PNAS
www.pnas.org

Supplementary Information for

Imaging-based pooled CRISPR screening reveals regulators of lncRNA localization

Chong Wang[a,b,c,1], Tian Lu[a,b,c,1], George Emanuel[a,b,c], Hazen P. Babcock[d], Xiaowei Zhuang[a,b,c,2]
[a] Howard Hughes Medical Institute, Harvard University, Cambridge, Massachusetts 02138, USA
[b] Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, USA
[c] Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA
[d] Center for Advanced Imaging, Harvard University, Cambridge, Massachusetts 02138, USA
[1] These authors contributed equally to the work
Correspondence: zhuang@chemistry.harvard.edu

This PDF file includes:

SI Materials and Methods
Figs. S1 to S5
References for SI reference citations

Other supplementary materials for this manuscript include the following:

Datasets S1 to S5

SI Materials and Methods

Cell culture. U-2 OS cells were cultured in EMEM medium (ATCC, HTB-96) supplemented with 10% FBS (Sigma, F4135-1L) and 1% Pen/Strep (Invitrogen, 15140122) antibiotics at 37 ℃. U-2 OS cells stably expressing Cas9-BFP were generated through lentivirus transduction followed by FACS sorting using BFP signal. To generate the lentivirus vector for Cas9-BFP, the Cas9-BFP sequence was PCR amplified from pLentiCas9-BFP (Addgene #78545) and cloned into pFUGW backbone with SVVF promoter. Two nucleus localization signal sequences were added to enhance the nucleus localization of Cas9.

Vector library cloning. The 12-digit barcodes were each comprised of twelve 30-nt sequences, each of the 30-nt sequence representing a trit, with a nucleotide 'A' separating adjacent trits. Oligos encoding each pair of adjacent 30-nt sequences were ordered from IDT in forward and reverse direction alternatively (i.e. trit1 + trit2, trit2 + trit3 reverse complement, trit3 + trit4, trit4 + trit5 reverse complement, …., trit9 + trit10, trit10 + trit11 reverse complement, trit11 + trit12, see Dataset S4). Each trit has three different values, represented by three distinct 30-nt sequences, hence each pair of adjacent 30-nt sequences has 9 possible different sequences, and total 11 x 9 = 99 oligos were needed to cover all possible pairs of adjacent 30-nt sequences. At both ends of the barcodes (represented by Oligos 1-9 and 91-99), two constant primer binding sequences were added for PCR amplification purpose. The sequences of these 99 oligos are described in Dataset S4. The whole barcode library was assembled by two-step overlapping PCR. First, 12 trits were divided into 3 segments, and each segment was generated by the following reactions: Segment 1. Oligos 1-36 as templates, Oligos 1-9 as forward primers, Oligos 28-36 as reverse primers; Segment 2. Oligos 37-72 as templates, Oligos 37-45 as forward primers, Oligos 64-72 as reverse primers; Segment 3. Oligos 73-99 as templates, Oligos 73-81 as forward primers, Oligo 100 as reverse primers. The three PCR products were gel purified. Then, the three segments were mixed and subjected to overlapping PCR using forward primer Oligo 101 and reverse primer Oligo 102. The reverse primer of this step contained a random sequence region of 20 bases which served as the unimolecular identifier (UMI) for the sequencing step. The sequences of oligos 100-102 are also described in Dataset S4. All PCR reactions were performed using real-time qPCR equipment to monitor the reactions so that the reactions were stopped at log-growth phase to reduce library skewing resulted from PCR bias. The PCR products were assembled into a modified pFUGW backbone through isothermal assembly. The assembled library was electroporated into Endura electrocompetent cells (Lucigen, 60242-2) which were then grown under ampicillin selection overnight to amplify the library. The amplified library was purified by mini prep. This library is named pFUGW_barcodes_UMI_backbone library.
The pFUGW_barcodes_UMI_backbone library was then used to generate a library that additionally contain a reporter gene (Luciferease-mCherry) for barcode imaging. The reporter cassettes containing CMV promoter and the reporter open reading frames were first generated in intermediate vectors. The open reading frames contain a luciferase-mCherry, a 2X HA or 2X Myc tag at the N-terminus, and a nuclear localization signal at the C-terminus. The reporter cassettes were PCR amplified from the intermediate vectors using Oligos 103 and 104 (sequence provided in Dataset S4). The

pFUGW_barcodes_UMI_backbone library was then digested with BstXI and treated with alkaline phosphatase and assembled with the reporter cassettes PCR products using isothermal assembly. The assembled libraries were electroporated into Endura electrocompetent cells which were then grown under ampicillin selection overnight for amplification. The cells were then diluted to include the desired number of constructs in each library. These bottlenecked libraries were then purified by mini prep. These libraries are named reporter gene_barcodes libraries.

Cloning of sgRNA-barcode libraries. We used the following strategy to clone the sgRNA-barcode libraries: first, the protospacer-sgRNA constant region-barcode cassette library was generated through multi-step overlapping PCR; then, this library was inserted into lentiviral vectors with U6 promoter placed downstream of the PPT sequence through isothermal assembly. To generate the protospacer-sgRNA constant region-barcode cassette library, the barcode segments were first generated using similar approaches as described in the "Cloning of barcode libraries for quantification of barcode decoding accuracy" section. The only difference is that the constant (primer) region in the 5' end was changed by substituting Oligos 1-9 with 105-113, since the barcodes were placed immediately adjacent to sgRNAs. The sgRNA constant region was PCR amplified using Oligos 114 and 115. The protospacer libraries were ordered from IDT with constant regions on both side of the protospacer for PCR amplification. In this work, we generated two proto-spacer libraries, one for essential ribosome genes and non-targeting sgRNA controls used to measure the recombination rate between sgRNAs and barcodes (Dataset S1) and the other for targeting genes that potentially regulate RNA localizations in the nucleus (Dataset S2). The protospacer libraries were PCR amplified using Oligos 116 and 117, and the PCR products were gel purified. Then, the proto-spacer, sgRNA constant region and barcodes PCR products were mixed and subjected to overlapping PCR using Oligo 116 and 118 as primers. The reverse primer, Oligo 118, contained a random sequence region of 20 bases which served as the unimolecular identifier (UMI) for the sequencing step. The sequences of oligos 105-118 are also described in Dataset S4. All PCR reactions were performed using real-time qPCR equipment to monitor the reactions so that the reactions were stopped at log growth phase. The PCR products were assembled into a modified pFUGW backbone with U6 promoter placed downstream of the PPT sequence through isothermal assembly. The assembled library was electroporated into Endura electrocompetent cells which were then grown on ampicillin selection plate overnight for amplification. Certain number of colonies (~3800 for essential ribosome gene library and ~2500 for RNA localization screening library) were scrapped off the plate with LB buffer and cultured in 200mL LB buffer overnight. The libraries were purified by maxi prep. These libraries are named sgRNA_barcodes libraries.

