High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization

Jeffrey R. Moffitt[a,b,1,2], Junjie Hao[a,b,1], Guiping Wang[a,b,1], Kok Hao Chen[a,b], Hazen P. Babcock[c], and Xiaowei Zhuang[a,b,c,d,2]

[a]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138; [b]Howard Hughes Medical Institute, Harvard University, Cambridge, MA 02138; [c]Center for Advanced Imaging, Harvard University, Cambridge, MA 02138; and [d]Department of Physics, Harvard University, Cambridge, MA 02138

Contributed by Xiaowei Zhuang, August 4, 2016 (sent for review July 7, 2016; reviewed by Arjun Raj and Aviv Regev)

Image-based approaches to single-cell transcriptomics, in which RNA species are identified and counted in situ via imaging, have emerged as a powerful complement to single-cell methods based on RNA sequencing of dissociated cells. These image-based approaches naturally preserve the native spatial context of RNAs within a cell and the organization of cells within tissue, which are important for addressing many biological questions. However, the throughput of these image-based approaches is relatively low. Here we report advances that lead to a drastic increase in the measurement throughput of multiplexed error-robust fluorescence in situ hybridization (MERFISH), an image-based approach to single-cell transcriptomics. In MERFISH, RNAs are identified via a combinatorial labeling approach that encodes RNA species with error-robust barcodes followed by sequential rounds of single-molecule fluorescence in situ hybridization (smFISH) to read out these barcodes. Here we increase the throughput of MERFISH by two orders of magnitude through a combination of improvements, including using chemical cleavage instead of photobleaching to remove fluorescent signals between consecutive rounds of smFISH imaging, increasing the imaging field of view, and using multicolor imaging. With these improvements, we performed RNA profiling in more than 100,000 human cells, with as many as 40,000 cells measured in a single 18-h measurement. This throughput should substantially extend the range of biological questions that can be addressed by MERFISH.

single-cell analysis | fluorescence | in situ hybridization | transcriptomics | multiplexed imaging

Single-cell transcriptomics, powered by next-generation RNA sequencing (RNA-seq), has transformed many aspects of cellular and tissue-scale biology (1–3). This capability has allowed researchers to address exciting questions ranging from the response of single immune cells to antigen (4–6) to the number of transcriptionally distinct cell types and the cellular heterogeneity within complex tissues (7–13). Recent advances in the automated handling of individual cells and the sequencing library preparation for these cells have substantially increased the number of cells that can be routinely characterized with these approaches; notably, state-of-the-art droplet-based RNA-seq approaches provide the ability to quantify the transcriptome of tens of thousands or more cells (14, 15). This throughput allows rare populations of cells to be characterized and transcriptionally distinct cell types within sizable tissue blocks to be mapped.

However, in most approaches to single-cell transcriptomics, cells are dissociated from tissues, and RNAs are extracted from cells; as a result, the native spatial context of these RNAs is lost. However, this spatial information is important for a complete understanding of many biological behaviors (16). For example, the spatial organization of individual cell types within most tissues is crucial to how tissue function or dysfunction arises from the behavior of individual cells. Likewise, the intracellular spatial organization of RNAs is a powerful form of posttranscriptional regulation; thus, it is often important to know not only how many RNA copies are present within a cell but also where they are located within that cell (17). Addressing questions such as these requires spatially resolved approaches to single-cell transcriptomics (16).

Recently we introduced an image-based approach to spatially resolved, single-cell transcriptomics, multiplexed error-robust fluorescence in situ hybridization (MERFISH) (18). In this approach, RNAs are identified via single-molecule FISH (smFISH) (19, 20), as opposed to alternative in situ methods using sequencing (21, 22). MERFISH uses error-robust barcoding schemes to encode RNA species and reads out these barcodes with sequential rounds of smFISH measurements (Fig. 1*A*). In our previous implementation of MERFISH (18), RNAs were encoded with binary barcodes and hybridized with complex sets of oligonucleotide probes termed "encoding probes" (Fig. S1). Each encoding probe contains a targeting sequence that binds a given cellular RNA and multiple readout sequences. The collection of readout sequences associated with a cellular RNA corresponds to the barcode of that RNA species. These barcodes then are read out through a series of smFISH measurements; in each round, the sample is stained with a readout probe complementary to one of the readout sequences, the sample is imaged, and the fluorescence signal is extinguished via photobleaching. This process then is repeated with a different readout probe, and the specific on/off pattern of fluorescence observed across multiple smFISH rounds defines the binary barcode ("1": readout probe bound, "0" readout

Significance

Image-based approaches to single-cell transcriptomics offer the ability to quantify not only the copy number of RNAs within cells but also the intracellular RNA location and the spatial organization of cells within cultures or tissues. Here we report advances in multiplexed error-robust fluorescence in situ hybridization (MERFISH) that increase the measurement throughput by two orders of magnitude and allow gene expression profiling of ∼40,000 human cells in a single 18-h measurement. This drastic increase in throughput should facilitate the identification and study of rare populations of cells as well as the characterization of transcriptionally distinct cell types within large tissue regions.

Author contributions: J.R.M., J.H., G.W., K.H.C., H.P.B., and X.Z. designed research; J.R.M., J.H., and G.W. performed research; J.R.M. and H.P.B. contributed new reagents/analytic tools; J.R.M., J.H., and G.W. analyzed data; and J.R.M., J.H., G.W., and X.Z. wrote the paper.

Reviewers: A. Raj, University of Pennsylvania; and A. Regev, MIT and Broad Institute.

