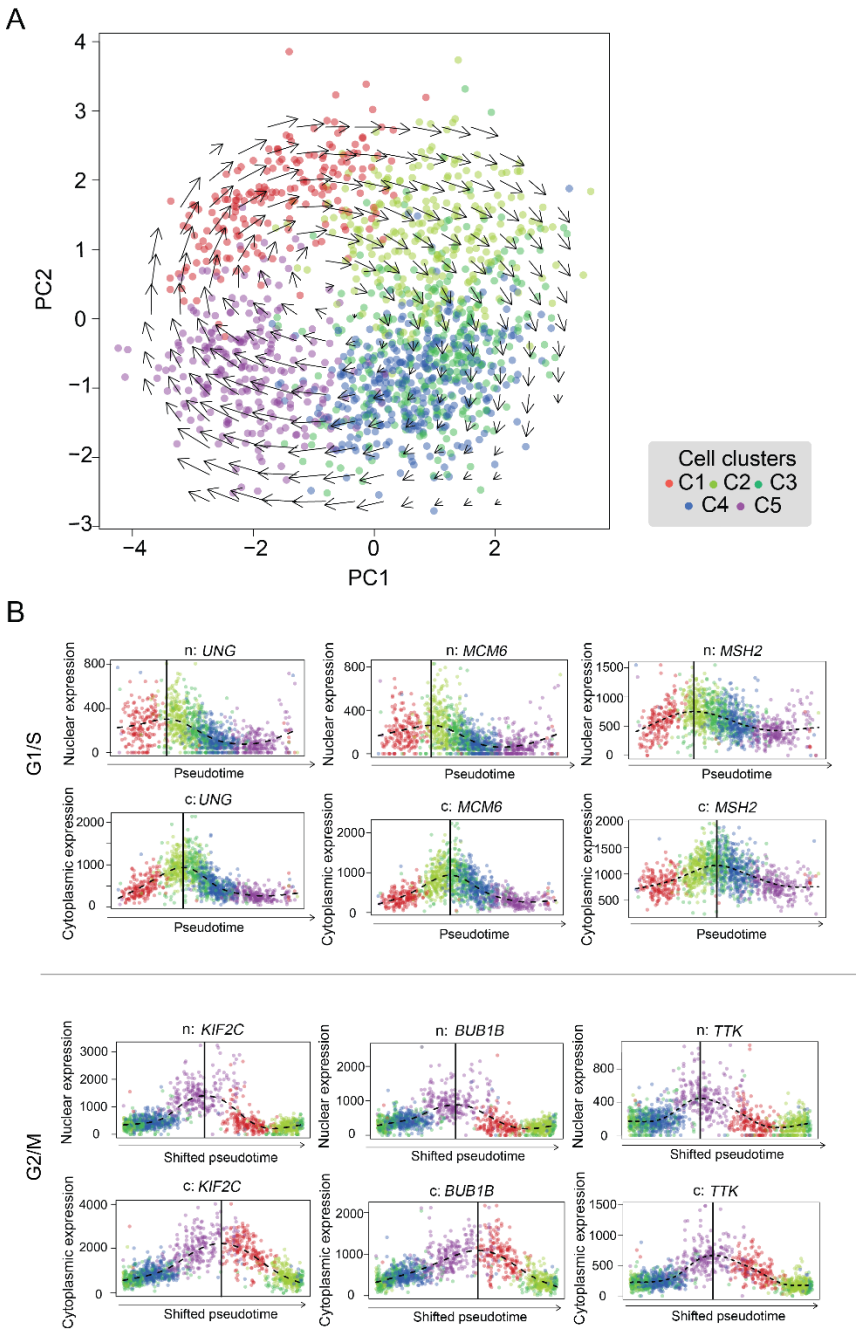

Fig. S9. Visualization of RNA velocity in principle component (PC) space and relationship of nuclear and cytoplasmic expressions of individual genes in pseudotime. (A) A two-dimensional PC embedding of the gene expression profiles for 1,368 cells measured by MERFISH. Each point is a cell and is colored by the Louvain cluster annotations of the cell. Projected velocity arrows show the RNA velocity. **(B)** CPM-normalized expression magnitude in nucleus (n, top panels) or cytoplasm (c, bottom panels), with the top and bottom 0.1% winsorized, versus pseudotime for the genes highly expressed in G1/S (top panel) and G2/M (bottom panel) phases. A smooth-spline curve is fitted for each gene (dashed line) and the pseudotime corresponding to the maximum expression of the fitted curve is determined (vertical solid line). A phase-shifted pseudotime is shown for G2/M genes to allow easier visualization of the expression peak.

