Spatial transcriptome profiling by MERFISH reveals subcellular 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], and 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; and [c]Department of Physics, Harvard University, Cambridge, MA 02138

The expression profiles and spatial distributions of RNAs regulate many cellular functions. Image-based transcriptomic approaches provide powerful means to measure both expression and spatial information of RNAs in individual cells within their native environment. Among these approaches, multiplexed error-robust fluorescence in situ hybridization (MERFISH) has achieved spatially resolved RNA quantification at transcriptome scale by massively multiplexing single-molecule FISH measurements. Here, we increased the gene throughput of MERFISH and demonstrated simultaneous measurements of RNA transcripts from ∼10,000 genes in individual cells with ∼80% detection efficiency and ∼4% misidentification rate. We combined MERFISH with cellular structure imaging to determine subcellular compartmentalization of RNAs. We validated this approach by showing enrichment of secretome transcripts at the endoplasmic reticulum, and further revealed enrichment of long noncoding RNAs, RNAs with retained introns, and a subgroup of protein-coding mRNAs in the cell nucleus. Leveraging spatially resolved RNA profiling, we developed an approach to determine RNA velocity in situ using the balance of nuclear versus cytoplasmic RNA counts. We applied this approach to infer pseudotime ordering of cells and identified cells at different cell-cycle states, revealing ∼1,600 genes with putative cell cycle-dependent expression and a gradual transcription profile change as cells progress through cell-cycle stages. Our analysis further revealed cell cycle-dependent and cell cycle-independent spatial heterogeneity of transcriptionally distinct cells. We envision that the ability to perform spatially resolved, genome-wide RNA profiling with high detection efficiency and accuracy by MERFISH could help address a wide array of questions ranging from the regulation of gene expression in cells to the development of cell fate and organization in tissues.

single-cell transcriptomics | spatial transcriptomics | fluorescence in situ hybridization | MERFISH | RNA velocity

Single-cell transcriptome imaging allows quantitative measurements of both the gene expression profiles of individual, spatially localized cells and the intracellular distributions of the transcripts. Such information can help answer a wide range of biological questions. At the subcellular level, compartmental distributions of RNAs within cells provide an efficient way to produce proteins at the location of function and in response to local stimuli, such as at synapses in neurons (1). In addition, spatial organization is also important for the function of noncoding RNAs, such as in chromatin structure organization and gene expression regulation (2). At the tissue level, cell-specific gene expression defines cell types and cell states, the spatial organization of which is tightly coupled to both the development and function of normal tissues and to the pathogenesis and prognosis of tissue pathology from patients (3, 4). Therefore, the ability to perform spatially resolved, single-cell transcriptome profiling will provide important insight into many biological systems.

Various spatially resolved transcriptomics approaches have been developed, including methods based on multiplexed single-

molecule fluorescence in situ hybridization (smFISH) (5–7) and in situ sequencing (8–10), which provide single-cell resolution, as well as methods based on spatially resolved RNA capture followed by sequencing (11, 12). Among these techniques, multiplexed error-robust FISH (MERFISH) enables transcriptome-scale RNA imaging of individual cells by introducing the strategies of using error-robust barcodes to encode individual RNA species, physically imprinting the barcodes on RNAs using combinatorial oligonucleotide labeling, and then measuring these barcodes through sequential rounds of imaging (6). In each round, RNAs are imaged by FISH at the single-molecule level (13, 14), and the use of error-robust barcodes allows errors accumulated through multiple imaging rounds to be detected and/or corrected (6). MERFISH has previously achieved single-cell RNA profiling in both cultured cells and brain tissues with high detection efficiency (6, 15, 16) and demonstrated transcriptome-scale imaging of as many as ∼1,000 RNA species in single cells (6).

Significance

The spatial organization of RNAs within cells and spatial patterning of cells within tissues play crucial roles in many biological processes. Here, we demonstrate that multiplexed error-robust FISH (MERFISH) can achieve near-genome-wide, spatially resolved RNA profiling of individual cells with high accuracy and high detection efficiency. Using this approach, we identified RNA species enriched in different subcellular compartments, observed transcriptionally distinct cell states corresponding to different cell-cycle phases, and revealed spatial patterning of transcriptionally distinct cells. Spatially resolved transcriptome quantification within cells further enabled RNA velocity and pseudotime analysis, which revealed numerous genes with cell cycle-dependent expression. We anticipate that spatially resolved transcriptome analysis will advance our understanding of the interplay between gene regulation and spatial context in biological systems.

Here, we increased the gene throughput of MERFISH by 1 order of magnitude, demonstrating simultaneous imaging of RNA transcripts from ∼10,000 genes in single cells with ∼80% detection efficiency and ∼4% misidentification rate. We achieved this high level of multiplexity by using a 69-bit, error-correcting encoding scheme with 23 hybridization rounds and 3-color imaging per round, in combination with expansion microscopy (17) and an improved decoding algorithm to allow measurements of RNA molecules at high density. Combined with imaging of cellular compartments, such as the endoplasmic reticulum (ER) and nucleus, MERFISH allowed near-genome-wide, quantitative characterization of subcellular localizations of RNAs. Using the known localization of secretome RNAs to the ER, we demonstrated both high sensitivity and low false discovery rate in determining the spatial compartmentalization of RNAs. We further observed that long noncoding RNAs (lncRNAs), RNAs with intron retention, and a subgroup of protein-coding mRNAs were highly enriched in the nucleus. Moreover, using the spatially resolved, transcriptome quantification of individual cells and leveraging the knowledge that mRNAs are produced in the nucleus before export to the cytoplasm, we developed an approach to determine RNA velocity in situ using the balance of nuclear versus cytoplasmic RNA counts, which allowed us to order cells in pseudotime, identify cells in distinct cell-cycle states, and observe ∼1,600 genes with putative cell cycle-dependent expression. Finally, our spatially resolved analysis of transcriptionally distinct cells identified both cell cycle-dependent and cell cycle-independent spatial heterogeneity in cell populations.

Results

Spatially Resolved RNA Quantification of 10,050 Genes with High Detection Efficiency and Accuracy.
MERFISH allows spatially resolved gene expression profiling of single cells in intact biological samples by assigning error-robust barcodes to individual RNA species followed by barcode detection through sequential imaging. In our original, most adopted implementation of MERFISH (6), RNA transcripts are stained with a mixture of encoding probes, each containing a target sequence complementary to 1 of the target transcripts and 1 or multiple readout sequences drawn from a collection of readout sequences that each represents a specific bit of the barcodes. The collection of readout sequences bound to an RNA then determines which bit of its barcode reads "1," and hence the barcode identity. Next, fluorescently labeled readout probes complementary to each readout sequence are sequentially hybridized to the sample, allowing the readout sequences present on each RNA transcript, and hence the identity of its barcode, to be determined. To allow error detection and/or correction, we require that each barcode has at least 2 bits that are different from any other valid barcode in the codebook (i.e., Hamming distance [HD] $\geq$ 2). An HD of 2 allows error detection and an HD of 3 or greater additionally allows error correction. For example, we have previously used a 16-bit, HD4 (HD = 4) code to simultaneously detect 140 genes with a detection efficiency of nearly 100% and a 14-bit HD2 code to simultaneously detect 1,000 genes with a lower detection efficiency in individual cells (6).
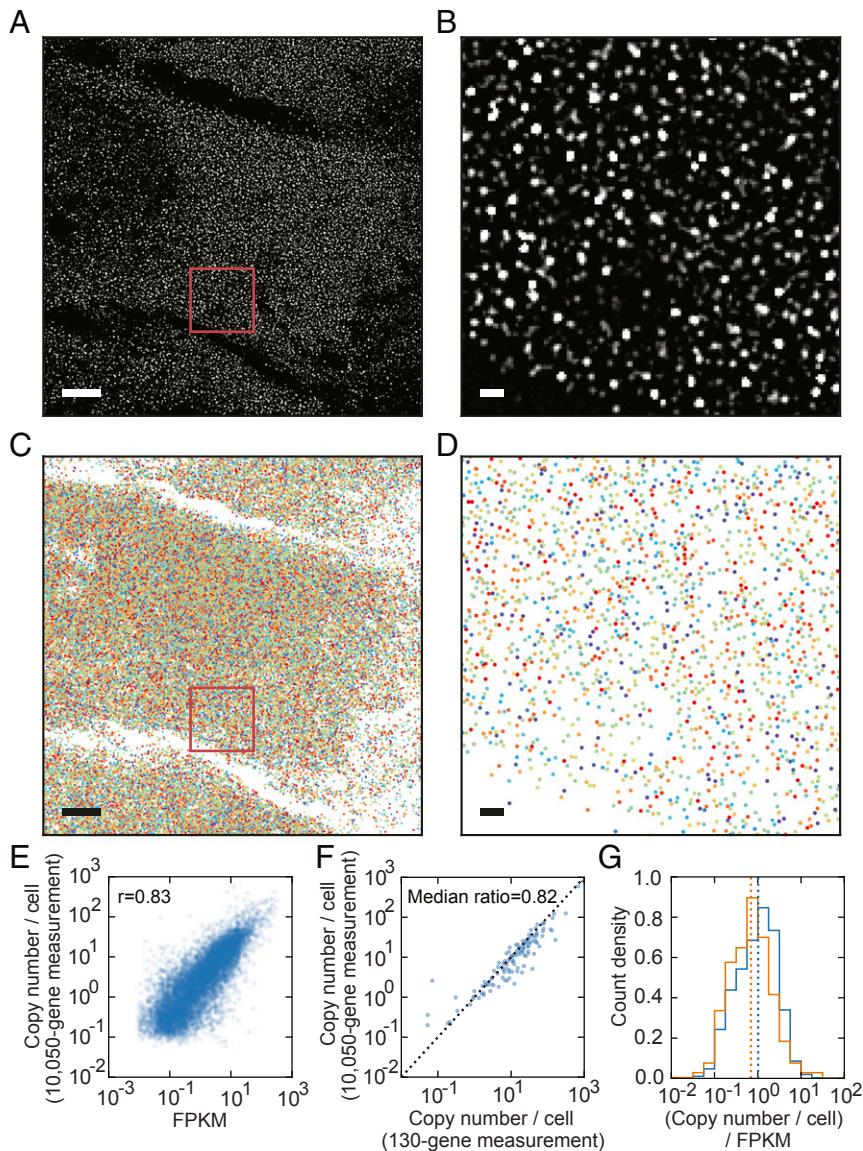
Here, to enable simultaneous imaging of ∼10,000 genes, we constructed a 69-bit HD4 code with a Hamming weight (HW) of 4 (HW = number of "1" bits per barcode), which yielded a total of 12,903 barcodes (Dataset S1). We then randomly selected 10,050 of these barcodes to uniquely encode 10,050 genes, leaving the remaining 2,853 barcodes unassigned to serve as blank controls for misidentification quantification (Dataset S1). To physically imprint the assigned barcodes onto their corresponding target RNAs, we designed a pool of encoding probes, each containing a 30-nt target sequence and 3 readout sequences drawn from the 69 unique readout sequences corresponding to the 69 bits (Dataset S2), such that the collection of readout sequences bound to each