Conflict of interest statement: X.Z., J.R.M., and K.H.C. are inventors on a patent applied for by Harvard University that covers the MERFISH method.

Freely available online through the PNAS open access option.

[1]J.R.M., J.H., and G.W. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: zhuang@chemistry.harvard.edu or lmoffitt@mcb.harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1612826113/-/DCSupplemental.

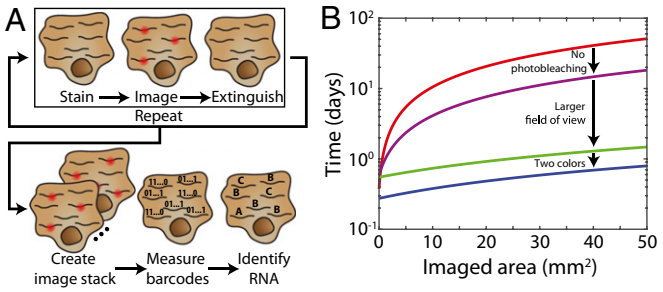

Fig. 1. Approaches to improve the measurement throughput of MERFISH. (*A*) Simplified schematic of a MERFISH readout protocol. Target RNAs are stained with encoding probes that contain a barcode comprising a combination of readout sequences unique to each RNA species. The barcode then is identified through successive rounds of smFISH, each with a readout probe complementary to one readout sequence. A registered stack of smFISH images for each sample produces an ensemble of fluorescence spots with on/off patterns that define binary barcodes ("1" represents fluorescent signal on, and "0" represents fluorescent signal off) which allow individual RNA species to be identified. A more detailed hybridization and imaging procedure is shown in Fig. S1. (*B*) The time required to perform a MERFISH experiment for a given sample area for the published protocol (18, 23) that uses photobleaching to remove smFISH signal (red line), a modified protocol without photobleaching (purple line), a modified protocol without photobleaching and a larger FOV (green line), and a modified protocol without photobleaching, a large FOV, and two-color imaging (blue line).

probe not bound) used to identify each RNA. We use error-robust barcodes that allow measurement errors to be identified and, in some cases, corrected to ensure high-accuracy MERFISH measurements (18). Using this approach we have previously demonstrated the ability to image 140 RNA species with an 80% detection efficiency using 16 rounds of smFISH imaging with an encoding scheme capable of detecting and correcting errors and to image 1,000 RNA species with a 30% detection efficiency with an encoding scheme capable of detecting but not correcting errors (18). In both cases, we were able to quantify the copy number and spatial distribution of these RNAs within ∼100 human fibroblast cells in a single ∼18-h measurement. However, for many biological questions, such as the study of rare populations of cells or the survey of sizeable volumes of tissues, it is highly desirable to increase the throughput of MERFISH so that many more cells can be measured.

Here we present an improved MERFISH method that drastically increases the throughput of this technique, simplifies several aspects of this protocol, and increases the measurement accuracy. With these improvements, we demonstrated the ability to perform spatially resolved gene expression profiling of ∼40,000 cultured human osteosarcoma (U-2 OS) cells in a single 18-h experiment. As a simple illustration of the benefits of this increased throughput, we characterized 130 genes in ∼100,000 cells, identified a subpopulation of cells undergoing DNA replication or cell division, and characterized both the expression profile and the spatial distribution of the cells that comprised this subpopulation.

Results

Increasing the Throughput of MERFISH Measurements. The total time required for a MERFISH measurement can be divided into an area-dependent time that scales with the total imaged area and an area-independent time that does not. The area-dependent time includes the time required to position, focus, and image each field of view (FOV). In addition, because of the high illumination intensity required to photobleach the fluorescence signals between consecutive rounds of smFISH, each FOV must be photobleached individually; thus, this time is also a part of this area-dependent time. The area-independent time includes buffer-exchange times and incubation times required for sample staining and thus scales with the number of rounds of smFISH that must be performed. Fig. 1*B* illustrates the scaling of the duration of a MERFISH measurement with the imaged area (red line). For 16 rounds of hybridization and imaging, the total area-independent time amounts to several hours; however, this area-independent time is exceeded by the area-dependent time when the imaged sample area is larger than ∼1 mm^2.

To improve the throughput of MERFISH, we first sought to decrease the area-dependent time. In our previously published MERFISH protocols (18, 23), imaging an FOV of ∼40 × 40 μm required only 0.1 s, but photobleaching of this same FOV required a significantly longer exposure, ∼3 s. Thus, we devised a scheme in which the smFISH signal from the entire sample could be extinguished simultaneously by chemical reaction instead of photobleaching. Specifically, we reasoned that fluorescent dyes conjugated to readout probes via a disulfide linkage could be cleaved from these probes rapidly with a mild reducing agent such as Tris(2-carboxyethyl)phosphine (TCEP) (Fig. 2*A*).