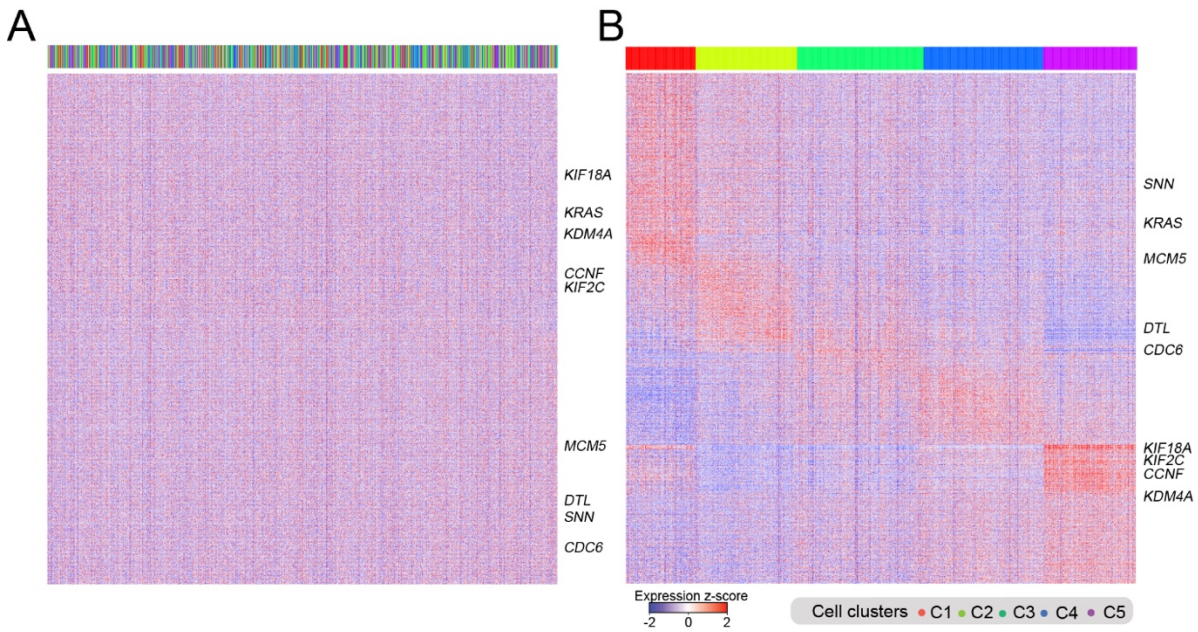

Fig. S10. Heatmap of genes that exhibit cell-cycle-dependent expression after pseudotime randomization. (A) Cells were randomly shuffled in pseudotime and a smooth spline curve was fit to the resulting CPM-normalized expression magnitude versus randomized pseudotime for each gene. Genes (y axis) are ordered based on their maximum expression time points on the fitted curves along the randomized pseudotime axis (x-axis). The color bar on top denotes the original cluster annotation for each cell. Selected cell-cycle genes are labeled. **(B)** Same as (A) but with cells randomly shuffled in pseudotime within individual cell clusters.

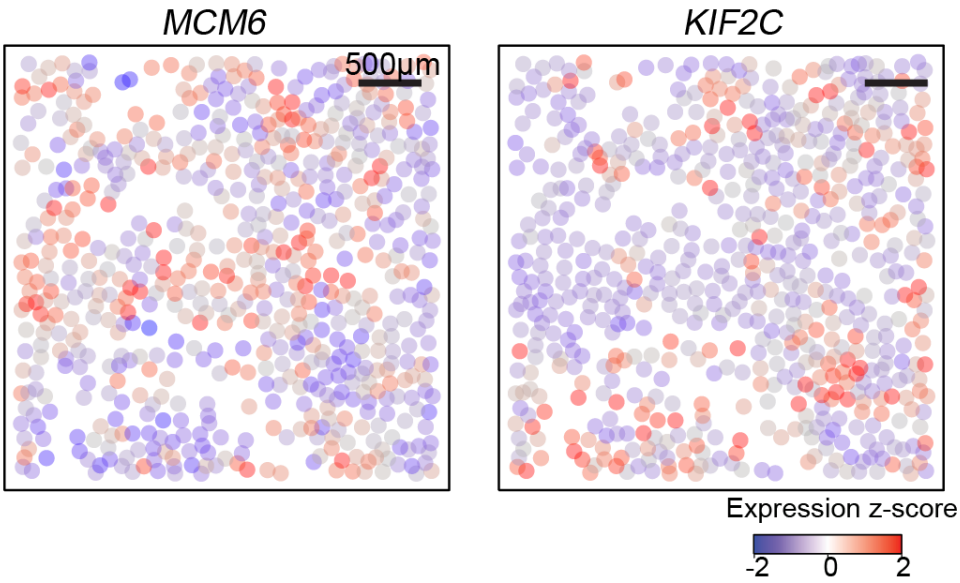

Fig. S11. Spatial expression profile of cell-cycle related genes. Each point indicates the spatial position of a cell and is colored based on the z-scored CPM-normalized expression magnitude for a known cell-cycle marker, *MCM6* or *KIF2C*.

