

METHODS

Mammalian cell culture

HEK-293T cells from the ATCC (passages < 25) were cultured in a 1:1 DMEM/MEM mixture (Cellgro) supplemented with 10% fetal bovine serum, 100 units/mL penicillin, and 100 mg/mL streptomycin at 37°C under 5% CO₂ (Hung et al., 2016). Mycoplasma testing was not performed before experiments. For fluorescence microscopy imaging experiments, cells were grown on 7×7-mm glass coverslips in 48-well plates. For qPCR and RNA-seq experiments, cells were grown on 10-cm glass-bottomed Petri dishes (Corning). To improve the adherence of HEK-293T cells, we pretreated glass slides with 50 µg/mL fibronectin (Millipore) for 20 minutes at 37°C before cell plating and washing three times with Dulbecco's PBS (DPBS) (pH 7.4).

Generation of HEK-293T cells stably expressing different APEX2 constructs

APEX2 fusion constructs in [Figure S2A](#) were cloned into pLX304 vector via Gibson assembly. For preparation of lentiviruses, HEK-293T cells in 6-well plates were transfected at ~60%–70% confluency with the lentiviral vector pLX304 containing the gene of interest (1,000 ng), the lentiviral packaging plasmids dR8.91 (900 ng) and pVSV-G (100 ng), and 8 µL of Lipofectamine 2000 for 4 hr (Hung et al., 2016). About 48 hours after transfection the cell medium containing lentivirus was harvested and filtered through a 0.45-mm filter. HEK-293T cells were then infected at ~50% confluency, followed by selection with 8 mg/mL blasticidin in growth medium for 7 days before further analysis.

APEX labeling in living cells

18-24 hours after plating HEK-293T cells stably expressing the corresponding APEX2 fusion construct, APEX labeling was initiated by changing the medium to fresh medium containing 500 µM biotin-phenol (Iris Biotech GmbH). This was incubated at 37°C under 5% CO₂ for 30 minutes. H₂O₂ (Sigma Aldrich) was then added to each well to a final concentration of 1 mM, and the plate gently agitated for 1 minute (Hung et al., 2016). The reaction was quenched by replacing the medium with an equal volume of 5 mM Trolox, 10 mM sodium ascorbate and 10 mM sodium azide in Dulbecco's phosphate-buffered saline (DPBS). Cells were washed with DPBS containing 5 mM Trolox and 10 mM sodium ascorbate three times before proceeding to imaging, RT-qPCR or RNA-seq experiments. The unlabeled controls were processed identically, except that the H₂O₂ addition step was omitted.

Immunofluorescence staining and fluorescence microscopy

Cells were fixed with 4% paraformaldehyde in PBS at room temperature for 10 minutes. Cells were then washed with PBS three times and permeabilized with cold methanol at –20°C for 5 minutes. Cells were washed again three times with PBS and blocked for 1 hour with 3% BSA in PBS (“blocking buffer”) at room temperature. Cells were then incubated with primary antibodies (Mouse anti-V5 antibody, Life Technologies, 1:2000 dilution; Mouse anti-FLAG antibody, Sigma, 1:1000 dilution) in blocking buffer for 1 hour at room temperature. After washing three times with PBS, cells were incubated with secondary antibodies (AlexaFluor488, Invitrogen, 1:1000 dilution; neutravidin-AlexaFluor647, Invitrogen, 1:1000 dilution) in blocking buffer for 30 minutes. Cells were then washed three times with PBS and imaged. Fluorescence confocal microscopy was performed with a Zeiss AxioObserver microscope with 60X oil immersion objectives, outfitted with a Yokogawa spinning disk confocal head, Cascade II:512 camera, a Quad-band notch dichroic mirror (405/488/568/647), and 405 (diode), 491 (DPSS), 561 (DPSS) and 640 nm (diode) lasers (all 50 mW). CFP (405 laser excitation, 445/40 emission), Venus/Alexa Fluor488 (491 laser excitation, 528/38 emission), AlexaFluor568 (561 laser excitation, 617/73 emission), and AlexaFluor647 (640 laser excitation, 700/75 emission) and differential interference contrast (DIC) images were acquired through a 60x oil-immersion lens. Acquisition times ranged from 100 to 2,000 milliseconds (ms).