Sequencing library preparation and analysis. To determine the identity of the barcodes presented in the library as well as to establish the correspondence between sgRNAs and barcodes, we analyzed the library using high-throughput sequencing. We found that PCR amplification of the barcode region can lead to recombination of the barcodes due to homologous regions among the barcodes. Thus, we used ligation-based approach to install sequencing adaptors to the barcode library. In this approach, the regions subjected to sequencing were digested from the library and then ligated to adaptors using T4 ligase.

To determine the barcodes in the reporter gene-barcodes libraries, the libraries were digested using BstXI and BamHI at 37 ℃ for 2 to 3 hours, and the resulting fragments were purified using Zymo DNA purification kit (ZD4002). To generate adaptors with sticky ends for ligation, Oligos 119 -124 (sequence provided in Dataset S4) were mixed at 0.5 μM each, and subjected to 5 cycle of PCR reaction, and products were purified using Zymo DNA purification kit to produce double-stranded sequences with 5' and 3' adaptors separated by BstXI and BamHI digestion sites. The purified products were digested with BstXI and BamHI for 37℃ for 2 to 3 hours, and then purified. The resulting mixture contained adaptors with sticky ends for ligation and was mixed with the purified library fragment mixtures described above. T4 ligase were added and the reactions were kept in room temperature for 2 to 4 hours. The reaction mixtures were directly subjected to 2% agarose gel electrophoresis and a band corresponds to a size of ~400 bp was excised and purified. The purified DNA samples were used for concentration measurement and high-throughput sequencing using V2-MISeq kit (Illumina, MS-103-1003).
To determine the sgRNA-barcode correspondence of the sgRNA_barcodes libraries, we generated two sequencing libraries because the length from proto-spacer to the end of the barcodes is longer than 500 bps, which exceeds the length range optimal for high quality sequencing by the V2-MISeq kit. In one library, the ligation sites were generated using BstXI and BamHI, which located 5' to the proto-spacer and 3' to the UMI, respectively. Sequencing of this library covered the proto-spacer region, part of the barcode region and the UMI region because the middle part the barcode could not be reached by sequencing from either end for this library. In the second library, the ligation sites were generated using KpnI and BamHI (the KpnI site was placed right after sgRNA and before barcodes). Sequencing of this library covered the whole barcode region as well as the UMI region. UMI sequences were used to identify the proto-spacer and barcode in the same construct from these two libraries. Oligos 125-130 and Oligos 131-136 (sequence provided in Dataset S4) were used to generate adaptors for the first and the second library, respectively. The procedures were the same as described for generating sequencing libraries for reporter gene-barcode libraries.
UMI, protospacers and barcode sequences were extracted from sequencing reads. The reads were then grouped by common UMI and barcode to generate a codebook for sgRNA and barcode correspondence. The reads with incorrect protospacers or with barcodes assigned to multiple sgRNAs were excluded from further analysis.
In order to use sequencing to determine the distribution of protospacers and UMIs from cell populations at different time points after lentivirus transduction in the experiments to determine recombination rates, the sequencing libraries were prepared by PCR amplification from purified genomic DNA using Oligos 137-140 and Oligos 141-144 (sequence provided in Dataset S4) as forward and reverse primers, respectively.

Lentivirus production and transduction. Lentivirus were produced in LentiX cells (Takara, 632180) using Lenti-X™ Packaging Single Shots (VSV-G) (Takara, 631276). The produced viruses were concentrated using Lenti-X™ Concentrator (Takara, 631231) and stored at -80 ℃. For transfection, the amount of virus was controlled so that 10-30% of the cells were transduced to ensure most infected cells were infected by only 1 virus particle. The virus transductions were performed using 10 μg/mL polybrene (Sigma, TR-1003-G). The virus titer for the construct with U6-sgRNA-barcode array placed after PPT

did not show obvious reduction compared to that for the construct without insertion after PPT, indicating that the insertion did not impair the lentivirus transduction.

siRNA knockdown. All siRNAs were purchased from Dharmacon, and siRNA knockdown was performed according to Dharmacon's protocol. Briefly, U-2 OS cells were plated on imaging coverslips in 12-well plate at 30,000 cells per well. For siRNA transfection, 1.5 μL 20 μM siRNA was added in 100 μL serum-free, antibiotics-free medium in one tube and 1 μL Dharmacon transfection reagent (Dharmacon, T-2001-01) was added in 100 μL serum-free medium in a separate tube. Two tubes were incubated for 5 minutes and then mixed gently to incubate for another 20 minutes at room temperature. 800 μL antibiotics-free medium with serum was mixed into the 200 μL siRNA and transfection reagent mix described above to generate the 1 mL transfection medium. Cell culture medium was replaced with the 1 mL transfection medium. The cells were incubated at 37℃ for 72 hours before phenotype measurements.

Imaging coverslip silanization. Imaging coverslips were first cleaned by 1M KOH and pure methanol, washed by 70% ethanol and dried in the oven. For silanization, coverslips were covered in silanization buffer (500 mL distilled water, 1500 μL Bind-silane (Sigma, GE17-1330-01) and pH adjusted to 3.5 by glacier acetic acid) for an hour at room temperature. The coverslips were then washed by water and dried to store. Before plating the cells, the silanized coverslips were coated by 1% poly-D-lysine (Sigma, P0899) in 60 mm diameter cell-culture dishes for 30 min followed by a single one-hour wash with water.

Imaging sample preparation. U-2 OS cells were plated on the coverslips two days before fixation. For phenotype imaging in the experiments that screen for factors involved in regulating nuclear RNA localization, U-2 OS cells were fixed 6 days after lentivirus transduction. The samples were fixed by 4% paraformaldehyde (EMS,15714) in PBS for 15 min and permeabilized in 0.5% Tritton-X (Sigma, X100) for 30 mins. Next, samples were incubated in block buffer (500 μL block buffer: 50 μL 10x PBS, 200 μL RNAse free BSA (ThermoFisher, AM2618), 50 μL 25 mg/ml yeast tRNA (ThermoFisher, 15401029), 5 μL Murine RNAse inhibitor (NEB, M0314L), 1 μL 25% Triton-X and Rnase-free water to 500 μL) for one hour and stained with 1:100 primary antibody, anti-SON (Abcam, ab121759), in block buffer for one hour at room temperature. The samples were washed three times with 1xPBS and incubated with 1:300 oligonucleotide-labeled secondary antibody for one hour. The oligonucleotide-labeled secondary antibody can be later probed by readout probes with sequence complementary to the oligonucleotide sequence on the antibody. The samples were washed three times with 1xPBS and post-fixed with 4% PFA for 30 minutes. Then the samples were equilibrated in 30% formamide in 2x SSC for 5 minutes before FISH staining. The FISH hybridization buffer contains 30% formamide (ThermoFisher, AM9342), 60% stellaris RNA FISH hybridization buffer (Biosearch, SMF-HB1-10), 10% 25 mg/mL Yeast tRNA and 1:100 murine RNase inhibitor. The samples were stained with 300 nM FISH probes for the reporter gene, 300 nM FISH probes for RNA phenotype (i.e, 6 RNA species) imaging, and 100 nM primary amplification probes for barcode imaging at 37 ℃ overnight. The FISH probes for the reporter gene each contained a 30-nt targeting sequence that can bind to the reporter gene mRNA and three 20-nt readout sequence that allows the binding of complementary, fluorescently labeled