SI Datasets

Dataset S1. MERFISH codebook indicating the 69-bit barcodes assigned to each of the genes and blank controls. The first column is the gene name, the second column is the Ensembl transcript ID, and the following columns indicate the binary values for each of the 69 bits indicated by name of the corresponding readout sequence. Barcodes that were used as blank controls are denoted by a gene name that begins with "Blank-." The 9,050 genes that were labeled with the non-overlapping encoding probe strategy are in rows 2 to 9,051 and the 1,000 genes that were labeled with the overlapping encoding probe strategy are in rows 9,280 to 10,279.

Dataset S2. Encoding probe information. For each encoding probe, the encoding probe sequence and the Ensembl ID of the target transcript are indicated.

Dataset S3. Readout probe information. For each of the bits, the bit number, the readout probe sequence name, the readout probe sequence, and the conjugated dye are indicated.

Dataset S4. RNA enrichment at the endoplasmic reticulum (ER). The 'ER-enriched genes" worksheet shows the log2(Fold-Change) for ER versus non-ER cytoplasm normalized expression, Bonferroni-corrected p-value, mean CPM-normalized ER expression, mean CPM-normalized non-ER cytoplasm expression, and ENSEMBL biotype for each RNA that exhibits statistically highly significant enrichment (log2(fold-change)>0, p-value < 1e-10) at the ER. The "All genes" worksheets shows the same information for all 9,050 genes encoded with the non-overlapping encoding probe strategy and all 2,853 blank controls.

Dataset S5. Rank-based gene set enrichment analysis results for genes that exhibit statistically highly significant enrichment at ER. Only RNA species that exhibit statistically highly significant enrichment (p-value < 1e-10) at the ER are considered. Significantly enriched gene sets (GO terms) among these genes, their adjusted p-values, and enrichment scores are listed.

Dataset S6. RNA enrichment in the nucleus. The 'Nuclear-enriched genes" worksheet shows the log2(Fold-Change) for nuclear versus cytoplasmic normalized expression, Bonferroni-corrected p-value, mean CPM-normalized nuclear expression, mean CPM-normalized cytoplasmic expression, and Ensembl biotype for each RNA that exhibits statistically highly significant enrichment (log2(fold-change) > 2, p-value < 1e-10) in the nucleus. The "All genes" worksheets shows the same information for all 9,050 genes encoded with the non-overlapping encoding probe strategy and 2,853 blank controls.

Dataset S7. Differentially upregulated genes in each cell cluster. Results for each cell cluster are depicted in one worksheet, and the differential expression Bonferroni-corrected p-value and log2(fold-change) for each cluster versus all other clusters are listed for statistically significantly upregulated genes (Bonferroni-corrected p-value < 1e-10) within each cell cluster.

Dataset S8. Genes exhibiting cell-cycle-dependent expression. For each identified gene that exhibits putative cell-cycle-dependent expression as described in the main text, its adjusted p-value in terms of significance for the proportion of variance explained by pseudotime and its pseudotime of maximal, smooth-spline-fitted expression are listed.

Dataset S9. Pseudotime dependence of nuclear enrichment for two classes of genes. The "instantaneous recovery" worksheet lists genes that exhibit instantaneous re-establishment of nuclear enrichment after mitosis and the "gradual recovery" worksheet list genes that exhibit gradual re-establishment of nuclear enrichment after mitosis. For each gene, the minimum and maximum smooth-spline-fitted log2(nuclear enrichment fold-change) through the post-mitotic to G1/S phase, and the mean CPM-normalized gene expression across all cells are listed. Specifically, we define instantaneous re-establishment of nuclear enrichment as having a minimum fitted log2(fold-change) > 1.5. We define gradual re-establishment of nuclear enrichment as having a minimum fitted log2(fold-change) < 1.5 and a difference between maximum and minimum degree of nuclear enrichment > 1.5. Other nuclear-enriched genes that do not satisfy these requirements are not listed.

Dataset S10. Cell-cycle-independent over-dispersed genes. For each gene, the log (CPM-normalized mean expression), log (CPM-normalized variance), residual variance after normalization by regression to the GAM fit line, and log (adjusted p-value) are listed.