RNA extraction for RT-qPCR or RNA-seq

To labeled and unlabeled (controls) HEK-293T cells in 10-cm plates, we added ~1 mL DPBS containing 5 mM Trolox and 10 mM sodium ascorbate, as well as ~4 µL Ribolock RNase inhibitor (Thermo Fischer). The cells were then scrapped off 10-cm plates using cell lifters (Corning), transferred to 2-mL Eppendorf tubes, and spun at ~300G for 4 minutes to pellet cells. The supernatant was removed, and the cells were processed using the RNeasy plus mini kit (Qiagen) following the manufacture protocol, including adding

β -mercaptoethanol to the lysis buffer. The cells were sent through the genomic DNA (gDNA) eliminator column supplied with the kit. One modification to the protocol was replacing the RW1 buffer with RWT buffer (Qiagen) for washing. The extracted RNA was eluted into RNase-free water, and RNA integrity was checked using the Agilent bioanalyzer 2100 using the RNA pico assay. Only RNA with a RIN (RNA integrity number) > 8.5 was used for subsequent experiments. RNAs shorter than 100 nt were not efficiently recovered. RNA concentrations were determined using the Nanodrop (Thermo Fischer).

APEX labeling streptavidin dot blot experiment

RNA from labeled NES-APEX2 (cytosol) HEK-293T cells was treated with Turbo DNase (Thermo Fischer) at 37°C for 30 minute, followed by purification using the RNA clean and concentrator -5 kit (Zymo Research). ~500 ng of purified RNA was blotted on the Amersham Protran 0.45 nitrocellulose (NC) membrane, and the membrane allowed to sit for at least 15 minutes to allow liquid to dry. The RNA was crosslinked to the membrane using 2500 μ J energy (254 nm wavelength, UV Stratalinker 2400)(Spitale et al., 2015). The membrane was then wet with ~5 mL PBST (PBS-TWEEN 20), followed by incubation with 15 mL PBST containing 1 μ L LI-COR Streptavidin IRDye 800CW (green). The membrane was washed thrice with PBS and imaged on the LI-COR Odyssey CLX. For the RNase digestion, we treated the RNA with RNase cocktail enzyme mix (Ambion) for 30 minute at room temperature (RT), followed by purification using the RNA clean and concentrator kit.

All RNA experiments were carried out using standard protocols to minimize and eliminate RNase contamination. These included using a dedicated work area for RNA, using filtered pipette tips, wiping all surfaces with RNase Zap (Invitrogen), using certified RNase-free buffers and reagents, and testing buffers for RNase contamination using RNase Alert (Ambion). When appropriate, ~1-2 μ L of Ribolock RNase inhibitor (Thermo Fischer) was added per 100-200 μ L of buffer/solution.

Horseradish peroxidase (HRP) *in vitro* labeling

For *in vitro* labeling, 100 μ g of yeast tRNA extract (Thermo Fischer) were incubated with 500 μ M BP, 1 mM H₂O₂, and 2.25 μ M HRP (Thermo Fischer) in PBS for 1 minute. The reaction was quenched by adding a PBS solution with final concentration of 10 mM sodium azide, 10 mM sodium ascorbate, and 5 mM Trolox. The reaction was cleaned up by the RNA clean and concentrator -5 kit (Zymo Research). For RNA digestion, 15 μ g of labeled RNA was incubated with 2.5 μ g RNase A (Thermo Fischer) in total volume 25 μ L in water. After 1 hour at room temperature, the reaction was cleaned up by RNA clean and concentrator kit. For proteinase K digestion, 15 μ g of labeled RNA was incubated with 50 μ g of Protease K (Ambion) in a total volume 25 μ L in PBS. After 1 hour at 37°C, the reaction was cleaned up by RNA clean and concentrator kit. 1 μ g of RNA was spotted for each condition and then dot-blotted as described above.