readout probes. The FISH probes for each RNA target in phenotype imaging each contained a 30-nt targeting sequence that can bind to the RNA target and one or two 20-nt readout sequences that allows the binding of complementary, fluorescently labeled readout probes. Each primary amplification probe for barcode imaging contained a 30-nt targeting sequence that can bind to one of the 30-nt trit sequence on the barcodes, as well as four additional 30-nt identical sequences that allows the binding of secondary amplification probes (Fig. 1A). Then the samples were washed in 30% formamide in 2x SSC twice and stained with 100 nM secondary amplification probes for barcode imaging in 10% hybridization buffer (10% formamide, 80% stellaris RNA FISH hybridization buffer, 10% 25 mg/mL Yeast tRNA and 1:100 murine RNase inhibitor) for an hour at 37 ℃. Each secondary amplification probe contained a 30-nt targeting sequence that can bind to the primary amplification probes, and four additional 20-nt identical readout sequences that allows the binding of complementary, fluorescently labeled readout probes. This amplification scheme thus allows a maximum of 16-fold signal amplification. The samples labeled with FISH probes for phenotype imaging and reporter gene mRNA imaging, and primary and secondary amplification probes for barcode imaging were washed twice in 30% formamide in 2x SSC, and then embedded in 4% polyacrylamide gel, followed by incubation with protein digestion buffer (for 50 mL digestion buffer: 5 mL 8M Guanidine-HCL (ThermoFisher, 24115), 2.5 mL 1 M Tris pH 8.0 (ThermoFisher, 15569025), 100 μL 0.5 M EDTA (ThermoFisher,15575020), 0.25 mL Triton-X and 1:100 proteinase K (ThermoFisher, AM2548)) at 37 ℃ overnight to remove proteins and lipids from the sample. We refer to this step as the sample clearing step below. The protease K cleavage led to protein digestion (including the digestion of mCherry protein), and therefore the mCherry fluorescence signal was eliminated after digestion and did not interfere with FISH signal detection using 561 nm channel. The FISH probes for polyA-containing RNAs, 7SK, MRP, U2 snRNA, and the oligonucleotides linked to the secondary antibody for SON staining were conjugated with acrydite, which can crosslink to the polyacrylamide gel and retain these probes as well as their bound RNA within the gel during the sample clearing step. The FISH probes for MALAT1 and pre-ribosome were not labeled by acrydite because both MALAT1 and pre-ribosome are large in size and thus were retained in the gel during sample clearing. The reporter gene mRNAs were linked to the gel through the acrydite labeled poly T probes that can bind to the poly A tails of the reporter mRNAs, thereby allowing the the FISH probes for the reporter gene and the FISH probes for barcode imaging to be retained in the the gel during clearing. The sample clearing step substantially reduces background signal due to cell autofluorescence and nonspecific binding of FISH probes to proteins and lipids (1). The samples were then washed by 2x SSC and left in 2x SSC for imaging. Sequences for used FISH probes are listed in Dataset S5.

For experiments that were used to measure the barcoding identification error using two known phenotypes (expression of HA or Myc tagged reporter genes), U-2 OS cells were fixed 6 days after transduction. The tags were stained by primary antibodies (anti-Myc (Abcam, ab9132), anti-HA (Abcam, ab9110)), and then Alexa 405 labeled anti-mouse secondary antibody (Abcam, ab175658) and Alexa 488 labeled anti-rabbit secondary antibody (Invitrogen, A21206). The samples were incubated in 25 mM MA-NHS (Sigma, 730300) in 2x SSC for one hour before gel embedding, therefore, MA-NHS labeled antibodies were linked to the gel during gel polymerization. After sample cleaning, antibodies were digested into fragments and the dyes were linked to gels via crosslinked

antibody fragments (2). The dyes Alexa 405 and Alexa 488 can survive the polymerization reaction during gel embedding (2). The rest of sample preparation including immunostaining and barcode staining is described as above.

Antibody labeling by oligonucleotide. We used the following strategy to label antibodies with oligonucleotide. Antibodies were first mixed with DBCO-NHS which conjugate DBCO to antibodies and the DBCO-labeled antibodies were then mixed with azide-labeled oligonucleotide to conjugate oligonucleotide to antibodies. Specifically, 100μg anti-rabbit antibody (ThermoFisher, 31210) was first buffer exchanged into 100 μL PBS using 50 KD protein concentrator (Millipore, UFC510024). NaHCO3 and DBCO-NHS ester (Kerafast, FCC310) were added into the antibody solution so that their final concentrations were 50 mM and 100 μM, respectively. The reaction was allowed to proceed for 1h at room temperature to make DBCO-labeled antibodies and excess DBCO was removed through buffer exchange with PBS using 50 KD protein concentrator. Then PBS buffer was added to DBCO-labeled antibodies to make the solution volume 100μL and 25 μL azide-labeled oligonucleotide (100 μM, Dataset S3) was added. The reaction was allowed to proceed at 4℃ overnight. After the reaction finished, the excess oligonucleotide was removed through buffer exchange using PBS and the final oligonucleotide-labeled antibody was aliquoted and stored at -80℃.

Imaging setup and sequential imaging. The imaging setup was as described previously (3). Briefly, a peristaltic pump (Gilson, MINIPULS 3) pulled liquid into Bioptech's FCS2 flow chamber with sample coverslips and three valves (Hamilton, MVP and HVXM 8-5) were used to select the input fluid. A custom microscope built around a Nikon Ti-U microscope body with a Nikon CFI Plan Apo Lambda 60× oil immersion objective with 1.4 NA was used for imaging. Solid-state single-mode lasers (405 nm laser, Obis 405 nm LX 200 mW, Coherent; 488 nm laser, Genesis MX488-1000, Coherent; 560 nm laser, 2RU-VFL-P-2000-560-B1R, MPB Communications; 647 nm laser, 2RU-VFL-P-1500-647-B1R, MPB Communication; and 750 nm laser, 2RU-VFL-P-500-750-B1R, MPB Communications) were used for illumination. Acousto-optic tunable filter (AOTF) were used to control the intensities of the 488 nm, 560 nm, and 647 nm lasers; the 405 nm laser was modulated by a direct digital signal; the 750 nm laser were switched by mechanical shutters. A custom dichroic (Chroma, zy405/488/561/647/752RP-UF1) and emission filter (Chroma, ZET405/488/461/647-656/752m) were used to separate the excitation illumination from the fluorescence emission. The emission was imaged onto the Hamamatsu digital CMOS camera. During acquisition, the sample was translated using a motorized XY stage (Ludl, BioPrecision2) and kept in focus using a home-built autofocus system.