target RNA corresponds to the 4 bits that read "1" in the barcode assigned to this RNA. Among the 10,050 genes, 9,050 genes were 1,440 nt or longer, and we designed up to 48 encoding probes with nonoverlapping target sequences for each of these genes. For the additional 1,000 genes that were either shorter than 1,440 nt or did not have enough targetable regions to accommodate 48 nonoverlapping encoding probes, we used an overlapping encoding-probe design that allowed the encoding probes to target overlapping regions on the target RNA (20-nt overlap between adjacent probes), a strategy that we have previously shown not to substantially reduce signal detected from individual RNA molecules, presumably because individual RNA molecules are typically not simultaneously bound by all cognate encoding probes (16, 18). This latter labeling strategy allowed us to detect RNAs as short as 500 nt in length using 48 encoding probes per RNA.

A challenge associated with imaging such a large number of genes is the high density of RNA molecules, which prevents neighboring RNA molecules from being resolved from each other. To overcome this challenge, we used expansion microscopy (17) to physically separate the transcripts while anchoring them to an expandable gel by hybridizing acrydite-modified poly-dT probes to the poly-A tails of the RNAs (*SI Appendix*, Fig. S1) (18). Additionally, to allow simultaneous imaging of cellular structures in order to determine the subcellular localization of RNAs, we stained the cellular structure with antibodies conjugated to acrydite-modified oligonucleotides and treated these oligonucleotides as the readout sequences during MERFISH imaging (*SI Appendix*, Fig. S1). After antibody and encoding-probe staining, gel embedding, and expansion, we iteratively hybridized the expanded sample with fluorescent readout probes (Dataset S3) complementary to the readout sequences, and imaged 3 readout probes at a time using 3-color imaging (*SI Appendix*, Fig. S1). After 23 rounds of hybridization to image all 69 bits of the barcodes, we decoded the images to identify the RNA transcripts (Fig. 1 *A–D*) (*SI Appendix, Materials and Methods*).

We performed the 10,050-gene MERFISH experiments on human osteosarcoma (U-2 OS) cells. On average, we identified ∼92,000 ± 32,000 (mean ± SD) transcripts per cell. Only 1.1% of the detected barcodes per cell were blank control barcodes, which suggests a barcode misidentification rate of ∼4% (*SI Appendix, Materials and Methods* and Fig. S2). The average copy number per cell detected for individual RNA species by MERFISH was highly correlated with the RNA abundance measured by bulk RNA sequencing (Pearson correlation coefficient $r = 0.83$) (Fig. 1E). Furthermore, the copy number per cell of each RNA species was highly reproducible between replicate MERFISH experiments (Pearson correlation coefficient $r = 0.99$ to 1.00; median copy-number ratio, 0.98 to 1.03) (*SI Appendix*, Fig. S3).
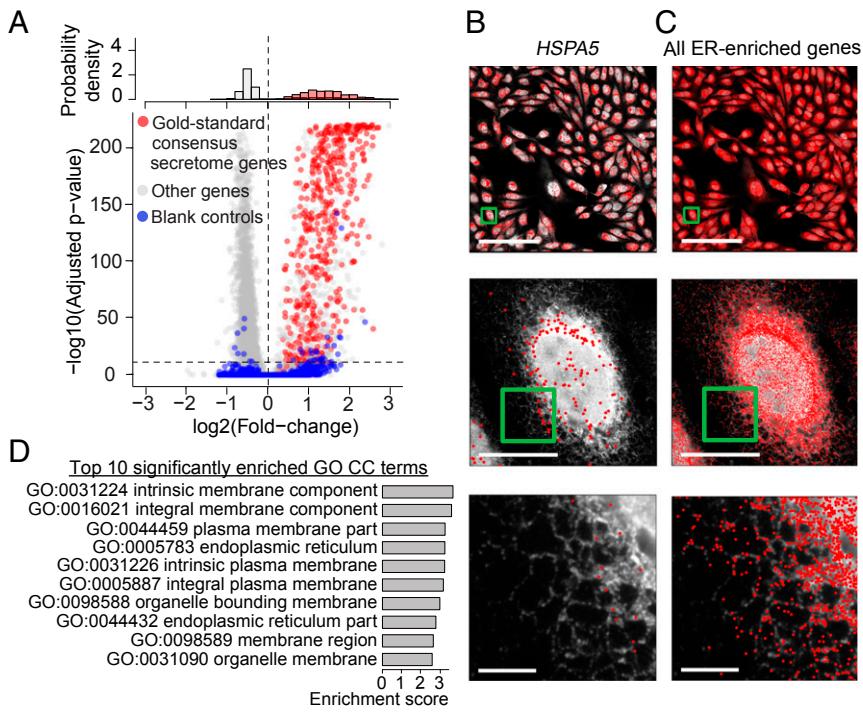
To estimate the detection efficiency of our 10,050-gene measurements, we compared the transcript counts with the results from 130-gene MERFISH measurements using the 16-bit HD4 HW4 code with 92 encoding probes per gene, which we have previously established to have a detection efficiency of 96% as compared with smFISH measurements (15, 19). Of these 130 genes, 128 were included in our 10,050-gene measurements. To enable accurate comparison, we performed the 130- and 10,050-gene measurements using the same sample preparation and expansion protocols, and imaged the same number of z slices using the same setup, except that the 130-gene measurements were performed using the 16-bit HD4 HW4 code with 92 encoding probes per gene and hence required only 6 rounds of hybridization. The median ratio of transcript counts per cell for these 128 genes determined in our 10,050-gene measurements to the numbers determined in our 130-gene measurements was 82% (Fig. 1F), thus indicating a detection efficiency of 79% (0.82 × 96%) for our 10,050-gene measurement. All of these 128 genes were longer than 1,440 nt and hence the detection efficiency determined was for the 9,050 genes labeled with the nonoverlapping

Fig. 1. MERFISH imaging of 10,050 genes in individual U-2 OS cells. (*A*) A high-pass-filtered, single-*z* slice, single-bit image of a U-2 OS sample stained with encoding probes targeting 10,050 RNA species, imprinting a 69-bit binary barcode onto each RNA species, and with an Alexa 750-labeled readout probe that detects 1 of the 69 bits of the barcodes. (Scale bar: 10 μm.) (*B*) A zoomed-in image of the region marked with the red box in *A*. (Scale bar: 1 μm.) (*C*) All identified RNA molecules (colored markers) detected from all 6 imaged *z* slices in the region depicted in *A*. (Scale bar: 10 μm.) (*D*) A zoomed-in image of all identified RNA molecules detected from all 6 imaged *z* slices in the region marked with the red box in *C*. (Scale bar: 1 μm.) In *C* and *D*, different colored dots mark RNA molecules with distinct barcodes. (*E*) The copy number per cell of RNA transcripts of each gene determined by MERFISH versus the fragments per kilobase of transcript per million mapped reads determined by bulk RNA sequencing for the 10,050 measured genes. The Pearson correlation coefficient is 0.83. (*F*) The copy number per cell detected in the 10,050-gene measurements vs. the copy number per cell in the 130-gene measurements for the 128 genes that are commonly present in both experiments. The median value of the ratios of the 10,050-gene measurements to the 130-gene measurements for each of the 128 common genes is 0.82. (*G*) Normalized histograms of the FPKM-normalized copy number per cell for the 9,050 genes that were labeled using the nonoverlapping encoding-probe design (blue) and for the 1,000 genes that were labeled using the overlapping encoding-probe design (orange). The medians of the FPKM-normalized copy number per cell for the 2 groups of genes are indicated by the blue and orange dotted lines, respectively. A high-resolution version of Fig. 1 can be found on Zenodo (DOI: 10.5281/zenodo.3380442).

encoding-probe strategy. To determine the impact of the overlapping encoding-probe design on the detection efficiency for the remaining 1,000 genes, we compared the measured transcript abundance per gene (normalized by the fragments per kilobase of transcript per million mapped reads [FPKMs] determined from bulk sequencing) between the two encoding-probe design strategies (Fig. 1*G*). We found the median FPKM-normalized copy number per cell for the 1,000 genes detected with the overlapping encoding-probe design to be 68% of that for the 9,050 genes detected with the nonoverlapping probe design.