Dataset S11. Characterization of spatial heterogeneity by Moran's I for over-dispersed genes that do not exhibit cell-cycle-dependent expression. For each gene, the observed Moran's I statistic (using K = 3 nearest neighbors in space), the expected Moran's I statistic under a null hypothesis of no spatial heterogeneity, the expected Moran's I standard deviation, adjusted p-value, and mean CPM-normalized expression are listed. The genes are classified into three groups: i) exhibiting highly significant spatial heterogeneity (Bonferroni-corrected p-value < 1e-10), ii) exhibiting moderately significant spatial heterogeneity (1e-10 < Bonferroni-corrected p-value < 0.05), or iii) exhibiting insignificant spatial heterogeneity (Bonferroni-corrected p-value> 0.05), and are presented in three separate worksheets.

Dataset S12. Total RNA count matrix. Total RNA copy number for each gene in each cell are listed. Each gene is in a row and each cell is in a column. All 10,050 genes and 2,853 blank controls are listed. The batch numbers and cell numbers are indicated in the first row.

Dataset S13. Matrix of RNA counts colocalized with the ER. The copy numbers of RNA molecules colocalized with the ER for each gene in each cell are listed. Each gene is in a row and each cell is in a column. All 10,050 genes and 2,853 blank controls are listed. The batch numbers and cell numbers are indicated in the first row.

Dataset S14. Matrix of RNA counts colocalized with the nucleus. The copy numbers of RNA molecules colocalized with the nucleus for each gene in each cell are listed. Each gene is in a row and each cell is in a column. All 10,050 genes and 2,853 blank controls are listed. The batch numbers and cell numbers are indicated in the first row.

Dataset S15. Cell positions information. Centroid X and Y coordinates (in the unit of microns) for each cell. The three batches (replicates) are presented in three separate worksheets.

References

1. Hamming RW (1950) Error Detecting and Error Correcting Codes. *Bell System Technical Journal* 29(2):14.
2. Chen KH, Boettiger AN, Moffitt JR, Wang S, & Zhuang X (2015) RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* 348(6233):aaa6090.
3. Moffitt JR*, et al.* (2016) High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc Natl Acad Sci U S A* 113(39):11046-11051.
4. Walz S*, et al.* (2014) Activation and repression by oncogenic MYC shape tumour-specific gene expression profiles. *Nature* 511(7510):483-487.
5. Wang G, Moffitt JR, & Zhuang X (2018) Multiplexed imaging of high-density libraries of RNAs with MERFISH and expansion microscopy. *Sci Rep* 8(1):4847.
6. Moffitt JR*, et al.* (2018) Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* 362(6416):aau5324.
7. Moffitt JR*, et al.* (2016) High-performance multiplexed fluorescence in situ hybridization in culture and tissue with matrix imprinting and clearing. *Proc Natl Acad Sci U S A* 113(50):14456-14461.
8. Chen F*, et al.* (2016) Nanoscale imaging of RNA with expansion microscopy. *Nat Methods* 13(8):679-684.
9. Hua B*, et al.* (2014) An improved surface passivation method for single-molecule studies. *Nat Methods* 11(12):1233-1236.
10. Pau G, Fuchs F, Sklyar O, Boutros M, & Huber W (2010) EBImage--an R package for image processing with applications to cellular phenotypes. *Bioinformatics* 26(7):979-981.
11. Jones TR, Carpenter A, & Golland P (2005) Voronoi-based segmentation of cells on image manifolds. *Lect Notes Comput Sc* 3765:535-543.
12. Kaewsapsak P, Shechner DM, Mallard W, Rinn JL, & Ting AY (2017) Live-cell mapping of organelle-associated RNAs via proximity biotinylation combined with protein-RNA crosslinking. *Elife* 6:e29224.
13. Subramanian A*, et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102(43):15545-15550.
14. Fan J (2019) *Differential Pathway Analysis , Computational methods for single-cell data analysis* (Springer Science+Business Media, New York, NY) p pages cm.
15. Storey JD (2002) A direct approach to false discovery rates. *J Roy Stat Soc B* 64:479-498.
16. Hicks SC, Townes FW, Teng M, & Irizarry RA (2018) Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* 19(4):562-578.
17. Johnson WE, Li C, & Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8(1):118-127.
18. van der Maaten L & Hinton G (2008) Visualizing Data using t-SNE. *J Mach Learn Res* 9:2579-2605.
19. Blondel VD, Guillaume JL, Lambiotte R, & Lefebvre E (2008) Fast unfolding of communities in large networks. *J Stat Mech-Theory E* 2008(10):P10008.
20. La Manno G*, et al.* (2018) RNA velocity of single cells. *Nature* 560(7719):494-498.
21. Moran PA (1950) Notes on continuous stochastic phenomena. *Biometrika* 37(1-2):17-23.