Enrichment of biotinylated RNA

To enrich biotinylated RNAs we used Pierce streptavidin magnetic beads (Thermo Fischer), using 10 μ L beads per 25 μ g of RNA. In general, RNA from 1/2 10-cm plate (~30-50 μ g) was sufficient for generating high-quality polyA+ RNA-seq libraries. The beads were washed 3 times in B&W buffer (5 mM Tris-HCl, pH = 7.5, 0.5 mM EDTA, 1 M NaCl, 0.1% TWEEN 20 (Sigma Aldrich)), followed by 2 times in Solution A (0.1 M NaOH and 0.05 M NaCl), and 1 time in Solution B (0.1 M NaCl). The beads were then suspended in ~100-150 μ L 0.1 M NaCl and incubated with ~100-125 μ L RNA (diluted in water) on a rotator for 2 hours at 4°C. The beads were then placed on a magnet and the supernatant discarded. Beads were washed 3 times in B&W buffer and resuspended in 54 μ L water. A 3X proteinase digestion buffer was made (1.1 mL buffer contained 330 μ L 10X PBS pH = 7.4 (Ambion), 330 μ L 20% N-Lauryl sarcosine sodium solution (Sigma Aldrich), 66 μ L 0.5M EDTA, 16.5 μ L 1M dithiothreitol (DTT, Thermo Fischer) and 357.5 μ L water). 33 μ L of this 3X proteinase buffer was added to the beads along with 10 μ L Proteinase K (20 mg/mL, Ambion) and 3 μ L Ribolock RNase inhibitor. The beads were then incubated at 42°C for 1 hour, followed by 55°C for 1 hour on a shaker. The RNA was then purified using the RNA clean and concentrator -5 kit (Zymo Research). The RNA was typically not bioanalyzed but used as is for downstream applications.

APEX-seq library preparation

RNA-seq libraries were prepared from enriched RNA (corresponding to ~30-50 μ g of pre-enriched RNA) using the Illumina TruSeq stranded mRNA preparation kit, which included polyA+ selection. The prepared libraries were stranded and were quality-controlled by sequencing on the Illumina MiSeq. Good libraries (> 80% unique reads) were sequenced on the Illumina HiSeq 4000 at ~40 million paired (2 x 75) reads per

library. The polyA+ libraries (41 ± 2 million paired reads, mean \pm S.E.M) had high mapping in both targets ($90 \pm 1\%$ uniquely mapped reads) and controls ($86 \pm 1\%$ uniquely mapped reads) (Figure S2B). The correlation between biological replicates was high (between 0.96 and 1). For the MITO APEX-seq we also generated total RNA samples (by omitting the polyA+ selection step in the TruSeq protocol), as well as a polyA+ selected technical replicate for 1 of the labeled samples.

APEX RT-qPCR experiments (MITO)

To test for APEX RNA enrichment, we designed primers against positives (*MTND1* and *MTCO2*) and negatives (*GAPDH*, *SSR2*, *XIST*, *FAU*). The sequences of the primers (purchased from Elim Biopharmaceuticals) were as follows:

MTND1 forward

CACCTCTAGCCTAGCCGTTT

MTND1 reverse

CCGATCAGGGCGTAGTTTGA

MTCO2 forward

AACCAAACCACTTTCACCGC

MTCO2 reverse

CGATGGGCATGAAACTGTGG

GAPDH forward

TTCGACAGTCAGCCGCATCTTCTT

GAPDH reverse

GCCCAATACGACCAAATCCGTTGA

XIST forward

CCCTACTAGCTCCTCGGACA

XIST reverse

ACACATGCAGCGTGGTATCT

SSR2 forward

GTTTGGGATGCCAACGATGAG

SSR2 reverse

CTCCACGGCGTATCTGTTCA

FAU forward

TCCTAAGGTGGCCAAACAGG

FAU reverse

GTGGGCACAACGTTGACAAA

For the RT-qPCR experiments, the enriched MITO-APEX2 RNA was first reverse transcribed following the Superscript III reverse transcriptase (Thermo Fischer) protocol using random hexamers as primers (Kaewsapsak et al., 2017). The resulting cDNA was then tested using qPCR using the primers above in 2X SYBR Green PCR Master Mix (Thermo Fischer), with data generated on Lightcycler 480 (Roche). For each RNA we calculated the ratio of RNA recovered in the labeled target relative to unlabeled controls. We then calculated enrichment as recovery of positives relative to negatives, correcting for primer efficiency ($> 85\%$ for all primers).