These experiments demonstrate that MERFISH allows spatially resolved expression profiling of ∼10,000 genes in individual cells with a detection efficiency that is substantially higher than the typical detection efficiency (5 to 40%) of single-cell RNA sequencing (scRNA-seq) measurements (4). The moderate reduction in detection efficiency compared with our 130-gene measurements is likely caused by the decrease in the number of encoding probes used, which led to reduced signals from individual RNA molecules, and the dramatic increase in total RNA counts per cell, which could lead to a small fraction of

Fig. 2. Identification of RNAs enriched at the endoplasmic reticulum. (*A*) Quantification of the ER enrichment for each RNA species. The fold change between count-per-million–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. In cpm normalization, the abundance of each RNA species was divided by the abundance of all RNA species in the corresponding cellular compartment and multiplied by a million for each cell. *P* values are determined based on a 2-sided pairwise Wilcoxon rank-sum test across all cells and adjusted for multiple testing using Bonferroni correction. (*A, Bottom*) The scatterplot 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, gray, and blue, respectively. The horizontal dashed line indicates the *P* = 1e-10 significance threshold and the vertical dashed line indicates log2(fold change) = 0. (*A, Top*) Histograms of the fold changes for the gold-standard consensus secretome genes (red) and other genes (gray). For the other genes, only those with *P* < 1e-10 are shown in the histogram. (*B*) Spatial distribution of an example gene, *HSPA5*, that is enriched at the ER, overlaid on the ER image. (*C*) Spatial distribution of all genes identified as highly significantly enriched (log2[fold change] > 0, *P* < 1e-10) at the ER overlaid on the ER image. Each red point in *B* and *C* represents the position of a transcript detected by MERFISH from all 6 imaged *z* slices. The ER images in *B* and *C* are from 1 of the 6 imaged *z* slices. In *B* and *C*, *Middle* and *Bottom* panels are zoomed-in images of the boxed regions in the *Top* and *Middle* panels, respectively. (Scale bars: *B* and *C, Top*, 500 μm; *B* and *C, Middle*, 50 μm; *B* and *C, Bottom*, 10 μm.) (*D*) Top 10 significantly enriched (FDR < 0.05) GO cellular component terms among the genes highly significantly enriched at the ER (as described above), ordered by GO term enrichment score.

unresolved molecules despite our use of expansion microscopy and a greater number of hybridization rounds to reduce the molecular density per round.
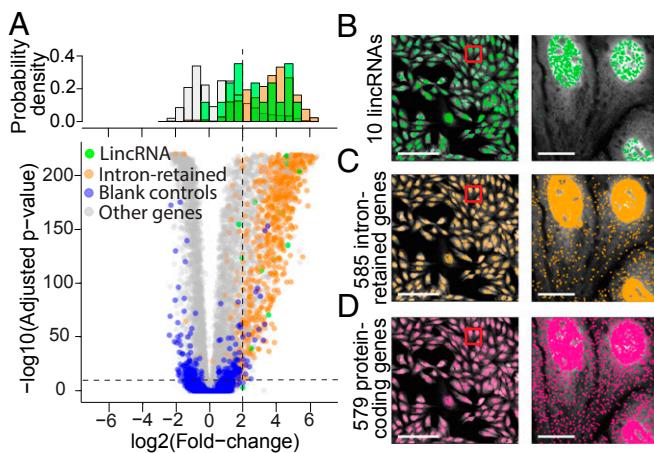
Identification of RNAs Enriched at the Endoplasmic Reticulum. Next, to demonstrate the ability to determine the subcellular compartmentalization of the transcriptome with MERFISH, we first sought to identify which RNAs were enriched at the ER. The translation of mRNAs that encode secreted, glycosylated, and/or transmembrane proteins, collectively termed the secretome, has been shown to take place on the rough ER (20). Moreover, the association of RNAs with the ER has been studied by several different methods, including cell fractionation with ribosome profiling (21), proximity-specific ribosome profiling (22), APEX-RIP (23), and APEX-seq (24). Thus, the ER provides an ideal test case for such a proof-of-concept demonstration.

To this end, we combined MERFISH with KDEL immunolabeling, which marks the ER, DAPI staining, which marks the nucleus, as well as poly-dT staining, which marks total RNAs with poly-A tails. We then used these signals to computationally segment the cells, as well as the ER and nucleus in each cell, and quantified the number of RNA molecules colocalized with these compartments for each gene in each cell (*SI Appendix, Materials and Methods*).

To identify genes enriched at the ER, we performed 2-sided pairwise Wilcoxon rank-sum tests on count-per-million (cpm) normalized RNA counts localized to the ER versus those localized

in the non-ER region of the cytoplasm in individual cells (*SI Appendix, Materials and Methods*) for each gene and for the blank controls (Fig. 2*A* and Dataset S4). We first restricted our analysis to the 9,050 genes detected with a nonoverlapping encoding-probe design. We identified 1,006 genes as highly significantly enriched at the ER (log2[fold change between ER and non-ER cytoplasm expression] > 0, Bonferroni-corrected *P* < 1e-10) (Dataset S4). Visual inspection indeed confirmed preferential localization of these RNAs to the ER (Fig. 2 *B* and *C*). Next, to characterize known gene sets overrepresented by these genes, we performed gene set enrichment analysis (25). For interpretability, we restricted analysis to gene sets within the cellular component (CC) class of gene ontology (GO) terms (Dataset S5). As expected, we identified canonical ER, ER parts, and membrane-associated terms among the most significantly enriched GO CC terms (Fig. 2*D*). We note that a stringent *P* value threshold was used here to increase the confidence of detecting ER-enriched genes, although some true ER-enriched genes may be excluded by such a stringent criterion and a more inclusive identification of ER-enriched genes could be obtained with a less stringent *P* value threshold using the all-gene data provided in Dataset S4.

To further assess the accuracy of our approach, we compared our results with ER-associated mRNAs identified by computationally derived databases, Phobius (26) and SignalP (27), as well as by orthogonal experimental approaches, proximity-specific ribosome profiling (22) and APEX-RIP (23). Between the two ribosome profiling approaches, cellular fractionation with ribosome

Fig. 3. Identification of RNAs enriched in the nucleus. (*A*) Quantification of nuclear enrichment for each RNA species. The fold change between cpm-normalized transcript counts in the nucleus versus those in the cytoplasm and the corresponding *P* values were calculated for each gene. *P* values are determined based on a 2-sided pairwise Wilcoxon rank-sum test across all cells and adjusted for multiple testing using Bonferroni correction. (*A, Bottom*) The scatterplot of the *P* value versus fold change for each gene. LincRNA, RNA with retained introns, other genes, and blank control barcodes are marked in green, orange, gray, and blue, respectively. The horizontal dashed line indicates the $P = 1e-10$ significance threshold and the vertical dashed line indicates the $\log2(\text{fold change}) = 2$ threshold used to define genes with significant nuclear enrichment. (*A, Top*) Histograms of the fold changes for lincRNA (green), RNA with retained introns (orange), and other genes (gray). Only genes with $P < 1e-10$ are shown in these histograms. (*B–D*) Spatial distribution of all identified lincRNAs (*B*), intron-retained RNAs (*C*), and protein-coding RNAs (*D*) that are highly significantly enriched in the nucleus ($\log2[\text{fold change}] > 2$; $P < 1e-10$), overlaid onto the poly-dT staining image. Each colored point in *B–D* indicates the spatial position of a detected transcript. In *B–D*, *Right* panels are zoomed-in images of the boxed regions in the *Left* panels. (Scale bars: *B–D*, *Left*, 500 μm; *B–D*, *Right*, 50 μm.)

profiling was previously noted to be noisier than proximity-specific ribosome profiling (23); between the two APEX approaches, the annotated APEX-RIP datasets provides a more complete coverage of secretome genes (23, 24), and thus we chose to compare our data with proximity-specific ribosome profiling and APEX-RIP results. Both APEX-RIP and proximity-specific ribosome profiling were performed on HEK293T cells, while our MERFISH measurements were performed on U-2 OS cells. We thus restricted our comparisons to genes expressed at greater than 1 FPKM (by bulk RNA sequencing) in both U-2 OS and HEK293T cells, resulting in 6,268 commonly expressed genes that were included in our MERFISH measurements. Among these, we defined a consensus set of gold-standard ER-enriched genes as genes identified to be ER-associated by more than 3 of the 5 approaches (Phobius, SignalP, APEX-RIP, proximity-specific ribosome profiling, and MERFISH), resulting in a set of 590 gold-standard ER-enriched genes. Of these, MERFISH correctly identified 507 as being significantly ER-enriched, while APEX-RIP identified 588 and proximity-specific ribosome profiling identified 469, resulting in sensitivity estimates of 86, >99, and 80% for the 3 approaches, respectively.