To test this approach, we hybridized encoding probes containing readout sequences to the filamin A (*FLNA*) mRNA in human lung fibroblast (IMR-90) cells and then stained this sample with a readout probe that was conjugated to a Cy5 dye via a disulfide bond. As expected, the sample exhibited bright fluorescent spots representing individual molecules of the *FLNA* mRNA, and these fluorescent spots reduced in brightness and eventually disappeared upon treatment with 50 mM TCEP (Fig. 2*B*). When averaged across thousands of RNAs, the brightness of these spots decayed exponentially (Fig. 2*C*) with a half-life of 1.17 ± 0.07 min (95% confidence interval). This half-life did not depend on the sequence of the readout probe or the dye to which it was conjugated (Fig. 2*D*). After ∼15 min of TCEP treatment, the average brightness of each RNA spot and the number of detected RNA

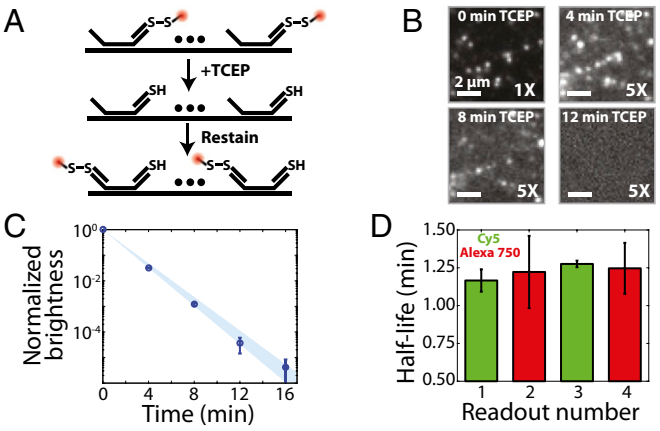

Fig. 2. Reductive cleavage of disulfide-linked fluorophores removes the fluorescent signal efficiently. (*A*) Schematic diagram of the use of TCEP to extinguish the fluorescence signal via cleavage of a disulfide bond linking a fluorescent dye to a readout probe. (*B*) Images of a region of a human fibroblast (IMR-90) stained with an encoding probe for the *FLNA* RNA and a readout probe linked to Cy5 via a disulfide bond as a function of time exposed to 50 mM TCEP. Each panel represents the same portion of an FOV. (Scale bars: 2 μm.) Except for the upper left panel, the contrast has been increased fivefold to illustrate better the fluorescent signal remaining in the sample after TCEP treatment. (*C*) The average brightness of readout probe 1 bound to encoding probes targeting *FLNA* (normalized to the brightness before TCEP exposure) as a function of the total time of exposure to 50 mM TCEP. Error bars represent SEM (*n* provided in Fig. S2*B*), and the blue region represents the 95% confidence interval for a fit to an exponential decay. (*D*) The measured half-life for the average brightness when exposed to 50 mM TCEP for four readout probes (1–4), each with a different sequence and linked to either Cy5 (green) or Alexa750 (red). Error bars represent the 95% confidence interval for the fit to an exponential decay shown in *C* for readout probe 1 and in Fig. S2*A* for readout probes 2–4.

SYSTEMS BIOLOGY

spots were reduced by 10^5-fold and 10^4-fold, respectively (Fig. 2*C* and Fig. S2 *A* and *B*). Furthermore, the TCEP treatment did not inhibit the ability of the next round of readout probes to bind to the sample (Fig. S2*C*). Our calculation shows that the use of this chemical approach to remove fluorescence signals between successive rounds of smFISH should reduce measurement time and increase throughput substantially (Fig. 1*B*, purple line).

Next, we reasoned that, without the requirement for high illumination intensities needed for efficient photobleaching, it should be possible to decrease the area-dependent time further by expanding the size of the imaging FOV. To explore this idea, we designed and constructed a microscope that uses a 2,048 × 2,048 pixel, scientific complementary metal-oxide semiconductor (sCMOS) camera in combination with a high numerical aperture (NA = 1.3) and a high-magnification (60×) silicone oil objective (*SI Materials and Methods*). We used a silicone oil objective because we found that it had less field curvature than comparable oil immersion 60× objectives. With this optical configuration, we could image an FOV of 223 × 223 μm, an area~25-fold larger than our previously reported FOV, with an exposure time of 0.5 s. This increase in the size of the FOV should further increase imaging speed, and hence measurement throughput, substantially (Fig. 1*B*, green line).

As a third step to reduce measurement time, we used multicolor imaging. Specifically, we stained the sample simultaneously with two readout probes per hybridization round, each probe conjugated to one of two spectrally distinct dyes, and used two-color imaging to reduce the number of imaging rounds by half, thereby cutting the area-independent time required to stain, wash, and extinguish signals (Fig. 1*B*, blue line). We used Cy5 and Alexa750 dyes because of the low cellular autofluorescence observed in the red and near-infrared spectral ranges. In total, the use of reductive cleavage to extinguish fluorescence signal between successive imaging rounds in combination with the increase in the FOV area and the use of two-color imaging should dramatically reduce the time required to perform MERFISH for a given area and increase the sample area that can be measured in a given time (Fig. 1*B*, blue line vs. red line).

Improving the Performance of MERFISH Measurements. We also made a series of protocol changes aimed at simplifying measurement procedures and improving the robustness of the measurement. First, we found that readout probes can bind to encoding probes at room temperature with rates similar to those observed at 37 °C (Fig. S3*A*). Room-temperature hybridization avoids any variation in measurement results associated with nonuniform sample heating. Second, we shortened readout probes from 30 to 20 nt, allowing us to include more readout sequences on each encoding probe without increasing the total length of the probe. This modification allows us either to increase the brightness of signals from single mRNA molecules by preserving the number of encoding probes per RNA or to achieve the same signal brightness with fewer encoding probes per RNA, allowing shorter RNAs to be targeted. Third, we created readout probes that bind to readout sequences with rates comparable to those of our previous probes but at 10-fold lower concentrations. Specifically, we exploited the published observation (24) that oligonucleotide sequences that contain only three of the four nucleotides have significantly less secondary structure than sequences that use all four nucleotides and thus have faster hybridization rates (Fig. S3*B*). Fourth, we replaced the toxic RNA denaturing agent formamide used in the readout hybridization and wash buffers with nontoxic ethylene carbonate (25); we found that this substitution also moderately increased the rate of readout hybridization (Fig. S3*C*).