Before loading into the flow chamber, the sample was stained an Atto565-labeled, 20-nt readout probe (Dataset S5) which has a sequence complementary to the readout sequence on the FISH probes for reporter gene mRNA imaging. The staining was performed in hybridization buffer (10% ethylene carbonate (Sigma, E26258) in 2x SSC), with a readout probe concentration of 3 nM. The readout probe for the reporter gene was introduced only once but was imaged repetitively during for all hybridization rounds. The readout probes for the 7 molecular targets (SON protein and 6 RNA targets) for phenotype imaging and the readout probes for barcode imaging were introduced in sequential rounds of

hybridizations. For phenotype and barcode imaging, 3 nM 20-nt readout probes (Bio-Synthesis Inc., Dataset S5), complementary to the oligonucleotide sequence on the SON antibody (Abcam, ab121759), or to the readout sequences on the FISH probe for the 6 RNA targets, or to the readout sequences on the secondary amplification probes for barcode imaging, in hybridization buffers (10% ethylene carbonate in 2x SSC) were flowed into the chamber, left for 15 minutes and followed by hybridization buffer wash. The dyes for these probes, Alexa 488, Cy5 or Alexa 750, were linked to the oligos via a cleavable disulfide bond (Biosynthesis, Dataset S5). The sample were imaged in anti-bleach buffer (For 50 mL anti-bleach buffer: 50 mg gluco-oxidase (Sigma, G2133), 50 mg (±)-6-Hydroxy-2,5,7,8-tetramethylchromane-2-carboxylic acid (Trolox) (Sigma, 238813) (4), 300 μL catalase (Sigma, C100-500MG), 10% w/v glucose (Sigma, G8270), 5 mL 500 μM Trolox quinone and 50 μL murine RNase inhibitor). For each round of hybridization, fluorescence signals from four color channels (488 nm, 561 nm, 647 nm and 750 nm, if phenotype imaging was included in the round) or three color channels (561 nm, 647nm and 750 nm, if phenotype imaging was not included in the round) were imaged. After each round, the dyes on the readout probes were cleaved by 10% tris (2-carboxyethyl) phosphine (TCEP; Sigma, 646547-10X1ML), followed by hybridization of the readout probes for next round.

Reporter gene mRNA signal was detected using the 561 nm channel in every round for the sake of quantification of the colocalization ratio between reporter gene signal and barcode signal, and for image registration. The barcode signals were measured through sequential rounds of hybridization and imaging using 647 nm and 750 nm channels with cleavable Cy5 and Alexa 750 dyes in rounds 1-18, which allowed all 36 values of the 12-trit barcodes to be imaged. The signals for SON and 6 RNA targets in phenotype imaging were measured through sequential rounds of hybridization and imaging using the 488 nm channel with cleavable Alexa 488 dye in rounds 1-7. For phenotype imaging, the images were collected at a slightly higher focal plane (2-3 μm) optimal for signals from interior of the nuclei.

For experiments measuring the barcoding identification error using two known phenotypes (expression of HA or Myc tagged reporter genes), Myc and HA tags were stained with Alexa 405-dye and Alexa 488-dye labeled secondary antibodies and imaged in 405 nm and 488 nm channels, respectively.

DAPI staining was imaged and used for cell segmentation and nucleus identification. For experiments measuring two know phenotypes (HA or Myc tagged reporter genes), DAPI staining was imaged at the last round of imaging. For the experiments to screen for factors regulating nuclear RNA localization, DAPI staining was imaged at the first round of imaging.

The sequences for dye labeled readout probes are listed in Dataset S5.

Transcription inhibition. For transcription inhibition, 50 μM DRB (Sigma, D1916-10MG) was mixed in EMEM and incubated with the cells for an hour before fixation.

Barcode decoding analysis. To corrected for non-uniformity in illumination, every image for a give color channel was divided by the mean-intensity image for all images for that illumination color. Images of multiple rounds were registered using uncleavable signals of the reporter gene mRNA. Cells were segmented by watershed algorithm using DAPI staining as seed and cell autofluorescence (for the experiments to evaluate barcode

decoding accuracy and lentivirus recombination) or poly-A containing RNAs staining (for the experiments to screen for factors regulating nuclear RNA localization) for cell boundary identification. Single-molecule signals for reporter gene mRNA and barcodes across all hybridizations were identified using a previously describe spot finding algorithm (5). The single-molecule FISH spots were assigned to cells, and the colocalization ratio for each of the three values of a trit in the barcode was calculated as the number of reporter-gene smFISH spots that were colocalized with barcode smFISH signal divided by total number of reporter-gene smFISH spots within the cells. To determine the value of each trit for each cell, cells were clustered based on the three colocalization ratios of that trit by k-means clustering, and the trit value was assigned to each cluster based on which of the three mean colocalization ratio was the highest for that cluster. The same process was repeated for all 12 trits, so that each cell was assigned a 12-trit barcode. For each value of a trit, the average colocalization ratio for the population of cells assigned that value was measured to be 0.4; whereas the average colocalization ratio due to random colocalization with non-specifically bound probes, assessed from the two populations of cells not assigned that trit value, was measured to be 0.1.

To decode the cells based on the barcode signals alone (i.e. without consideration of the colocalization between barcode signals and reporter gene signals), cells were clustered based on the numbers of barcode-signal spots detected for the three trit values within each cell. A k-means clustering algorithm was used to partition the cells into three populations, and the trit value was assigned to each population based on which one of the three trit values had the highest mean spot numbers. This same process was repeated for all 12 trits. To estimate the barcode misidentification rate, since only 0.4% of all possible barcodes were present in the libraries due to our bottlenecking strategy, the probability that any erroneously decoded barcode would match the barcodes in the libraries is only 0.4%. Thus, among the 57% exact-matched barcodes, only approximately p = 0.3% could arise from barcode misidentification (solving from (p*57%)/(1-57%) = 0.4%/(1-0.4%)).