In addition, we interpreted genes uniquely identified by a single approach with no support by any other approach as potential false positives. Based on this, MERFISH identified 45 such false positives, while APEX-RIP identified 47, and proximity-specific ribosome profiling identified only 4. Here, we used relatively stringent criteria to identify both the gold-standard, consensus ER-enriched gene set and false positives in order to increase our confidence of their classification (as being ER-enriched or non–ER-enriched). Although these criteria may have excluded some true and false positives, impacting the absolute estimates of
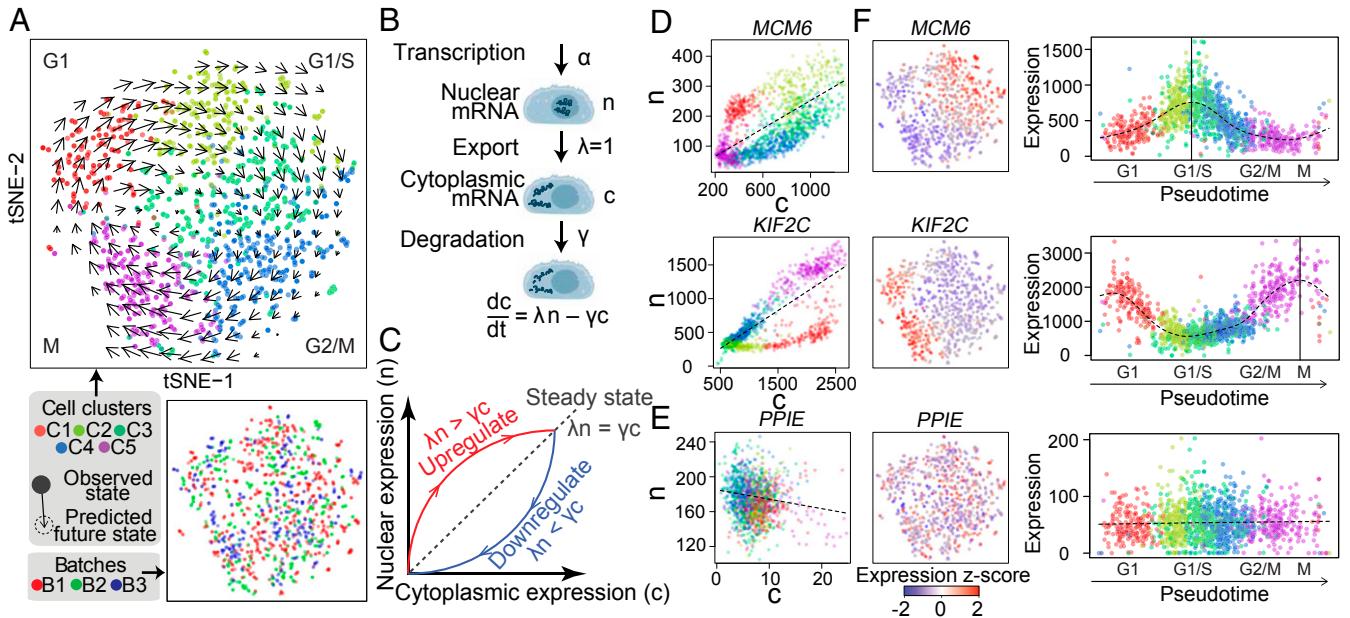
sensitivity and false discovery rates, the same criteria were applied uniformly to all 3 experimental approaches, thus offering a reasonable comparison between approaches.

In addition, we used the blank control barcodes included in our MERFISH measurements to further gauge the rate of misidentification. Out of the 2,853 blank barcodes, we identified only 30 (~1%) to be significantly enriched in the ER, suggestive of a low misidentification rate. Extension of our analysis to all 10,050 genes, including the 1,000 genes detected using the overlapping encoding-probe design, yielded similar detection sensitivity and false discovery rate (*SI Appendix*, Fig. S4). Therefore, MERFISH allows the identification of RNAs associated with subcellular compartments with both high sensitivity and a low false discovery rate at a genome-wide scale. Moreover, MERFISH provides single-cell resolution whereas the purification-based approaches have yet to achieve single-cell resolution.

Identification of RNAs Enriched in the Nucleus. Having established that MERFISH can effectively identify genes enriched in subcellular compartments, we next applied MERFISH to characterize enrichment of RNA species in the nucleus, which has a less well characterized transcriptome. In this and the following sections, we limited our analysis to the 9,050 genes detected with the nonoverlapping encoding-probe design. Because nuclear enrichment is known to vary as a function of cell cycle, we applied a more stringent fold-change criterion for enrichment ($\log2[\text{fold change between cpm-normalized nuclear and cytoplasmic counts}] > 2$) to minimize identification of nuclear enrichment due to nascent transcripts. We identified 1,488 genes as being highly significantly enriched in the nucleus ($\log2[\text{fold change}] > 2$; Bonferroni-corrected $P < 1e-10$; Fig. 3*A* and Dataset S6) and additional nuclear-enriched genes could be identified with less stringent criteria on fold change and *P* value using the all-gene data provided in Dataset S6. Because certain RNA species may be enriched in the perinuclear region outside the nucleus, such as the ER, we further performed a more stringent nuclear segmentation by eroding away ~1 μm around the nuclear circumference. Still, after such conservative segmentation, 1,484 of the 1,488 (>99%) identified genes remained significantly enriched (Bonferroni-corrected $P < 1e-10$), with highly correlated fold-change numbers between the 2 segmentation criteria (*SI Appendix*, Fig. S5). Out of the 2,853 blank control barcodes, we identified only 17 as being significantly enriched ($\log2[\text{fold change}] > 2$; Bonferroni-corrected $P < 1e-10$) in the nucleus, again suggestive of a low misidentification rate. In addition, among the 507 gold-standard consensus ER-enriched RNA species identified by MERFISH, none were identified to be significantly nuclear-enriched by our criteria, suggesting that the ER-associated RNAs enriched in the perinuclear region did not contaminate the nuclear-enriched set appreciably.

Because we designed encoding probes specifically to preferentially target 1 isoform of each gene in our 10,050-gene library (*SI Appendix, Materials and Methods*), we were able to distinguish the biotype of each targeted RNA. With this ability, we observed several specific categories of RNA species enriched in the nucleus. For example, long noncoding RNAs are known to often be localized in the nucleus (23, 28). Indeed, of the 17 long intergenic noncoding RNAs (lincRNAs) that satisfied Bonferroni-corrected $P < 1e-10$ in our experiments, 10 of them were observed to be highly enriched in the nucleus with our stringent fold-change threshold (Fig. 3*B* and Dataset S6), including the extensively studied, nuclear-enriched lncRNA *MALAT1* (29) (*SI Appendix*, Fig. S6*A*). Many other measured lncRNAs were also enriched in the nucleus (Dataset S6).

In addition, as the nuclear pore complex acts as a gate to prevent export of unspliced RNA (30), we anticipated that RNAs with retained introns will be preferentially enriched in the nucleus. Among the intron-retained RNAs that satisfied Bonferroni-corrected $P < 1e-10$ in our experiments, 85% (585 genes) were

BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Fig. 4. Characterizations of transcriptionally distinct cell states and RNA velocity. (*A*) A 2D tSNE 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 (*Top*) and by batch/replicate (*Bottom*). Projected velocity arrows show the RNA velocity as described in *B*. (*B*) Transcriptional dynamics model used to estimate RNA velocity. For each gene, the RNA velocity, defined as the rate of change in cytoplasmic RNA abundance (dc/dt), is modeled based on the transcription rate of the gene (α), rate constant of nuclear export (λ), and rate constant of RNA degradation (γ) given the observed mRNA abundance within the nucleus (n) and cytoplasm (c). Based on the RNA velocity, the future expression state of a cell can be predicted, and the projection from the current to future state is presented as arrows in *A*. (*C*) Schematic phase portrait of the nuclear versus cytoplasmic abundance of an RNA for various scenarios. The expected steady-state nuclear and cytoplasmic abundances for different transcription rates α fall on the diagonal (dotted line) with the slope γ/λ. The expected nuclear versus cytoplasmic abundance upon transcriptional up-regulation and down-regulation is depicted by the red and blue lines with arrows, respectively. (*D*) Phase portrait for 2 known cell cycle-related genes, *MCM6* and *KIF2C*. Each point represents a cell and is colored by the cluster annotations as in *A* and, for each cell, mean cpm-normalized nucleus and cytoplasm expression magnitudes from 50 nearest-neighbor cells in PC space are shown, similar to phase portrait calculations in ref. 40. The steady state (dotted line) is estimated as the ratio between n and c using pooled cells from the lower- and upper-fifth extreme expression quantiles (*SI Appendix, Materials and Methods*). (*E*) Phase portrait for a known housekeeping gene, *PPIE*, as in *D*. (*F*) Expression visualization for the 3 genes described in *D* and *E*. (*F, Left*) tSNE embedding as in *A* colored by z-scored, cpm-normalized total cell expression level of the indicated gene. (*F, Right*) Cpm-normalized total gene expression magnitude versus pseudotime for each gene. Each point is a cell colored by its cluster annotation described in *A*. The top and bottom 0.1% of expression magnitude was winsorized. The dashed line indicates the smooth-spline fitted curve. The vertical solid line indicates the pseudotime at which the expression level reaches a maximum (from the fitted curve).
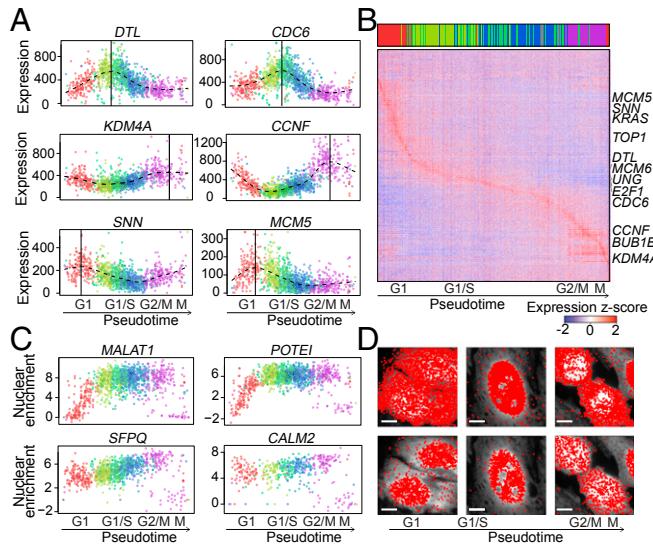
observed to be highly enriched in the nucleus (Fig. 3*C* and Dataset S6). Such nuclear retention of RNAs with retained introns has a number of proposed functions. For example, RNAs with retained introns can regulate expression of some intronic noncoding RNAs (31), and can also serve to regulate gene expression in the cytoplasm by activity-dependent splicing (32). Indeed, we observed that *NASP* was enriched in the nucleus (*SI Appendix*, Fig. S6*B*), consistent with the previous observation that this RNA is enriched in the nucleus under normal conditions but undergoes rapid splicing and export under CLK (CDC-like kinase) inhibition (32). In addition, we observed nuclear enrichment of the intron-retained form of *EIF4A2* (ENST00000485101.5; *SI Appendix*, Fig. S6*C*), which contains 5 different snoRNAs and 1 miRNA in its introns, and may thus regulate the expression of these small noncoding RNAs.

Interestingly, although protein-coding mRNAs are translated in the cytoplasm, among the RNAs classified as the protein-coding biotype that satisfied Bonferroni-corrected $P < 1e-10$, we identified 15% (579 protein-coding mRNAs) to be highly enriched in the nucleus (Fig. 3*D* and Dataset S6). These include known nuclear-enriched protein-coding mRNAs, such as *JRK* and *MDM4* (24), as well as many previously unknown ones. A number of functions have been proposed for the accumulation of mature protein-coding mRNAs in the nucleus. For example, retention of mRNAs in the nucleus may help buffer noise generated by stochastic mRNA production (33, 34). Moreover, longer genes that take more time to be transcribed and exported may be enriched in the nucleus so that cells can quickly respond to stimuli by

exporting these RNAs to the cytoplasm to up-regulate translation, bypassing the transcription step (35).

Identification of Transcriptionally Distinct Cell-Cycle States and Their Progression by RNA Velocity and Pseudotime Analysis. Single-cell transcriptomic analysis enables the identification of novel cell types and cell states in a systematic and quantitative manner (4, 36–38). The ability to perform spatially resolved RNA profiling could further facilitate cell-type and cell-state identification.

To illustrate this, we first performed single-cell clustering analysis to identify cell populations based on the gene expression profiles of individual cells. Briefly, we filtered out lowly expressed genes, performed batch correction and cpm and variance normalization, identified 1,598 overdispersed genes, and applied principal-component analysis (PCA) to identify 30 PCs that capture the greatest variance (*SI Appendix, SI Materials and Methods* and Fig. S7 *A–D*). We then applied graph-based Louvain clustering in the PC space to identify cell clusters (Fig. 4*A* and *SI Appendix*, Fig. S7 *E–G*). Overall, among 1,368 cells measured across 3 batches (replicates), we identified 5 transcriptionally distinct cell clusters (Fig. 4*A* and *SI Appendix*, Fig. S7*E*). Given that our measurements were performed on a single cell type, these clusters likely represent distinct cell states at different stages of the cell cycle. Indeed, we observed significant enrichment in expression of known cell-cycle markers in different clusters (Dataset S7). For example, *MCM5*, a known marker of the G1 phase, exhibited expression up-regulation in the C1 cluster; *UNG* and *DSCC1*,

Xia et al.

Fig. 5. Variations in gene expression and nuclear enrichment across cell-cycle phases. (*A*) Cpm-normalized total expression magnitude, winsorized as in Fig. 4*F*, versus pseudotime for 3 selected genes identified to have cell cycle–dependent expression by the pseudotime analysis (*Left*) shown together with 3 known cell-cycle genes with similar pseudotime dependence (*Right*). 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). (*B*) Heatmap of the cpm-normalized and *z*-scored expression magnitude of the 1,654 genes identified to have cell cycle–dependent expression as a function of pseudotime. The cells are sorted by the pseudotime and the genes are sorted by the pseudotime point of maximum expression (as determined from the fitted curve). Selected cell-cycle genes are labeled. The colored bar at the top denotes the cell's cluster annotation. (*C*) Nuclear enrichment (log2 of cpm-normalized nuclear and cytoplasmic expression ratio) versus pseudotime for 4 select genes. Each point represents a single cell colored by its cluster annotation. *MALAT1* and *POTEI* (*Top*) exhibit gradual reestablishment of nuclear enrichment after mitosis whereas *SFPQ* and *CALM2* (*Bottom*) exhibit instantaneous reestablishment of nuclear enrichment after mitosis. (*D*) Spatial distribution of all genes determined to have gradual recovery of nuclear enrichment similar to *MALAT1* and *POTEI* (*Top*) or instantaneous recovery of nuclear enrichment similar to *SFPQ* and *CALM2* (*Bottom*) for cells at 3 distinct pseudotime points (columns). Each red dot indicates the position of 1 transcript detected by MERFISH overlaid on the poly-dT staining image. (Scale bars: 20 μm.)

known to be up-regulated in G1/S–phase progression, exhibited enrichment in C2 and C3 clusters, respectively; *BCL2L1*, known to be relevant to G2/M transition, exhibited expression up-regulation in the C4 cluster; and *CCNF*, a known M-phase checkpoint marker, exhibited enrichment in the C5 cluster (*SI Appendix*, Fig. S8 *A* and *B*) (39). Similarly, we observed expression up-regulation of specific gene sets involved in different cell-cycle phase transitions in distinct clusters (*SI Appendix*, Fig. S8C).
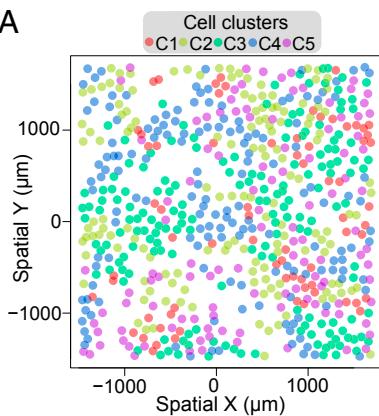
Further understanding of these cell states will benefit from quantification of temporal changes of gene expression profiles across the cell cycle. However, like scRNA-seq analysis, MERFISH measurements capture only static snapshots in time. To address this limitation, we sought to place cells on a pseudotime axis by analysis of the RNA velocity, namely the time derivative of the gene expression state. Previously, La Manno et al. (40) described an approach to estimate RNA velocity by distinguishing between unspliced and spliced mRNAs identified from scRNA-seq data under the assumption that unspliced mRNAs are indicative of nascent transcription while spliced mRNAs are indicative of the mature version. As mRNAs are transcribed within the nucleus and then exported to the cytoplasm upon maturation, we reasoned that an alternative approach for inferring RNA velocity is to