We also found that these modified readout probes and readout hybridization protocols improved MERFISH performance by reducing the variance in staining quality among different rounds of readout hybridization as compared with our previous protocols (Fig. S3*D*). Of the multiple changes made above, the modified readout sequences likely account for the majority of this improvement, because we previously have observed that some of the variability across different readout staining rounds (Fig. S3*D*) can be attributed to sequence variations, presumably resulting from unanticipated secondary structures. By design, such secondary structures should be far less likely with the modified readout sequences that use only three of the four nucleotides (24). We anticipate these improvements will increase the accuracy of our MERFISH measurements because lower-quality (or varying-quality) readout hybridizations can result in dim fluorescence signals in some imaging rounds and increase the rate at which readout errors are made.

An Image Analysis Algorithm to Handle High-Throughput MERFISH Data. In parallel, we anticipated that our previous computational methods for MERFISH data analysis (18, 23), which typically required several hours to a day to analyze a single MERFISH dataset, would not be adequate for analyzing the two orders of magnitude higher data volume generated per experiment. Thus, we developed an analysis pipeline capable of handling this drastic increase in imaging throughput (*SI Materials and Methods*). The major advance in this pipeline is the adoption of a pixel-based decoding approach, as opposed to a spot-finding approach, to reduce computation time. Briefly, images of the same FOV from different imaging rounds are registered using images of fiducial beads collected in each round. These images are high-pass filtered to remove background and are deconvolved to sharpen and better resolve closely positioned spots. Previously we observed that signals from the same RNA often varied in position from round to round by ~100 nm (18). Thus, to connect signals from one round to another more accurately, we applied a low-pass filter with a kernel size of 100-nm radius. The intensities of each pixel across all 16 rounds of images then were used to form a 16-dimensional vector, which we normalized to unit amplitude. This vector then was compared with the set of unit vectors defined by all valid barcodes. The pixel was assigned to a given barcode if the Euclidean distance between its normalized intensity vector and the closest barcode vector was less than the distance defined by a single-bit error. Contiguous sets of pixels that matched to the same barcode were combined to form a single detected RNA. Background pixels mistakenly matched to a barcode were identified and removed based on their low brightness and small number of contiguous pixels matched to the same barcode (Fig. S4). With this pipeline, analysis of large MERFISH datasets (~40 mm^2 with ~40,000 human cells) can be completed in 2–3 d using multiple cores on a computer cluster.

High-Throughput MERFISH Measurements of Tens of Thousands of Cells. To demonstrate the substantial increase in imaging throughput made possible by the above advances, we measured 130 RNAs in cultured U-2 OS cells with a previously published 16-bit, modified Hamming distance-4 (MHD4) encoding scheme (18). In this encoding scheme, all barcodes used are separated by a Hamming distance of at least 4, and hence at least four bits must be read incorrectly to change one valid barcode to another. Therefore, every single-bit error produces a barcode uniquely close to a single valid barcode, allowing such errors to be detected and corrected. Two-bit errors also can be detected but are not correctable because the resulting barcode is no longer uniquely close to a single valid barcode. To account further for the fact that it is more likely to miss a hybridization event (1-to-0 error) than to misidentify a background spot as an RNA (0-to-1 error) in smFISH measurements, our MHD4 code contains a constant and relatively low number (four) of "1" bits. This 16-bit MHD4 encoding scheme includes 140 distinct barcodes in total (18). We assigned 130 of these barcodes to different RNA species, leaving 10 barcodes unused to serve as blanks (not corresponding to any RNA) for misidentification controls.

Fig. 3*A* illustrates one such measurement over an area of 3.2 × 6.2 mm. The cells were fixed, permeabilized, and labeled with encoding probes to 130 RNA species. We then performed eight

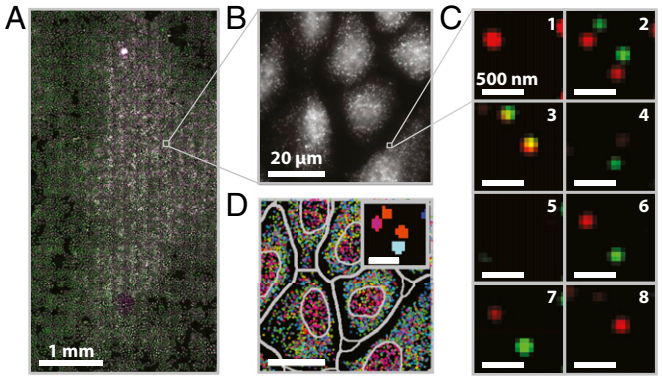

Fig. 3. A MERFISH measurement of an ~20 mm^2 sample area (~15,000 cells). (*A*) Mosaic image of a 3.2 × 6.2 mm region of cultured U-2 OS cells stained with DAPI (purple), encoding probes for 130 RNAs and a Cy5-labeled readout probe (green). (Scale bar: 1 mm.) (*B*) Image of the Cy5 channel in the first round of readout hybridization for the small portion of the field in *A* marked by the gray square. (Scale bar: 20 μm.) (*C*) Two-color images of the smFISH stains for all eight rounds of hybridization and imaging for the small portion of the field in *B* marked by the gray square after the application of a high-pass filter to remove background, deconvolution to tighten spots, and a low-pass filter to connect spots in different images more accurately (*SI Materials and Methods*). Green, red, and orange represent the Cy5 channel, the Alexa750 channel, and the overlay between the two, respectively. (Scale bars: 500 nm.) (*D*) The decoded barcodes for the region shown in *B*. Spots represent individual molecules color-coded based on their RNA species identities (barcodes). Both the nuclear boundaries and the boundaries used to assign RNAs to individual cells are depicted (gray). (Scale bar: 20 μm.) (*Inset*) An image of the barcode assignment (indicated by color) for each pixel in the images shown in *C*. (Scale bar: 500 nm.)