APEX RT-qPCR experiments (ERM)

To confirm enrichment of known secretory RNAs by ERM APEX-seq, we designed primers against known secretory (SSR2, TMX1 and SFT2D2) and non-secretory genes (FAU, SUB1, MTCO2). The sequences of the primers (purchased from Sigma Aldrich) were as follows:

SSR2 forward
GTTTGGGATGCCAACGATGAG

SSR2 reverse
CTCCACGGCGTATCTGTTCA

TMX1 forward
ACGGACGAGAACTGGAGAGA

TMX1 reverse
ATTTTGACAAGCAGGGCACC

SFT2D2 forward
CCATCTTCTCATGGGACCAG

SFT2D2 reverse
GCAGAACACAGGGTAAGTGC

FAU forward
TCCTAAGGTGGCCAAACAGG

FAU reverse
GTGGGCACAACGTTGACAAA

SUB1 forward
CGTCACTTCCGGTTCTCTGT

SUB1 reverse
TGATTTAGGCATCGCTTCGC

MTCO2 forward
AACCAAACCACTTTCACCGC

MTCO2 reverse
CGATGGGCATGAAACTGTGG

APEX-RIP RT-qPCR experiments with ERM-APEX2 stable cells were performed as described previously (Kaewsapsak et al., 2017). Briefly HEK 293T cells stably expressing ERM-APEX2 were incubated with BP for 30 minutes prior to 1-minute H₂O₂ labeling. 0.1% (v/v, in PBS) formaldehyde with 10 mM ascorbate and 5 mM Trolox was then added for 10 minutes to quench the reaction and crosslink the RNAs to proteins. The crosslinking reaction was terminated by the addition of 125 mM of glycine for 5 minutes. Following cell lysis in RIPA buffer, streptavidin beads were used to enrich the biotinylated material for 2 hours at 4°C. The crosslinked RNAs and proteins were then reverse crosslinked before protein digestion with proteinase K. The subsequently purified RNA was analyzed by RT-qPCR.

RT-qPCR of *in vitro* 5S RNA to map labeling positions

5S RNA was *in vitro* transcribed as follows. First, a gBlock Gene Fragments (Integrated DNA Technologies) was purchased with the human 5S RNA sequence adjacent to an overhang region (underlined):

ATATGCAAGCAACCCAAGTGGTCTACGGCCATACCACCCTGAACGCGCCCGATCTCGTCTGATCTC
GGAAGCTAAGCAGGGTTCGGGCCTGGTTAGTACTTGGATGGGAGACCGCCTGGGAATACCGGGTGC
TGTAGGCTTT

The DNA was amplified by PCR using Phusion High-fidelity DNA polymerase (NEB) and cleaned up using the QIAquick PCR purification kit (Qiagen). The primers used for amplification were as follows:

Human 5S forward:
ATATGCAAGCAACCCAAGTG

Human 5S reverse:
AAAGCCTACAGCACCCGGTA

The DNA template was used to synthesize RNA using the MEGAscript Transcription T7 Kit (Thermo Fischer), and the transcribed RNA was purified using the MEGAclear kit (Ambion). The integrity and size of the transcribed RNA was checked by running the products on a 6% native agarose gel, stained with SYBR Gold (Thermo Fischer), and imaged using Molecular Imaging System (Biorad).

We prepared labeled reverse primer following the USB Optikinase protocol (Affymetrix) using gamma-³²P ATP (Perkin Elmer). We added ³²P end-labeled reverse primer to ~100 ng RNA (labeled and controls), and the reaction mixture was heated at 95°C for 2 minutes followed by slow cool to 4°C at 10°C/min, to facilitate annealing of primer (1 µL) to the RNA. The primer extension reaction was then initiated with reaction mix, as previously described (Lee et al., 2017). Briefly, the reaction mix (20 µL total; 4 µL 5X first strand buffer (FS), 1 µL 100mM DTT, 1 µL Ribolock inhibitor, 10 µL RNA, and 2 µL water) was added and the mixture preincubated at 52°C for 1 minute before adding Superscript III (1 µL; 200 units, Invitrogen). Separately, similar reactions were carried out spiking in dATP with ddATP (dideoxyadenosine), and dCTP with ddCTP (dideoxycytidine), as described (Lee et al., 2017). Primer extension was carried out at 52°C for 30 minute, after which the reaction was stopped by heating to 95°C for 5 minute, followed by cooling to 4°C. The RNA was then hydrolyzed using NaOH (4M; 1 µL) and heating to 95°C for 5 minutes.