distinguish between nuclear and cytoplasmic mRNAs, leveraging the spatial information of transcripts obtained in our measurements. The first-order time derivative of the mRNA abundance in the cytoplasm can be determined by the balance between the export of the nuclear-localized mRNAs and the degradation of the mRNAs in the cytoplasm (Fig. 4*B* and *SI Appendix, Materials and Methods*), in a way analogous to how La Manno et al. (40) treated unspliced and spliced mRNAs. Upon active up-regulation of a gene, we anticipate a rapid increase in nuclear mRNA counts, followed by an increase in cytoplasmic mRNA counts due to nuclear export until a new steady state is reached (Fig. 4*C*). Conversely, active down-regulation in transcription would lead to a rapid reduction in nuclear mRNA counts as the nuclear export of mRNAs continues; the cytoplasmic mRNA will drop eventually because of the reduction in the nuclear RNA pool for export and the continued RNA degradation in the cytoplasm (Fig. 4*C*). As a validation, we observed that *MCM6*, a cell-cycle gene that is known to be important for progression through the S phase and thus is expected to be up-regulated among G1/S cells and down-regulated among G2/M transition cells (41), was indeed up-regulated in C1 and C2 clusters and down-regulated in C3 and C4 clusters (Fig. 4*D*). In contrast, the *KIF2C* gene exhibited up-regulation among C4 and C5 clusters and down-regulation among C1 and C2 clusters (Fig. 4*D*), which is consistent with the prior knowledge that *KIF2C* is important for the M phase of the cell cycle (42). In contrast, *PPIE*, a housekeeping gene, did not exhibit coordinated variability between nuclear and cytoplasmic expression (Fig. 4*E*).
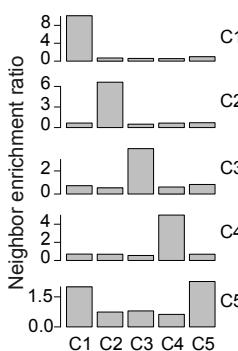
The balance of nuclear and cytoplasmic mRNA abundance can, therefore, be an indicator of the future state of cytoplasmic RNA abundance. We used this approach to determine the RNA velocity for each cell (considering only the RNAs with positive correlation between nuclear and cytoplasmic counts [*SI Appendix, Materials and Methods*]) and projected these velocities as arrows on the tSNE (Fig. 4*A*) and PCA (*SI Appendix*, Fig. S9A) plots. Consistent with expectations, our projected velocity arrows indicate that cells up-regulating G1 markers (C1 clusters) are transcriptionally moving toward cells up-regulating G1/S markers (C2 and C3 clusters), which are in turn transcriptionally moving toward cells up-regulating G2/M markers (C4 and C5 clusters), and the RNA velocity arrows form a circle on the tSNE and PCA plots (Fig. 4*A* and *SI Appendix*, Fig. S9A). We thus interpreted this cell ordering along the circle as the pseudotime. Examination of gene expression of known cell-cycle markers across pseudotime further validated our pseudotime ordering (Fig. 4*F* and *SI Appendix*, Fig. S8D). Furthermore, consistent with our model, we find that the pseudotime points of maximal nuclear expression preceded the pseudotime points of maximal cytoplasmic expression for many cell cycle-related genes (*SI Appendix*, Fig. S9B).

Pseudotime Analysis Reveals Novel Cell Cycle-Related Genes and Nuclear-Retention Dynamics. Armed with pseudotime ordering of cells across cell-cycle phases, we sought to more systematically identify potentially novel cell cycle-related genes. To this end, we used a linear regression model to quantify the variance in gene expression explained by pseudotime and identified 1,654 genes with putative cell cycle-dependent expression magnitude, defined as genes with a statistically significant (Bonferroni-corrected $P <$ 0.05) proportion of expression-level variance that was captured by the pseudotime (Dataset S8), albeit that a fraction of these genes had relatively weak dependence of expression on pseudotime. Among these putative cell cycle-related genes are 120 genes associated with the mitotic cell-cycle GO term (GO:0000278). In addition to these previously annotated cell-cycle genes, we also identified many other genes with expression levels dependent on the pseudotime. For example, we observed that *DTL* was most highly expressed at a pseudotime point where canonical G1/S markers, such as *CDC6*, are most strongly expressed (Fig. 5*A*),
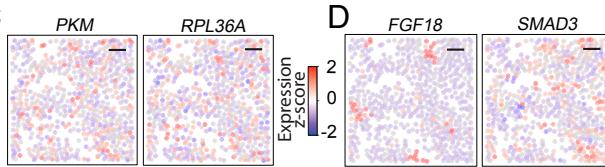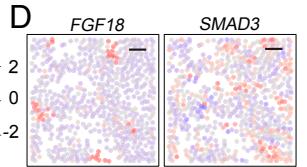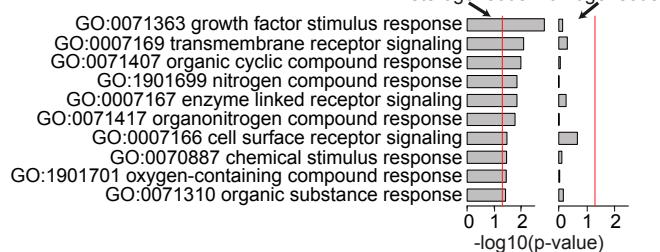
Fig. 6. Spatial heterogeneity of transcriptionally distinct cells. (*A*) Centroid positions of the cells measured in 1 MERFISH replicate (points) colored by the corresponding cell's cluster annotation. (*B*) Bar plots of neighbor enrichment ratios for cells assigned to each of the 5 clusters. For cells in the C1 cluster (*Top*), the neighbor enrichment ratio is calculated by first partitioning the cells into 2 sets: a neighbor cell set containing all cells that are within the 3 nearest neighbors of any cell in the C1 cluster and a nonneighbor cell set containing all other cells. The enrichment ratios for cells in a particular cluster (Cx, x = 1 to 5) in the neighborhood of C1 cells are defined as the ratio between the fraction of cells in the neighbor set that belong to the Cx cluster and the fraction of cells in the nonneighbor set that belong to the Cx cluster. The enrichment ratios in the neighborhood of C2, C3, C4, and C5 cells are determined similarly and shown in the 4 panels below. All clusters' spatial neighborhood exhibits significant enrichment (Fisher's exact $P < 1e-10$) for cells in the same cluster. (*C* and *D*) Spatial expression profiles for select overdispersed non-cell cycle-related genes that do not exhibit significant spatial heterogeneity in expression (Moran's I Bonferroni-corrected $P > 0.05$) (*C*) or exhibit highly significant spatial heterogeneity in expression (Moran's I Bonferroni-corrected $P < 1e-10$) (*D*). Each point indicates the spatial position of a cell and is colored based on the *z*-scored, cpm-normalized expression magnitude of *PKM* and *RPL36A* (*C*) and *FGF18* and *SMAD3* (*D*). (Scale bars: *C* and *D*, 500 μm.) (*E*) Gene set enrichment *P* values for a set of significantly enriched gene sets identified for genes that exhibit highly significant (Moran's I Bonferroni-corrected $P < 1e-10$) spatial heterogeneity (*Left*) and *P* values of the same gene sets for genes that appear spatially homogeneous, i.e., do not exhibit significant (Moran's I Bonferroni-corrected $P > 0.05$) spatial heterogeneity (*Right*). The red lines indicate the $P = 0.05$ significance threshold.

consistent with the previous observation that *DTL* down-regulates DNA replication factor *CDT1* in S phase to prevent rereplication (43). Likewise, we observed that *KDM4A*, a histone demethylase gene, was most highly expressed at a pseudotime point where canonical G2/M markers, such as *CCNF*, are most strongly expressed (Fig. 5*A*), suggesting that *KDM4A* may function to

regulate G2/M phase progression. This notion is supported by the observations that down-regulation of *KDM4A* arrests cells in the G2/M phase (44) and that *KDM4A* expression varies with cell cycle (39). Finally, we found that *SNN* (Stannin), a protein-coding gene with unknown cellular function, was most highly up-regulated at a pseudotime point where canonical G1 markers, such as *MCM5*, are up-regulated (Fig. 5*A*), suggesting a function of *SNN* in the G1 phase. Consistent with this, knockdown of *SNN* has been shown to significantly alter expression of genes associated with the G1 progression (45). These results highlight the ability of our pseudotime analysis to identify potentially novel cell cycle-related genes.

Interestingly, by ordering genes based on their maximum points of expression along pseudotime, we observed a gradual change of the transcription profile across cells along the pseudotime axis (Fig. 5*B*). In particular, across some of the cell-cycle stages, such as the G1 phase, many genes were up-regulated in succession. As a control, randomizing cells in pseudotime, or randomizing cells in pseudotime within each cluster, but still ordering the genes based on their maximum points of expression along the new randomized pseudotime axis did not generate such a pattern of gradual transcription profile change (*SI Appendix*, Fig. S10). These results suggest that a cell undergoes many gradual transcriptional changes as it progresses through some cell-cycle stages, rather than punctually transitioning from one cell-cycle stage to the next.

In addition to variations in the gene expression profile in pseudotime, we also studied how the intracellular distributions of RNAs may vary with pseudotime. In particular, we examined the degree of nuclear enrichment as a function of pseudotime for the 1,488 genes that we identified to be highly nuclear-enriched. As expected, we found that nuclear enrichment was lost during cell division due to the dissolution of the nuclear envelope (Fig. 5 *C* and *D*). We noticed that a number of these RNAs resumed nuclear enrichment immediately after mitosis (Fig. 5 *C* and *D*, *Bottom* and Dataset S9). Notably, some other genes, as exemplified by the lncRNA *MALAT1*, remained evenly distributed across the nucleus and cytoplasm even within postmitotic daughter cells, and the nuclear enrichment gradually recovered through the G1 phase (Fig. 5 *C* and *D*, *Top* and Dataset S9).