rounds of hybridization, imaging, and TCEP cleavage with 16 different readout probes; each round of imaging used two readout probes conjugated to Cy5 and Alexa750, respectively. Single-molecule spots were clearly observed across the entire imaged area in both Cy5 and Alexa750 channels in each round of smFISH staining and imaging (Fig. 3 *B* and *C*). The identities of individual RNA molecules then were decoded via the algorithm described above (Fig. 3*D*). To assign RNAs to individual cells, we used DAPI to identify cell nuclei and the local density of RNAs to define cellular boundaries (*SI Materials and Methods*). In total, Fig. 3 contains 15,181 cells. Among these, 12,607 segmented cells satisfied our conservative criteria for cell morphology designed to eliminate segmentation errors (*SI Materials and Methods*), and these properly segmented cells contained 9.7 million identified RNA molecules.

To determine the RNA decoding quality, we considered two types of errors for each RNA species. First, some RNAs can be misidentified as the wrong species, leading to a nonzero misidentification rate. Second, some RNAs can be missed, leading to a non-100% calling rate. To assess these errors, we first examined the fraction of decoded RNAs that required error correction (Fig. S5*A*). In our previous published MERFISH experiments using the same 16-bit MHD4 code, we observed that ~60% of all decoded RNAs required error correction (18). By contrast, with the protocols described here, only ~20% of RNAs required correction. Lower levels of error correction would suggest a lower level of misidentification and a higher calling rate. To test the level of misidentification, we examined the number of times that the blank barcodes were counted. Indeed, these barcodes were counted relatively infrequently: 120 of the 130 (92%) RNA species were counted more frequently than the most abundant blank barcode (Fig. 4*A*). In addition, we used an alternative metric, the confidence ratio, to assess the misidentification rate further. As previously defined (18), the confidence ratio for each measured barcode was determined as the number of RNA molecules exactly matching this barcode over the total number of exact matches and matches with single-bit errors for this barcode. We have previously shown that blank barcodes tend to have lower confidence ratio values than RNA-encoding barcodes (18). Indeed, here we found that 95% of the 130 RNA species had a confidence ratio higher than the maximum confidence ratio observed for the blank barcodes (Fig. S5*B*). Next, to examine the calling rate of these measurements, we first used the frequency with which errors were corrected at each bit to determine the average per-bit error rate, as described previously (18). Previously we observed an average 1-to-0 error rate of ~10% and an average 0-to-1 error rate of ~4% (18). By contrast, the MERFISH protocol described here produced substantially lower per-bit error rates, namely, a 1-to-0 error rate of ~1% and a 0-to-1 error rate of ~0.5% (Fig. S5*C*). With these per-bit error rates we would predict a very high calling rate of ~99%. To assess the calling rate experimentally, we determined the copy numbers of 10 different RNAs using conventional smFISH and compared them with our MERFISH results. We found that the average copy number per cell for these 10 RNAs determined with MERFISH correlated strongly with the values determined via smFISH (Fig. 4*B*). Moreover, the average ratio of copy numbers between the MERFISH and smFISH measurements was 0.94 ± 0.06 (SEM; $n = 10$), consistent with the high calling rate estimated from our observed per-bit error rates. Together, these metrics indicate a moderately lower misidentification error rate and higher calling rate than obtained with our previous lower-throughput MERFISH measurements (18).

We further compared the average copy number per cell determined by MERFISH with that determined from published bulk RNA-seq for U-2 OS cells (26). The values determined by MERFISH correlated with those determined from RNA-seq with a high Pearson correlation coefficient for the logarithmic abundances ($\rho_{10} = 0.86$) (Fig. 4*C*).

Finally, to demonstrate the reproducibility of these high-throughput measurements, we performed MERFISH measurements for a range of confluencies of cells and for two different sample areas, ~20 mm^2 and ~40 mm^2. Fig. S6 shows that the average RNA copy number determined by each of these measurements correlated strongly with those determined by the measurement presented in Fig. 3 ($\rho_{10} \geq 0.95$). Across all seven measurements we observed an average calling rate of 90% ± 10% (SEM across seven replicate measurements) by comparison with smFISH results. In total, we measured 105,966 cells with 87,632 cells segmented. The largest of these datasets contained 39,523 cells (35,873 segmented) in an area