To the cDNA, Gel Loading Buffer II (10 µL; Invitrogen) was added, and the products run on a 35-cm long denaturing 8% polyacrylamide gel with 7M Urea (Thermo Fischer) (Lee et al., 2017). The resulting gel was dried (Labconco Gel Dryer) after placing it on Whatman paper (Sigma Aldrich) and exposed to a storage phosphor screen (Molecular Dynamics) for ~24 hours, and then visualized by phosphorimaging (STORM, Molecular Dynamics). The lanes containing ddATP (corresponding transcribed RNA nucleotide U) and ddCTP (corresponding transcribed RNA nucleotide C) were used to determine the position of all RT stops. All experiments were carried out using biological replicates. RNA secondary structures were predicted using mFOLD software (Zuker, 2003).

Liquid-chromatography (LC)-mass-spectrometry (MS) characterization of *in vitro* reaction products

2 mM dG (Sigma Aldrich) was incubated with 100 µM pentachlorophenol (PCP, Sigma Aldrich) or BP in PBS at 37°C for 1 hour in the presence of 2.25 µM HRP or APEX2 and 100 µM H₂O₂. The reaction was diluted with water to final volume 100 µL and injected into an LC-MS with Zorbax Poroshell 120 SB-C18, 2.1 x 50 mm 2.7 µm column with Poroshell 120 SB-C18 2.1 x 5 mm 2.7 µm guard column. The gradient for LC is shown in the table below. Mass was determined by single quadrupole mass spectrometry with positive and negative atmospheric pressure chemical ionization (APCI) and electrospray ionization (ESI) modes for M/Z = 20-2000.

Time (minute)	Flow (mL/minute)	% solvent A	% solvent B
0	0.3	98	2
2	0.3	98	2
6	0.3	5	95
8	0.3	5	95
8.5	0.3	98	2
9.5	0.3	5	95
10.5	0.3	5	95
11	0.3	98	2
15	0.3	98	2

Solvent A = 0.1% formic acid in water. Solvent B = 0.1% formic acid in acetonitrile.

Mapping and visualizing APEX-seq data

The RNA-seq libraries generated were mapped to the genome rather than to annotated transcriptome, so we could investigate intron retention. The RNA-seq reads were initially subject to barcode removal and primer trimming using a published script(Flynn et al., 2016) based on Trimmomatic(Bolger et al., 2014): (<https://github.com/qczhang/icSHAPE/blob/master/scripts/trimming.pl>):

```
perl trimming.pl -1 $fastq-file1 -2 $fastq-file2 -p $trimmed-file1 -q $trimmed-file2 -l 0 -t 0 -c phred33 -a adapter.fa -m 36
```

The reads were then mapped using STAR(Dobin et al., 2013; Dobin and Gingeras, 2015) to the GRCh39 Ensembl genome, with Homo_sapiens.GRCh38.84.gtf annotations (<http://uswest.ensembl.org/index.html>), using the following command::

```
STAR --genomeDir ~/genome/human/star --runThreadN 8 --readFilesIn $trimmed-file1 $trimmed-file2 --outFileNamePrefix $output-samfile
```

The mapped reads were then counted using HTSEQ(Anders et al., 2014):

```
python -m HTSeq.scripts.count -m intersection-nonempty -s reverse -i gene_id $output-samfile ~/genome/human/Homo_sapiens.GRCh38.84.gtf > $txt
```

The mapped data was visualized using the UCSC browser track(Kent et al., 2002). To generate genome tracks we used samtools(Li et al., 2009) to generate stranded BAM files for each library from the SAM file. The BAM file was then used to generate a bedGraph following the command:

```
genomeCoverageBed -bg -split -ibam $bam -g ~/genome/human/star/chrNameLength.txt > $bed_file
```