Cell Cycle-Dependent and -Independent Spatial Heterogeneity of Cell Populations. The spatial information provided by MERFISH also allows spatially resolved analysis of transcriptionally distinct cell populations. Taking advantage of this spatial information, we observed that cells within each cell-cycle cluster tended to be spatially proximal to cells of the same cluster (Fig. 6 *A* and *B*). We speculate that this phenomenon can be at least partly attributed to the fact that spatially neighboring cells are more likely sibling cells from the same mother cells, and hence tend to be in similar cell-cycle stages. As such, the expression levels of some cell cycle-related genes also exhibited spatial heterogeneity across cells (*SI Appendix*, Fig. S11).

Notably, we also observed 1,053 overdispersed genes with expression variance that was not captured by the pseudotime (Dataset S10). We reasoned that for a portion of these genes, such high variance may be indicative of spatial patterning or spatial context-dependent expression. Thus, we examined whether the expression levels of these genes showed spatial heterogeneity (i.e., positive spatial autocorrelation) across cells using a one-sided Moran's I statistic (Dataset S11) (46). Indeed, among the 1,053 overdispersed, non-cell cycle-related genes, only a small fraction (272 genes) did not show statistically significant spatial heterogeneity in their expression levels (Moran's I Bonferroni-corrected $P > 0.05$), exemplified by housekeeping genes such as *RPL36A*, which encodes a ribosomal protein subunit, and *PKM*, which is involved in glycolysis (Fig. 6*C*). The majority of these 1,053 genes exhibited varying degrees of spatial heterogeneity. Among these, we identified 339 genes, the expression levels for which

showed highly significant (Moran's I Bonferroni-corrected $P <$ 1e-10) spatial heterogeneity across cells ([Dataset S11](#)).

We reason that the spatial heterogeneity in the expression levels of these genes across cells might be due to local environmental stimulation or cell–cell communication. In support of this notion, genes within this group tended to be associated with the GO terms "cellular communication" (GO:0007154) and "cellular response to stimulus" (GO:0051716). Examples include *FGF18*, a member of the fibroblast growth factor family of secreted proteins, and *SMAD3*, a member of the SMAD family of signal transducers and transcriptional modulators (Fig. 6*D*). Gene set enrichment analysis of the GO terms that are children of cellular communication and cellular response to stimulus in the GO term hierarchy indeed showed enrichment of many of these terms among the 339 spatially highly heterogeneous genes, including terms associated with cellular response to growth factor and chemical stimuli (Fig. 6*E*). In contrast, none of these GO terms were enriched among the 272 genes that did not show spatial heterogeneity in expression (Fig. 6*E*). These results illustrate the ability of spatially resolved single-cell transcriptomics to characterize the interplay between transcriptional and spatial heterogeneity.

Discussion

Here, we demonstrated spatially resolved RNA quantification of ~10,000 genes in single cells with ~80% detection efficiency and 1% blank control calling rate (suggestive of ~4% barcode misidentification rate) for cultured cells using MERFISH. By combining our method with ER and nucleus imaging, we characterized the enrichment of RNAs within these intracellular compartments. As a validation, our measurements detected enrichment of secretome genes at the ER with a high sensitivity and a low false discovery rate. We further observed enrichment of lncRNAs, RNAs with retained introns, and a subgroup of protein-coding mRNAs in the cell nucleus. Furthermore, our spatially resolved RNA profiling, in particular the discrimination of nuclear and cytoplasmic RNAs, provided an approach to compute the RNA velocity in situ, which in turn allowed us to project cells onto a pseudotime axis that corresponded to different stages of the cell cycle. Using this approach, we identified ~1,600 genes with putative cell cycle-dependent expression. Notably, we observed that the transcriptional profiles of cells undergo gradual changes across some cell-cycle states, with many genes being up-regulated in succession. Finally, we observed spatial heterogeneity of transcriptionally distinct cell populations arising through both cell cycle-dependent and cell cycle-independent mechanisms.

These results highlight the ability of MERFISH to perform near-genome-wide quantification and localization of RNAs in individual cells. We envision that an even greater number of RNA species could be imaged in individual cells. As the number of RNAs increases, the overall density of RNA molecules inside the cells will increase, which may require greater expansion of the sample (our current expansion factor is only ~2 in each dimension) or a larger number of hybridization rounds. Another important factor to consider is the lengths of the RNAs. Here we imaged RNAs that are 500 nt or longer with up to 48 encoding probes per gene. We recently demonstrated the use of branched DNA amplification in MERFISH, which has allowed RNA to be detected with as few as 16 encoding probes and, when combined with the overlapping encoding-probe design, should allow RNAs as short as ~100 to 200 nt to be detectable (47). This decrease in RNA length requirement should also further increase our ability to distinguish different isoforms of a gene.

Beyond providing spatial information of RNAs and cells, MERFISH presents a number of additional advantages compared with conventional single-cell sequencing-based approaches. The high detection efficiency of MERFISH enables accurate quantification of lowly expressed genes, such as transcription factors, which may be challenging to investigate using scRNA-seq due to

dropout effects (48). Furthermore, the ability to detect RNAs in situ can also avoid dissociation-induced perturbation to cells. On the other hand, the necessity to target genes with a predesigned library of oligonucleotide probes currently makes the de novo discoveries of gene fusions and single-nucleotide variations by MERFISH more challenging than by sequencing.

While our paper was in preparation, Eng et al. (7) reported the imaging of 10,000 genes, with ~50% detection efficiency for cultured cells, using a multiplexed smFISH method which they termed seqFISH+. Both Eng et al. and our current study used error-correcting barcodes to encode RNA species, a library of encoding probes to place readout sequences on individual RNA molecules to present the barcodes, and sequential hybridization with readout probes complementary to the readout sequences to detect the barcodes, as previously demonstrated in MERFISH (6). Distinct error-correcting codes were used in these two studies. Here, we used a 69-bit error-correcting code with Hamming distance 4 and Hamming weight 4 to image ~10,000 genes using 3-color imaging and 23 rounds of hybridization, whereas Eng et al. used 3 independent sets of 80-bit, Hamming weight 4, error-correcting codes [a kind that can also be considered as a 4-unit base-20 code (7)] to detect ~10,000 genes with 3-color imaging and 80 rounds of hybridization. To help separate neighboring RNA molecules, we used expansion microscopy (17), a strategy that we have previously combined with MEFFISH (18), whereas Eng et al. used a greater number of hybridization rounds, to reduce the RNA spot density per round. In the sequential FISH approach using a color-coding scheme that requires each RNA molecule to be "on" in one of the several true color channels in each imaging round, as previously reported in seqFISH (49), increasing the number of hybridization rounds does not decrease the RNA spot density per round. Instead, encoding schemes that include both "on" (1) and "off" (0) signals in barcodes and allow the fraction of "1" bits to be varied by changing the number of hybridization rounds (6, 7) can use this parameter to adjust the RNA spot density in each imaging round. When the number of "1" bits per barcode (i.e., Hamming weight) is kept the same (reducing the Hamming weight greatly reduces the coding capacity), the 2 strategies—sample expansion or hybridization number increase—increase the imaging time by approximately the same amount in order to achieve the same level of dilution of RNA spot density. However, by reducing the number of hybridization rounds required, sample expansion has the additional advantage of reducing the total time required for hybridization and signal removal during imaging. In addition, sample expansion allows physical separation of RNA molecules and dilutes the fluorescence background from cell autofluorescence or nonspecific binding of the probes, which may have, among other factors, contributed to the high detection efficiency and accuracy achieved in our study.

In addition to multiplexed FISH, in situ sequencing also allows transcriptome-scale RNA profiling of single cells. For example, fluorescence in situ sequencing (FISSEQ) has demonstrated the detection of several thousands of RNA species in an untargeted manner (9), but the detection efficiency is low (4, 9). Recently, spatially resolved transcript amplicon readout mapping (STARmap), another type of in situ sequencing method, has been developed to substantially increase the detection efficiency of in situ sequencing to a level comparable to or higher than single-cell RNA-seq, and has demonstrated simultaneous measurement of ~1,000 targeted genes (10).

We envision that the ability of MERFISH to provide quantitative and spatially resolved RNA measurements at the genome scale in single cells with high accuracy and detection efficiency will allow a wide array of biological questions to be addressed. The subcellular resolution of MERFISH, combined with the ability to simultaneously image other cellular structures as we demonstrated here, enables the determination of the intracellular distribution and

compartmentalization of RNAs and how this spatial organization changes as a function of cell states and in response to external stimuli. In addition, transcriptome imaging could also be combined with imaging of other molecular factors to probe, for example, chromatin structures and protein factors involved in transcriptional and posttranscriptional regulation. The quantitative single-cell expression profiling capability further allows distinct cell types and cell states to be identified and their spatial organizations to be determined in situ. This ability to spatially map transcriptionally distinct cells, in combination with in situ RNA velocity analysis powered by the discrimination of nuclear and cytoplasmic mRNAs, will further enhance our understanding of how the transcriptional properties and spatial patterns of cells evolve during differentiation, development, and disease progression.