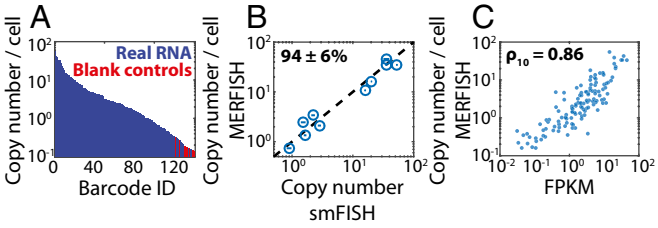

Fig. 4. Performance of the high-throughput MERFISH measurements. (*A*) The average RNA copy numbers per cell measured in Fig. 3 sorted from largest to smallest abundance. Barcodes assigned to real RNAs are marked in blue, and those not assigned to RNAs, i.e., blank controls, are marked in red. (*B*) The average RNA copy numbers per cell determined via MERFISH vs. that determined via conventional smFISH for 10 of the 130 RNAs. The dashed line represents equality. The average ratio of counts determined by MERFISH to that determined by smFISH indicates a calling rate (mean ± SEM) of 94 ± 6% ($n = 10$). Plotted error bars represent the SEM across the number of measured cells (>300 cells) for each gene measured via smFISH. (*C*) The average RNA copy number per cell determined by MERFISH vs. the abundance as determined by bulk sequencing. The Pearson correlation coefficient between the $\log_{10}$ values (ρ_{10}) is 0.86 with a *P* value of 6×10^{-39}. FPKM, fragments per kilobase per million reads.

of 40 mm^2 measured in less than 18 h. This throughput represents a 250-fold increase in the sample area imaged in a single 18-h measurement relative to previously published throughputs (18) and, because the U-2 OS cells used here are smaller than the IMR-90 cells used previously, a nearly 400-fold increase in the number of measured cells.

Characterization of a Subpopulation of Cells. One advantage of the significantly enhanced throughput is the ability to image potentially rare or transient subpopulations of cells with sufficient statistics to characterize the properties of such subpopulations. As a simple illustration of this ability, we identified a subpopulation of cells undergoing DNA replication or cell division in the three datasets collected at the highest confluency (total 78,815 cells). To identify this subpopulation, we determined the distribution of DAPI signal intensity observed in individual cells (Fig. 5*A*). A local minimum in this distribution divided the cells into two groups: group 1 cells contained lower DAPI levels, and group 2 cells contained roughly twice the DAPI signal of group 1 cells, suggesting that group 2 contained cells undergoing DNA replication or cell division. Group 2 represented a relatively small population of ∼20% of the measured cells; nonetheless, because of the large number of cells measured, this population contained 16,036 cells. To identify how the transcriptional profile of these 130 genes differed between group 2 and group 1, we determined for each gene a fractional expression level defined as the copy number of this RNA divided by the total copy number of all 130 RNAs detected in the cell. Fig. 5*B* displays the ratio of this fractional expression level in group 1 vs. group 2 cells for each gene, showing that some genes were up-regulated and some were down-regulated in group 2 cells. The large number of cells measured here allowed us to distinguish even small changes in expression levels with confidence. Fig. 5*C* plots the observed distribution of expression levels for both groups for the 10 most up-regulated (Fig. 5*C*, *Upper*) and 10 most down-regulated (Fig. 5*C*, *Lower*) genes. The most up-regulated genes included the centromere-binding protein CENPF, the spindle-binding protein CKAP5, the DNA polymerase POLQ, and the mitotic checkpoint protein BUB2, supporting the association of group 2 with cells undergoing DNA replication or cell division. Interestingly, the expression of these genes, in particular *CKAP5* and *CENPF*, also could be used to identify this subpopulation of cells without the DAPI signal information. The set of the most down-regulated genes included thrombosin (THBS1), fibrillin (FBN2), and tetraspanin (TSPAN3) as well as other genes involved in cell–cell interactions and adhesion. We speculate that the differential regulation of these proteins might facilitate the disruption and reformation of cell–cell interactions that must occur during cell division.

Finally, to illustrate the power of a spatially resolved measurement, we investigated the spatial distribution of the group 2 cells. To probe this organization, we examined the copy numbers of *CKAP5* and *CENPF*, the two RNAs most up-regulated in group 2 cells (Fig. 5*D*). As expected, we found that the expression levels of these RNAs were highly correlated and varied significantly among cells. Moreover, Fig. 5 *D* and *E* reveals that neighboring cells tended to express a similar level of these RNAs. Such spatial correlations could have been caused by a variety of potential mechanisms, e.g., neighboring cells likely share a common progenitor, resulting in an apparent synchronization of their cell cycles, or there may have been local cues that promoted or repressed cell division. The ability to reveal these cellular-scale spatial organizations directly is one of the benefits of an image-based approach to single-cell transcriptomics.

Discussion

Image-based approaches to single-cell RNA profiling, which identify RNAs via multiplexed smFISH (18, 27–31) or in situ sequencing (21, 22), can directly provide the native spatial context of individual RNAs both within cells and within the context of the culture or tissue. Recently we introduced MERFISH, which uses massively multiplexed smFISH to perform spatially resolved RNA profiling of single cells at the transcriptomic scale (18). However, the measurement throughput of these image-based approaches (i.e., the number of measured cells) has been relatively limited. Here we describe several advances in the MERFISH method that increase the throughput of this approach by two orders of magnitude: We profiled 130 RNAs across 40 mm^2 of sample containing as many as 39,000 human cells