Myc and HA signal quantification. To quantify the HA and Myc expression in the nucleus, the nuclear boundary of each cell was used as a mask to measure the intensity of the corresponding Myc or HA channel. To allow unambiguous assignment of HA and Myc expression to individual cells, we first determined the threshold values for HA and Myc expression, above which HA or Myc tag expression can be confidently detected. To determine these threshold values, a k-means clustering algorithm was used to cluster the cells into two groups based on their unthresholded HA and Myc tag staining intensity. This grouping allowed approximated separation of cells into HA- and Myc-expressing cells. To estimate the background stain level for the HA tag, the mean and standard deviation of the HA intensity values for cells in the Myc-expressing cluster was calculated and the threshold for HA signal was calculated as mean plus three standard deviations. The threshold value for Myc expression was determined similarly from the HA-expressing cluster. The cells with HA and Myc intensities that were both lower than their respective thresholds or both higher than their respective thresholds were discarded (197 out of 2336 cells). After removing these ambiguous cells, the remaining cells were clustered again using a k-means algorithm to obtain the final grouping as shown in Fig. 2C.

Recombination rate calculation. Calculation of the recombination rate a_i for the *i*th sgRNA (with the barcode or UMI) is based on following:

$$B_{i,day\ n} = P_{i,day\ 2} * \left((1 - a_i) * S_i + a_i * C\right)$$

$$S_i = \frac{P_{i,day\ n}}{P_{i,day\ 2}}$$

$$C = \sum_i \left(\frac{P_{i,day\ 2}}{\sum_j P_{j,day\ 2}} * S_i\right)$$

Where n is the number of days post transduction, which is equal to 21 or 28 in our experiments. $P_{i,day\ 2}$ is the normalized proto-spacer reads determined by sequencing for the *i*th sgRNA on day 2 post transduction (normalized by the total proto-spacer reads measured on day 2 post transduction). $P_{i,day\ n}$ is the normalized proto-spacer reads determined by sequencing for the *i*th sgRNA on day n post transduction (normalized by the total proto-spacer reads measured on day n post transduction). $B_{\ i,day\ n}$ is normalized cell numbers determined by barcode imaging or normalized UMI reads determined by sequencing (normalized by the total cell number or UMI reads on day n post transduction). S_i is the survival rate of the *i*th sgRNA. C is the average survival rate for all sgRNAs within the library, calculated by considering the abundance weight of different sgRNAs in the library, which is the mean survival rate if recombination happens.

Phenotype measurement quantification. Nucleus boundary were determined by DAPI signals. The cells whose nuclei were in contact with the edge of the imaging field-of-view were removed from further analysis. To identify the clusters of MRP, pre-ribosome, and SON, we subtracted the background intensity of the channel and used the functions regionprops (MatLab) and bwareaopen (MatLab) to identify the clusters. In detail, the pixels with intensity larger than a brightness threshold will be selected. The clusters of the selected pixels were identified by the bwareaopen function. The clusters within a bounded area range (20-3000 pixels for SON, 100-5000 pixels for pre-ribosome and 100-6000 pixels for MRP) were kept. The area ranges were determined by visual inspection of the raw image. In order to capture clusters with relatively wide variations in staining levels, this process was iterated using multiple brightness thresholds (from 0.9 x max (pixel intensity in the nucleus) to 0.1 x max (pixel intensity in the nucleus) with the decrement of 0.05 x max (pixel intensity in the nucleus)). For each iteration, lower brightness threshold will identify two types of clusters: (i) the dim clusters that cannot be detected at higher brightness threshold from the previous round and (ii) the larger clusters that completely include one or more clusters identified from previous round. For any cluster of type (i), it was kept only if its area was within the allowed area range described above. For any cluster of type (ii), if its area was within the allowed area range, it was kept; otherwise, it was removed and the smaller cluster(s) identified from the previous round that overlapped with this new cluster was kept instead. The number of the final identified clusters and the area of each cluster were measured using the regionprops function. For MRP, preribosome, and SON, we calculated the number of clusters, the mean area of clusters, and the cluster intensity (defined as the total signal within the cluster boundaries divided by total cluster area) for each cell. To quantify the nuclear speckle enrichment MALAT1, 7SK, U2 and poly-A containing RNAs, cluster boundaries from the SON staining were used as mask to measure the MALAT1, 7SK, U2 and poly-A containing RNAs signals within the SON

cluster boundaries. Nuclear speckle intensity of each of these RNAs was measured as the total signal of the said RNA within the SON cluster boundaries divided by the total area covered by SON clusters. The signal intensity outside the speckle was measured as the total signal of the RNA in the nucleus but outside nuclear speckles divided by the total area of the nucleus that was not in nuclear speckles. The nuclear speckle enrichment was determined as the ratio between the nuclear speckle intensity and the signal intensity outside the speckle.

To identify hits from the screening, we combined four replicates of experiments. The quantified values described above (i.e. cluster number, cluster area, cluster intensity, and nuclear speckle enrichment) of each replicate were normalized by the median values of all cells within each replicate before combination. We used Student's t test to calculate the p value by testing the measured values for the cells harboring one targeting sgRNA against the values measured from cells harboring all control, non-targeting sgRNAs. When at least two sgRNAs targeting a certain gene showed p values <0.05, the gene was listed as a hit. The sgRNAs that had less than 40 cells were removed from analysis.