Materials and Methods

Detailed descriptions of all experimental protocols and analysis methods are provided in *SI Appendix, Materials and Methods*.

In brief, U-2 OS cells (ATCC) were cultured, stained with cellular-structure markers and MERFISH encoding probes, gel-embedded, and expanded following protocols previously described (6, 18, 19). Samples were imaged on a custom fluorescence microscope with an automated fluidics system to exchange the hybridization, wash, imaging, and cleavage buffers for each hybridization round. RNA transcripts were decoded using a voxel-based decoding algorithm that assigned barcodes to individual voxels, aggregated adjacent voxels having the same barcode as putative transcripts, and removed barcodes containing identifiable errors.

Images of the ER, nucleus, and total poly-A RNA stains were segmented computationally, the total RNA counts and numbers of RNA molecules colocalized with the ER and nucleus were determined for each gene in each cell, and genes spatially enriched at the ER and nucleus were identified using a 2-sided pairwise Wilcoxon rank-sum test. The total RNA counts, and the counts colocalized with the ER and nucleus for each gene in each cell, are shown in Datasets S12, S13, and S14 and the position information of each cell is shown in Dataset S15. Clustering of cells based on single-cell gene expression profiles was performed using a graph-based Louvain community detection method (50). RNA velocity analysis was performed based on the balance of nuclear versus cytoplasmic RNA counts in a manner analogous to the previously developed approach based on unspliced versus spliced RNA counts (40). Spatial heterogeneity was quantified using Moran's I spatial autocorrelation statistic (46).

1. A. R. Buxbaum, G. Haimovich, R. H. Singer, In the right place at the right time: Visualizing and understanding mRNA localization. *Nat. Rev. Mol. Cell Biol.* **16**, 95–109 (2015).
2. J. M. Engreitz, N. Ollikainen, M. Guttman, Long non-coding RNAs: Spatial amplifiers that control nuclear structure and gene expression. *Nat. Rev. Mol. Cell Biol.* **17**, 756–770 (2016).
3. N. Crosetto, M. Bienko, A. van Oudenaarden, Spatially resolved transcriptomics and beyond. *Nat. Rev. Genet.* **16**, 57–66 (2015).
4. E. Lein, L. E. Borm, S. Linnarsson, The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. *Science* **358**, 64–69 (2017).
5. J. M. Levsky, S. M. Shenoy, R. C. Pezo, R. H. Singer, Single-cell gene expression profiling. *Science* **297**, 836–840 (2002).
6. K. H. Chen, A. N. Boettiger, J. R. Moffitt, S. Wang, X. Zhuang, Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
7. C. L. Eng *et al.*, Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568**, 235–239 (2019).
8. R. Ke *et al.*, In situ sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods* **10**, 857–860 (2013).
9. J. H. Lee *et al.*, Highly multiplexed subcellular RNA sequencing in situ. *Science* **343**, 1360–1363 (2014).
10. X. Wang *et al.*, Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691 (2018).
11. P. L. Ståhl *et al.*, Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).
12. S. G. Rodriques *et al.*, Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463–1467 (2019).
13. A. M. Femino, F. S. Fay, K. Fogarty, R. H. Singer, Visualization of single RNA transcripts in situ. *Science* **280**, 585–590 (1998).
14. A. Raj, P. van den Bogaard, S. A. Rifkin, A. van Oudenaarden, S. Tyagi, Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* **5**, 877–879 (2008).
15. J. R. Moffitt *et al.*, High-performance multiplexed fluorescence in situ hybridization in culture and tissue with matrix imprinting and clearing. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 14456–14461 (2016).
16. J. R. Moffitt *et al.*, Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, eaau5324 (2018).
17. F. Chen, P. W. Tillberg, E. S. Boyden, Expansion microscopy. *Science* **347**, 543–548 (2015).
18. G. Wang, J. R. Moffitt, X. Zhuang, Multiplexed imaging of high-density libraries of RNAs with MERFISH and expansion microscopy. *Sci. Rep.* **8**, 4847 (2018).
19. J. R. Moffitt *et al.*, High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 11046–11051 (2016).
20. T. Ast, G. Cohen, M. Schuldiner, A network of cytosolic factors targets SRP-independent proteins to the endoplasmic reticulum. *Cell* **152**, 1134–1145 (2013).
21. D. W. Reid, C. V. Nicchitta, Primary role for endoplasmic reticulum-bound ribosomes in cellular translation identified by ribosome profiling. *J. Biol. Chem.* **287**, 5518–5527 (2012).
22. C. H. Jan, C. C. Williams, J. S. Weissman, Principles of ER cotranslational translocation revealed by proximity-specific ribosome profiling. *Science* **346**, 1257521 (2014).
23. P. Kaewsapsak, D. M. Shechner, W. Mallard, J. L. Rinn, A. Y. Ting, Live-cell mapping of organelle-associated RNAs via proximity biotinylation combined with protein-RNA crosslinking. *eLife* **6**, e29224 (2017).
24. F. M. Fazal *et al.*, Atlas of subcellular RNA localization revealed by APEX-seq. *Cell* **178**, 473–490.e26 (2019).
25. A. Subramanian *et al.*, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15545–15550 (2005).
26. L. Käll, A. Krogh, E. L. Sonnhammer, A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **338**, 1027–1036 (2004).
27. T. N. Petersen, S. Brunak, G. von Heijne, H. Nielsen, SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nat. Methods* **8**, 785–786 (2011).
28. T. Derrien *et al.*, The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789 (2012).
29. M. N. Cabili *et al.*, Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biol.* **16**, 20 (2015).
30. M. J. Moore, N. J. Proudfoot, Pre-mRNA processing reaches back to transcription and ahead to translation. *Cell* **136**, 688–700 (2009).
31. J. J. Wong, A. Y. Au, W. Ritchie, J. E. Rasko, Intron retention in mRNA: No longer nonsense: Known and putative roles of intron retention in normal and disease biology. *Bioessays* **38**, 41–49 (2016).
32. P. L. Boutz, A. Bhutkar, P. A. Sharp, Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Genes Dev.* **29**, 63–80 (2015).
33. K. Bahar Halpern *et al.*, Nuclear retention of mRNA in mammalian tissues. *Cell Rep.* **13**, 2653–2662 (2015).
34. N. Battich, T. Stoeger, L. Pelkmans, Control of transcript variability in single mammalian cells. *Cell* **163**, 1596–1610 (2015).
35. B. W. Solnestam *et al.*, Comparison of total and cytoplasmic mRNA reveals global regulation by nuclear retention and miRNAs. *BMC Genomics* **13**, 574 (2012).
36. A. Tanay, A. Regev, Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331–338 (2017).
37. H. Zeng, J. R. Sanes, Neuronal cell-type classification: Challenges, opportunities and the path forward. *Nat. Rev. Neurosci.* **18**, 530–546 (2017).
38. K. D. Birnbaum, Power in numbers: Single-cell RNA-seq strategies to dissect complex tissues. *Annu. Rev. Genet.* **52**, 203–221 (2018).
39. M. L. Whitfield *et al.*, Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell* **13**, 1977–2000 (2002).
40. G. La Manno *et al.*, RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
41. S. E. Kearsey, D. Maiorano, E. C. Holmes, I. T. Todorov, The role of MCM proteins in the cell cycle control of genome duplication. *Bioessays* **18**, 183–190 (1996).
42. T. Maney, A. W. Hunter, M. Wagenbach, L. Wordeman, Mitotic centromere-associated kinesin is important for anaphase chromosome segregation. *J. Cell Biol.* **142**, 787–801 (1998).
43. C. L. Sansam *et al.*, DTL/CDT2 is essential for both CDT1 regulation and the early G2/M checkpoint. *Genes Dev.* **20**, 3117–3129 (2006).
44. T. D. Kim, S. Shin, W. L. Berry, S. Oh, R. Janknecht, The JMJD2A demethylase regulates apoptosis and proliferation in colon cancer cells. *J. Cell. Biochem.* **113**, 1368–1376 (2012).
45. B. E. Reese, D. Krissinger, J. K. Yun, M. L. Billingsley, Elucidation of stannin function using microarray analysis: Implications for cell cycle control. *Gene Expr.* **13**, 41–52 (2006).
46. P. A. Moran, Notes on continuous stochastic phenomena. *Biometrika* **37**, 17–23 (1950).
47. C. Xia, H. P. Babcock, J. R. Moffitt, X. Zhuang, Multiplexed detection of RNA using MERFISH and branched DNA amplification. *Sci. Rep.* **9**, 7721 (2019).
48. P. V. Kharchenko, L. Silberstein, D. T. Scadden, Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740–742 (2014).
49. E. Lubeck, A. F. Coskun, T. Zhiyentayev, M. Ahmad, L. Cai, Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods* **11**, 360–361 (2014).
50. V. D. Blondel, J. L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008).