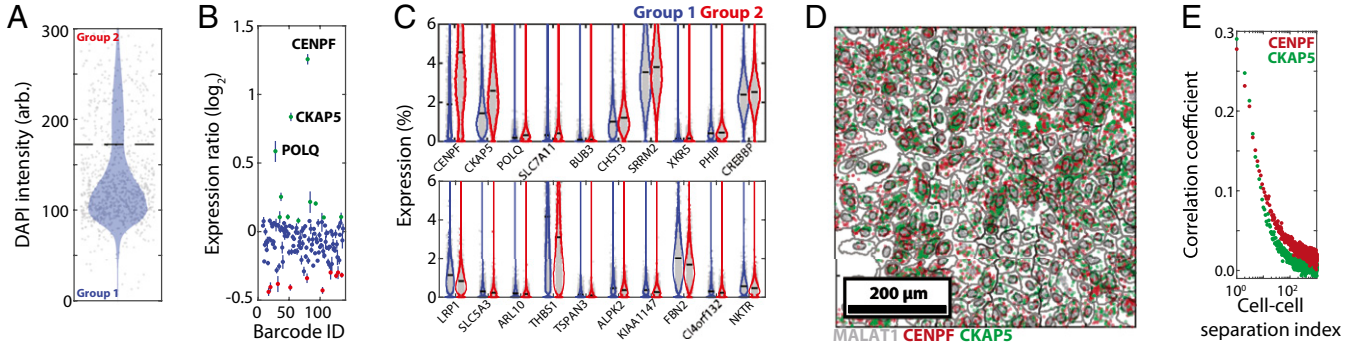

Fig. 5. Characterizing the expression differences of a subpopulation of cells undergoing DNA replication or cell division. (*A*) Violin plot of the distribution of total DAPI intensity for individual cells. The dashed line defines the intensity threshold (based on a local minimum) used to group cells with low DAPI signal into group 1 and cells with high DAPI signal into group 2. Gray dots indicate the values for individual cells, and the blue-shaded area represents the probability distributions. For clarity, only 1,000 randomly selected cells are displayed. (*B*) The $\log_2$ ratio of the mean fractional expression level of each RNA species in group 2 relative to that of group 1. The fractional expression level for an RNA species is defined as the copy number of that RNA divided by the total copy number of all 130 RNAs detected in the cell. The mean and SEM are computed across three biologic replicates. Green and red markers indicate genes further examined in *C*. (*C*) Violin plots of the distribution of expression levels for individual genes within group 1 (blue) or group 2 (red) for the 10 genes with the largest magnitude of up-regulation (*Upper*; marked green in *B*) or the 10 genes with the largest magnitude of down-regulation (*Lower*; marked red in *B*) in group 2 relative to group 1. The solid black lines represent the mean, and the colored curves represent the probability distributions. The gray dots represent the expression levels for 1,000 randomly selected cells. (*D*) A small region of one dataset showing the location of *MALAT1* (gray), *CENPF* (red), and *CKAP5* (green). The gray lines represent the boundaries of cells (segmented based on the density profile of all 130 measured RNAs). Note that *MALAT1* clearly defines the nucleus. (*E*) The Pearson correlation coefficient for the relative expression of *CENPF* (red) or *CKAP5* (green) observed between pairs of cells separated by various distances. The cell–cell separation index is defined as 1 for any given cell and its nearest neighbor, 2 for any given cell and its second closest neighbor, and so forth. These correlations were calculated for all cells within each of the three datasets and then were averaged across these datasets.

in only 18 h. In total, we performed such measurements in ~100,000 cells, generating a dataset comparable in size to those published using droplet-based single-cell sequencing approaches (14, 15). Previously, using a very similar experimental procedure but different encoding schemes, we have shown that MERFISH can be used to measure ~1,000 RNA species in individual cells (18). Thus we anticipate that this increase in throughput could be applied to the measurement of thousands of RNAs with MERFISH.

This substantial increase in throughput should extend the range of questions that can be addressed via MERFISH. For example, we demonstrate here the ability to identify a subpopulation of cells and to use the sizeable number of cells within this subpopulation to quantify the potentially small differences in their gene-expression profiles with statistical significance. We also envision that the increase in imaging throughput reported here will be instrumental in applying MERFISH to the de novo identification of cell types in sizable volumes of tissues. Finally, we anticipate that with further optimization of the hybridization protocol, utilization of faster fluorescence signal removal protocols, incorporation of more colors per imaging round, and additional improvements in camera, optics, and light sources to increase the FOV area and reduce the imaging time further, it will be possible to increase the throughput of MERFISH further and to characterize millions of individual cells in their native culture and tissue contexts. Given that the MERFISH experimental setup is, at its core, a simple fluorescence microscope with a sensitive camera in combination with an automated fluid handling system composed of commercially available components and controlled by open-source software (18, 23), we anticipate that this technique can be readily adopted by many laboratories.

Materials and Methods

Detailed protocols for all methods used in this work can be found in *SI Materials and Methods*. All software is available upon request.

Human U-S OS cells (American Type Culture Collection, ATCC) or human fibroblasts (IMR-90; ATCC) were fixed, permeabilized, stained with encoding probes, and coated with fiducial beads as described previously (18, 23). MERFISH imaging was done on a custom, high-throughput imaging platform built around an Olympus IX71 body, a 60× silicone oil 1.3 NA Plan Apochromat objective (UPLSAPO 60XS2; Olympus), and an sCMOS camera (Zyla 4.2; Andor). Automated fluid handling and sequential staining with readout probes were performed as described previously (18, 23) with the notable exception that the readout hybridization and wash buffers contained ethylene carbonate (E26258; Sigma-Aldrich) instead of formamide.

We created the high-diversity encoding probes by adopting and modifying the Oligopaint approach (32) with a high-yield enzymatic amplification protocol (18, 23) and a high-speed probe-design algorithm. The targeting regions of encoding probes were designed using the human transcriptome (hg38) sequences downloaded from Ensembl, published RNA abundances (26), and a custom probe-design algorithm and computational pipeline that selects target regions based on a narrow range of melting temperature (66–76 °C), GC content (43–63%), and a series of penalties associated with the presence of short homology regions within alternative isoforms of the same gene, all other genes, and abundant noncoding RNAs. A library of template oligonucleotides for making encoding probes was ordered from CustomArray (Dataset S1). Encoding probes were amplified from this library via a high-yield protocol as described previously (18, 23) with minor adjustments to nucleotide concentrations.