SI Figures

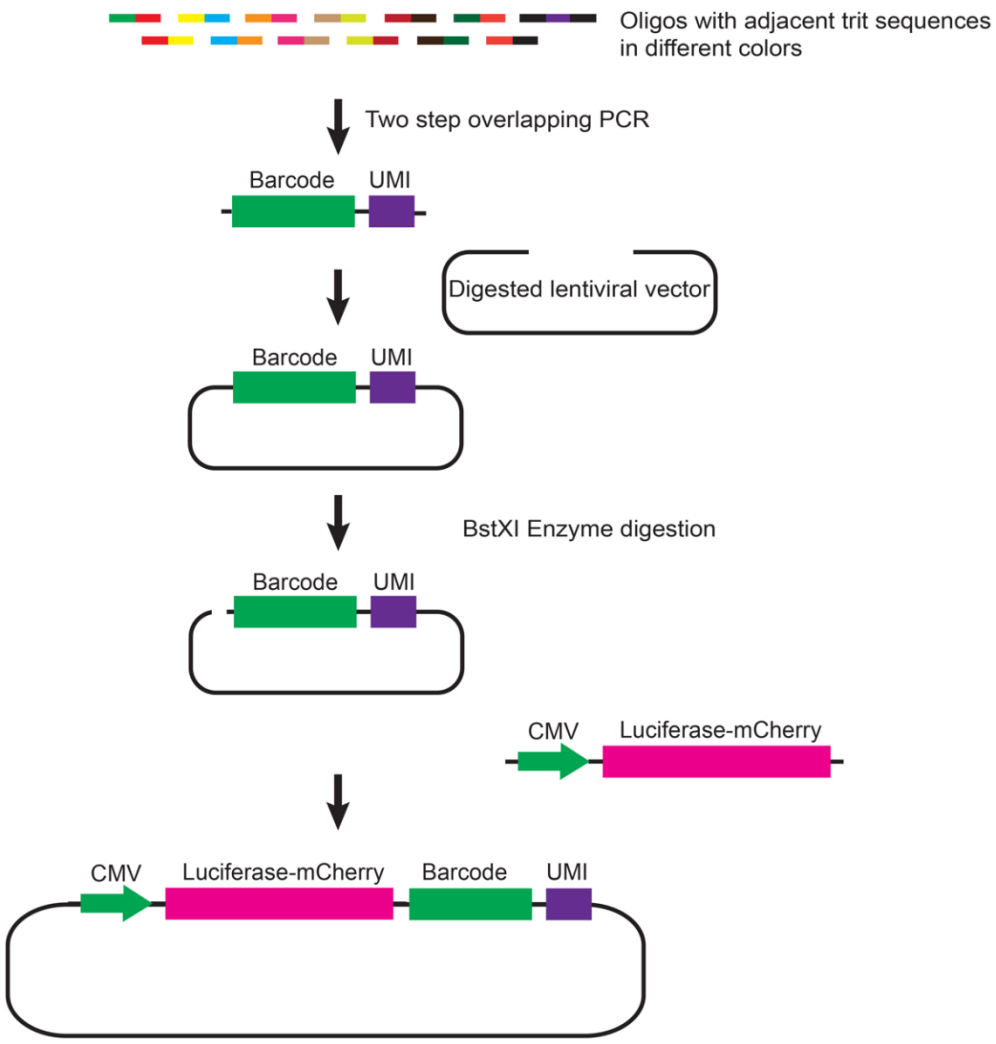

Fig. S1. Cloning strategy of libraries for evaluation of imaging-based barcode detection accuracy. In brief, the barcode and UMI are first assembled from individual pieces of DNA oligos through two-step overlapping PCR (see *SI Materials and Methods*). The colors for different oligos represent different trit sequences. This barcode-UMI library is then inserted into a digested lentiviral plasmid backbone to form a barcode-UMI lentiviral vector library. A reporter gene cassette is further inserted into the barcode-UMI lentiviral vector library to create the final reporter gene-barcode library.

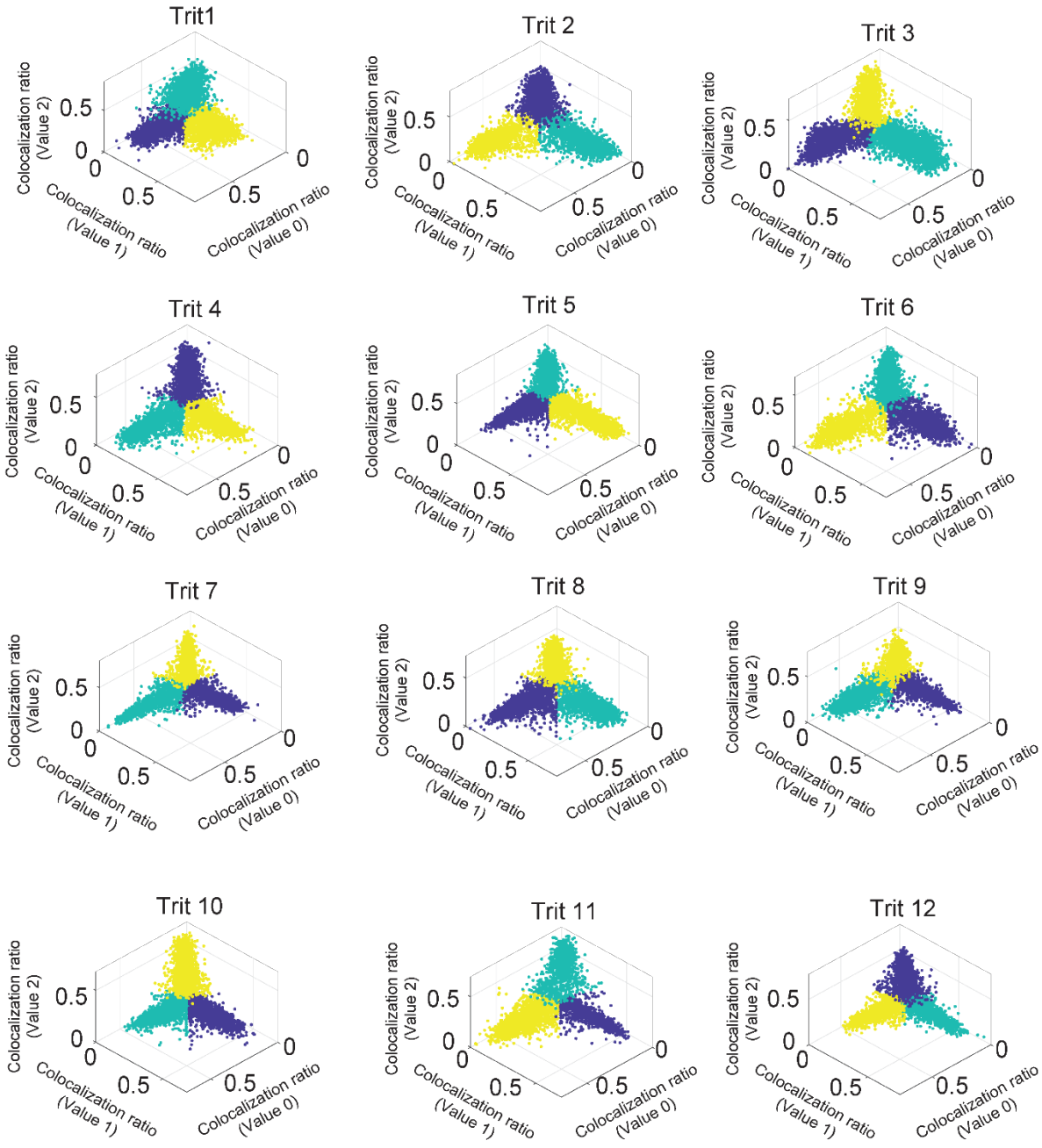

Fig. S2. Colocalization ratio analysis of all 12 trits. Colocalization ratio of the three values of individual trits measured for all cells are displayed for all 12 trits. Cell are partitioned into three clusters (shown in different colors) based on their colocalization ratios using a k-means clustering algorithm. Each cluster corresponds to cells with one trit value.

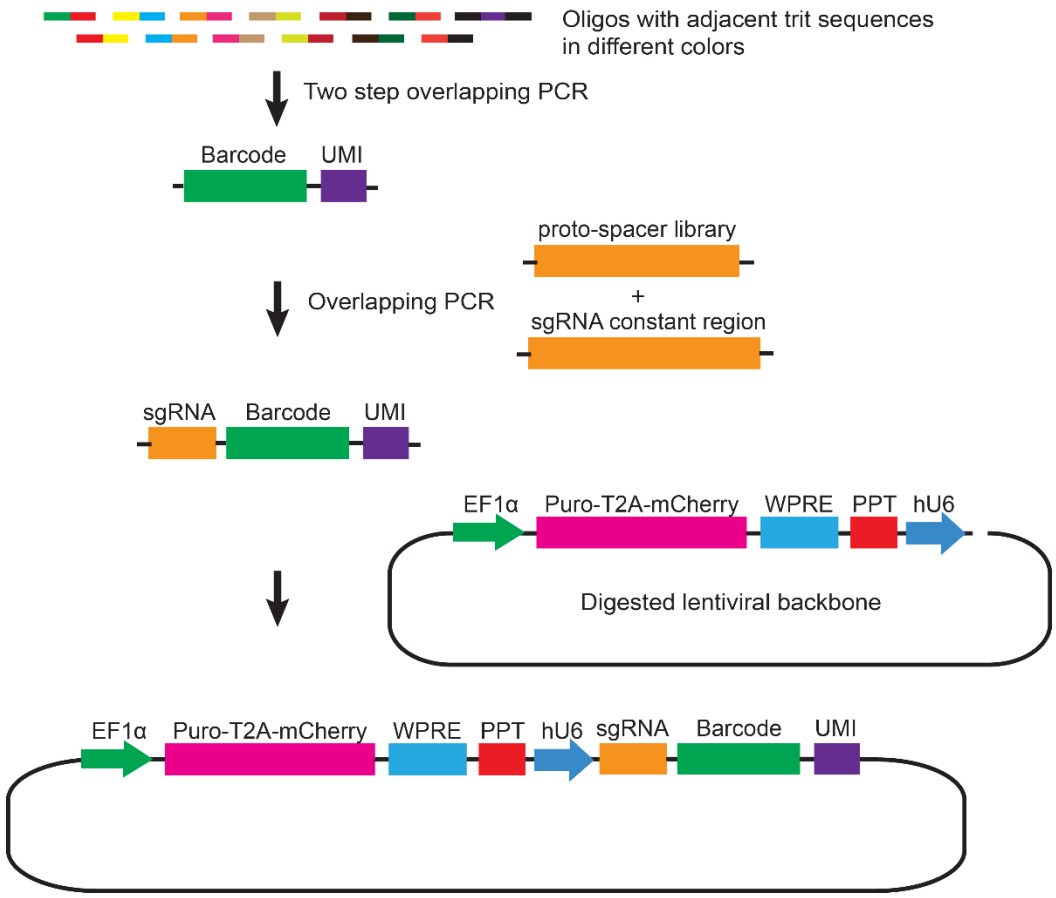

Fig. S3. Cloning strategy of lentiviral sgRNA-barcode delivery library. In brief, the barcode and UMI are first assembled from individual pieces of DNA oligos through two-step overlapping PCR and then assembled with the proto-spacer sequences and sgRNA constant region sequence using overlapping PCR to form a sgRNA-barcode-UMI cassette library. The colors for different oligos represent different trit sequences. This library is then inserted in to a digested, reporter gene containing lentiviral backbone with the hU6 promoter at the site downstream of the poly purine tract (PTT).