We generated the sequences of readout probes randomly, with an A/T probability of 25% and a C probability of 50%, and probes with significant homology to the human transcriptome, as determined via BLAST (33), were removed. Readout probes (Table S1) were purchased from Bio-Synthesis, Inc.

ACKNOWLEDGMENTS. We thank Alistair Boettiger, Yaron Sigal, George Emanuel, Bryan Harada, Pallav Kosuri, and Tian Lu for helpful discussions. This work was supported in part by the Howard Hughes Medical Institute (HHMI). X.Z. is an HHMI investigator.

1. Sandberg R (2014) Entering the era of single-cell transcriptomics in biology and medicine. *Nat Methods* 11(1):22–24.
2. Eberwine J, Sul J-Y, Bartfai T, Kim J (2014) The promise of single-cell sequencing. *Nat Methods* 11(1):25–27.
3. Shapiro E, Biezuner T, Linnarsson S (2013) Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat Rev Genet* 14(9):618–630.
4. Shalek AK, et al. (2013) Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* 498(7453):236–240.
5. Shalek AK, et al. (2014) Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* 510(7505):363–369.
6. Jaitin DA, et al. (2014) Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* 343(6172):776–779.
7. Treutlein B, et al. (2014) Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* 509(7500):371–375.
8. Patel AP, et al. (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 344(6190):1396–1401.
9. Hashimshony T, Feder M, Levin M, Hall BK, Yanai I (2015) Spatiotemporal transcriptomics reveals the evolutionary history of the endoderm germ layer. *Nature* 519(7542):219–222.
10. Achim K, et al. (2015) High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat Biotechnol* 33(5):503–509.
11. Satija R, Farrell JA, Gennert D, Schier AF, Regev A (2015) Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 33(5):495–502.
12. Zeisel A, et al. (2015) Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347(6226):1138–1142.
13. Petropoulos S, et al. (2016) Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell* 165(4):1012–1026.
14. Klein AM, et al. (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161(5):1187–1201.
15. Macosko EZ, et al. (2015) Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 161(5):1202–1214.
16. Crosetto N, Bienko M, van Oudenaarden A (2015) Spatially resolved transcriptomics and beyond. *Nat Rev Genet* 16(1):57–66.
17. Buxbaum AR, Haimovich G, Singer RH (2015) In the right place at the right time: Visualizing and understanding mRNA localization. *Nat Rev Mol Cell Biol* 16(2):95–109.
18. 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.
19. Femino AM, Fay FS, Fogarty K, Singer RH (1998) Visualization of single RNA transcripts in situ. *Science* 280(5363):585–590.
20. Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S (2008) Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods* 5(10):877–879.
21. Ke R, et al. (2013) In situ sequencing for RNA analysis in preserved tissue and cells. *Nat Methods* 10(9):857–860.
22. Lee JH, et al. (2014) Highly multiplexed subcellular RNA sequencing in situ. *Science* 343(6177):1360–1363.
23. Moffitt JR, Zhuang X (2016) RNA imaging with multiplexed error-robust fluorescence in situ hybridization (MERFISH). *Methods Enzymol* 572:1–49.
24. Zhang Z, Revyakin A, Grimm JB, Lavis LD, Tjian R (2014) Single-molecule tracking of the transcription cycle by sub-second RNA detection. *eLife* 3:e01775.
25. Matthiesen SH, Hansen CM (2012) Fast and non-toxic in situ hybridization without blocking of repetitive sequences. *PLoS One* 7(7):e40675.
26. Walz S, et al. (2014) Activation and repression by oncogenic MYC shape tumour-specific gene expression profiles. *Nature* 511(7510):483–487.
27. Levsky JM, Shenoy SM, Pezo RC, Singer RH (2002) Single-cell gene expression profiling. *Science* 297(5582):836–840.
28. Lubeck E, Cai L (2012) Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nat Methods* 9(7):743–748.
29. Levesque MJ, Raj A (2013) Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation. *Nat Methods* 10(3):246–248.
30. Jakt LM, Moriwaki S, Nishikawa S (2013) A continuum of transcriptional identities visualized by combinatorial fluorescent in situ hybridization. *Development* 140(1):216–225.
31. Lubeck E, Coskun AF, Zhiyentayev T, Ahmad M, Cai L (2014) Single-cell in situ RNA profiling by sequential hybridization. *Nat Methods* 11(4):360–361.
32. Beliveau BJ, et al. (2012) Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. *Proc Natl Acad Sci USA* 109(52):21301–21306.
33. Camacho C, et al. (2009) BLAST+: Architecture and applications. *BMC Bioinformatics* 10(1):421.
34. Rouillard J-M, Zuker M, Gulari E (2003) OligoArray 2.0: Design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Res* 31(12):3057–3062.
35. SantaLucia J, Jr, Hicks D (2004) The thermodynamics of DNA structural motifs. *Annu Rev Biophys Biomol Struct* 33(1):415–440.
36. Xu Q, Schlabach MR, Hannon GJ, Elledge SJ (2009) Design of 240,000 orthogonal 25mer DNA barcode probes. *Proc Natl Acad Sci USA* 106(7):2289–2294.
37. Babcock H, Sigal YM, Zhuang X (2012) A high-density 3D localization algorithm for stochastic optical reconstruction microscopy. *Opt Nanoscopy* 1(6):6.
38. Trapnell C, et al. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7(3):562–578.