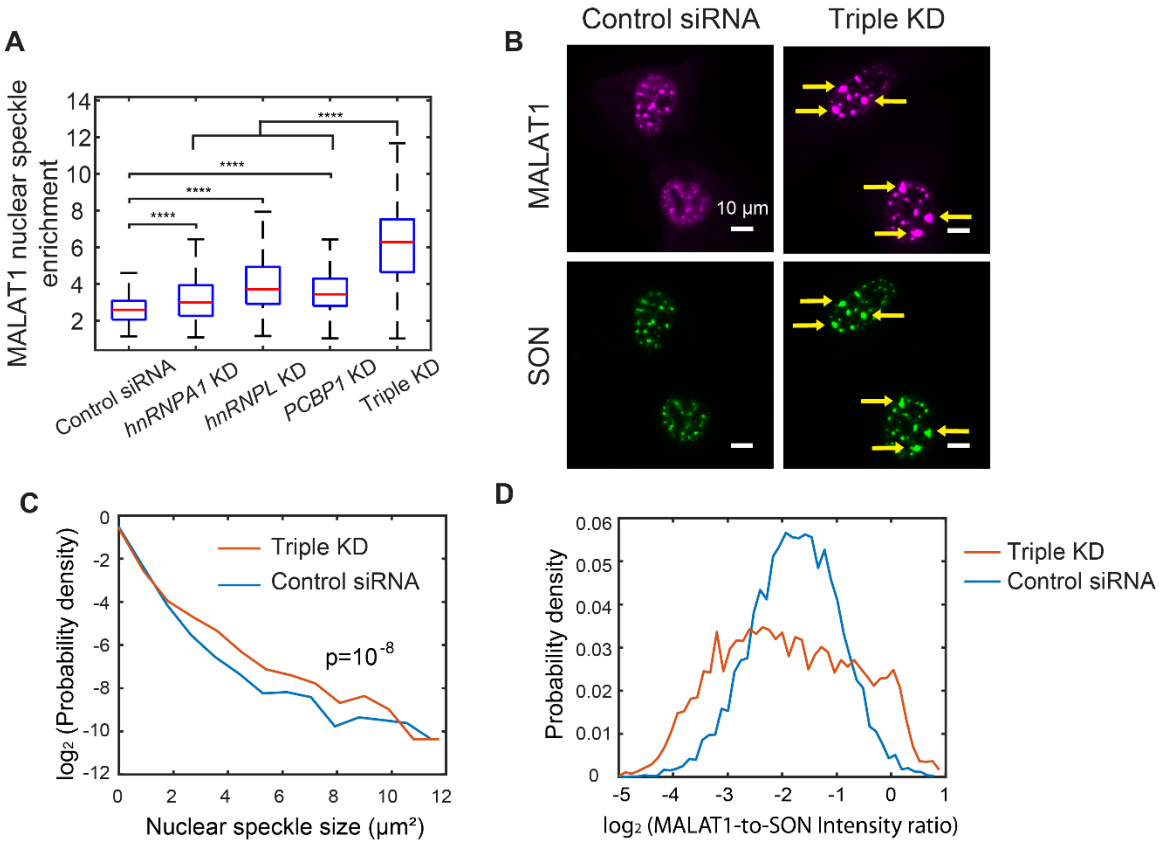

Fig. S4. Triple knockdown of *hnRNPA1*, *hnRNPL* and *PCBP1* affects the morphology and composition of nuclear speckles. **(A)** Boxplots showing the effect of control siRNA and *HNRNPA1*, *HNRNPL*, and *PCBP1* single and triple knockdown (KD) on MALAT1 localization. Boxplot elements are as described in Fig. 5. 100-300 cells are quantified for each condition. Student's t tests are performed between each single KD and the non-targeting control and between the triple KD and the *hnRNPA1*, *hnRNPL* or *PCBP1* single KD. ****, $p<0.0001$. **(B)** Example images for cells showing that some of the MALAT1-positive nuclear speckles are enlarged (highlighted by yellow arrows) after *hnRNPA1*, *hnRNPL* and *PCBP1* triple KD, as compared to the cells transfected with control nontargeting siRNA. Scale bars: 10 μm. **(C)** Distribution of nuclear speckle size shows that triple KD of *hnRNPA1*, *hnRNPL* and *PCBP1* increase the nuclear speckle size. Two-sample Kolmogorov-Smirnov test is used to test the difference between two distributions. **(D)** Distributions of $\log_2$ (MALAT1-to-SON intensity ratio) in each nuclear speckle for control siRNA and *hnRNPA1*, *hnRNPL* and *PCBP1* triple KD samples. ~300 cells and ~7000 speckles are measured for control and triple KD conditions, respectively.

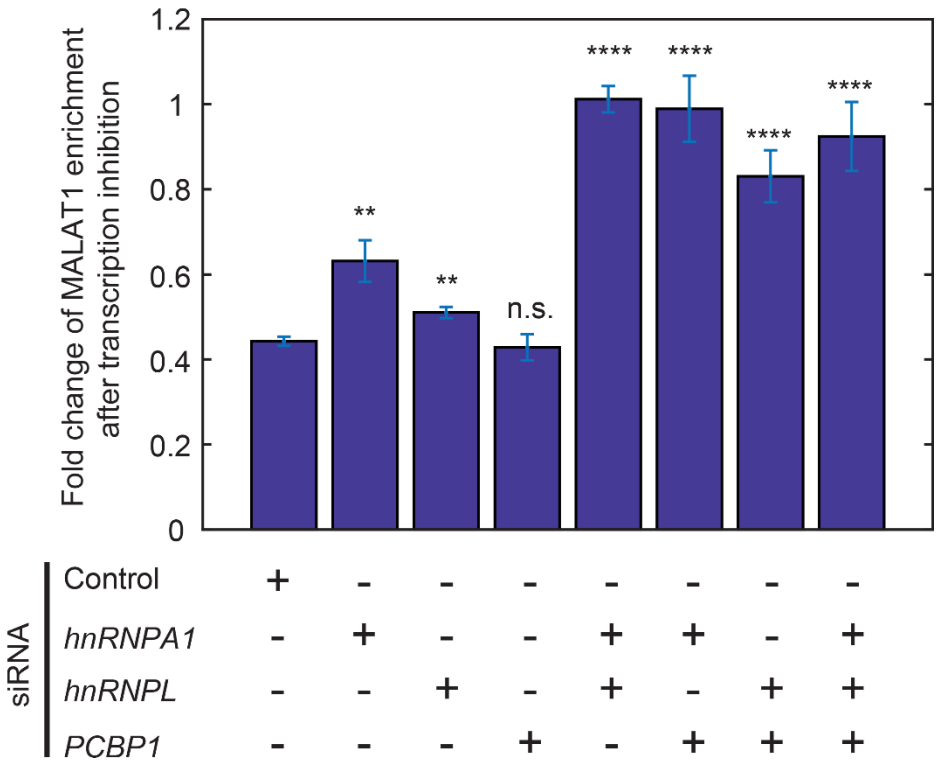

Fig. S5. Fold change of MALAT1 nuclear speckle enrichment after transcription inhibition under different knockdown conditions. Bar chart showing the fold change of MALAT1 nuclear speckle enrichment after transcription inhibition under the respective knockdown conditions. This fold change is defined as MALAT1 nuclear speckle enrichment after transcription inhibition divided by the enrichment before transcription inhibition under the same siRNA treatment. Control siRNA, *hnRNPA1*, *hnRNPL* and *PCBP1* single knockdown as well as *hnRNPA1*, *hnRNPL* and *PCBP1* double and triple knockdown conditions are shown. Error bar: SD. Student's t tests are performed for each condition with respect to control. **, $p<0.01$, ****, $p<0.0001$, n.s., not significant. n=3, each experiment contains 30-100 cells.

SI Datasets

Dataset S1. sgRNA library to evaluate the lentivirus design for reduced recombination effect.
This dataset lists the oligo sequences for the proto-spacers of 159 sgRNAs targeting essential ribosomal genes and 51 non-targeting sgRNAs.

Dataset S2. sgRNA library for genetic screen of factors regulating RNA localization in the nucleus.
This dataset lists the oligo sequences of the proto-spacers of 162 sgRNA targeting selected candidate genes for regulating RNA localization in the nucleus and 5 non-targeting sgRNAs.

Dataset S3. Features quantified and gene hits identified in the screen for factors regulating nuclear RNA localization.
This dataset lists the gene hits for individual phenotype features quantified in the screen.

Dataset S4. DNA oligo sequences used for library cloning and sequencing.
This dataset lists the DNA oligo sequences used for library construction and next-generation sequencing (for the determination of barcode identity and barcode-sgRNA correspondence within the libraries, and for the quantification of proto-spacer and UMI abundance in recombination quantification).

Dataset S5. FISH probe sequences for barcode, reporter gene and phenotype imaging.
This dataset includes the following separate lists of oligonucleotide probes:
The primary amplification probes for barcode imaging
The secondary amplification probes for barcode imaging
The FISH probes for polyA-containing RNA
The FISH probes for MALAT1
The FISH probes for 7SK
The FISH probes for MRP
The FISH probes for pre-ribosome
The FISH probes for U2 snRNA
The FISH probes for the reporter genes Puro-T2A-mCherry or mCherry-luciferase
The oligonucleotide probe attached to antibodies for SON
The readout probes for all of the above targets
Some probes were modified, and the modifications are shown in the probe sequences.

References

1. Moffitt JR*, et al.* (2016) High-performance multiplexed fluorescence in situ hybridization in culture and tissue with matrix imprinting and clearing. *Proc Natl Acad Sci U S A* 113(50):14456-14461.
2. Chozinski TJ*, et al.* (2016) Expansion microscopy with conventional antibodies and fluorescent proteins. *Nat Methods* 13(6):485-488.
3. Emanuel G, Moffitt JR, & Zhuang X (2017) High-throughput, image-based screening of pooled genetic-variant libraries. *Nat Methods* 14(12):1159-1162.
4. Rasnik I, McKinney SA, & Ha T (2006) Nonblinking and long-lasting single-molecule fluorescence imaging. *Nat Methods* 3(11):891-893.
5. Babcock H, Sigal YM, & Zhuang X (2012) A high-density 3D localization algorithm for stochastic optical reconstruction microscopy. *Opt Nanoscopy* 1: 6.