BedGraph files from multiple replicates were aggregated using bedtools(Quinlan and Hall, 2010) unionbedg and normalized each track to the same sequencing depth (30 million reads each). The averaged bedGraph files were converted to BigWig files using command bedGraphToBigWig(Kent et al., 2010) for visualization in the UCSC browser(Karolchik et al., 2004). Statistics from the mapped data was aggregated using MultiQC(Ewels et al., 2016). To calculate FPKM (fragments per kilobase per million reads), we obtained transcript lengths from Biomart Ensembl(Durinck et al., 2005) (Ensembl Genes 92, GRCh38), using the longest stable isoform for a gene as its length.

Transcript-level quantification:

Kallisto(Bray et al., 2016) (v 0.43.1) was used to quantify transcript-level abundances of the APEX-seq libraries. A fasta file corresponding to Homo_sapiens.GRCh38.89.gtf and hg38 was downloaded from the Ensembl website and a kallisto index was generated using the kallisto index command with default arguments. To quantify each pair of fastq files, the kallisto quant command with the -b 30 argument was used.

Data analysis using DESeq2

Differentially-expressed genes were determined using DESeq2(Love et al., 2014), using FDR < 0.05 and 18 controls. We tested the effect of imposing other FDR (false discovery rate) cutoffs, and found no appreciable increasing in precision with further decreasing FDR to 0.01. However, increasing FDR beyond 0.05 dramatically decreased precision. All P-values from DESeq2 are FDR adjusted for multiple-hypothesis testing.

For the OMM perturbation experiments, we used the same 18-control strategy, but replacing the unperturbed OMM APEX-seq and cytosol APEX-seq values with the corresponding drug-perturbed values. Using a strategy where we only had 4 controls per perturbation experiments (2 OMM unlabeled and 2 cytosol unlabeled) did not appreciably change the conclusions.

There were two exceptions to the 18-control approach in all our analysis. For the LMA gene in nuclear lamina, and the SENP2 gene in the nuclear pore construct, the unlabeled control had high counts for these genes in the RNA-seq data, as these genes were used to target APEX2 to the corresponding location. Therefore, to compensate, we replaced the LMA 18-control DESeq2 log₂fold-change with the 2-control log₂fold-change (using LMA controls) in the LMA data; we did the same for SENP2 in the nuclear

pore dataset. Otherwise, all other data was as is. As the default DESeq2 approach (Love et al., 2014) replaces outliers when there are > 7 replicates, with our 18-control experiment we didn't need to change the corresponding DESeq2 values of sub-compartments other than nuclear pore and nuclear lamina.

FPKM data sources

To obtain FPKM (reads per kilobase per million reads) of genes, we used two sources. First, we used published polyA+ data (Sultan et al., 2014) by averaging data from two protocols (Qiaquick and Trizol). The corresponding fastq files were downloaded from the European Nucleotide Archive (accession number PRJEB4197), and processed identically to the APEX-seq fastq files. Secondly, we estimated FPKM from the raw counts from our 18-control samples. Briefly we calculated the FPKM for each control sample and averaged the FPKM values from all 18 controls. As the RNA-seq reads from unlabeled samples were obtained after treatment with streptavidin beads, these FPKM values might suffer from sequence- or length-dependent biases. However, genome-wide we found the data to be highly-correlated ($r = 0.95$) with the published data (Sultan et al., 2014). For the correlation analysis, we considered FPKM > 1. For all analyses that used FPKM values, we also only considered genes with FPKM > 1.

Nucleus (NLS) APEX-seq extended analysis

To validate the NLS APEX-seq data we carried out nuclear fractionation of HEK-293T cells by following an established protocol (Gagnon et al., 2014). The protocol uses a detergent nonidet P40 (NP-40) to separate ER contaminants from the nucleus, and we confirmed this strategy was effective by imaging isolated nuclei, staining for ER using ER tracker Red (BODIPY TR Glibenclamide – Thermo Fischer), and visualizing on fluorescence microscope (Zeiss Observer Z1). We modified the protocol so that the extracted RNA was not purified by trizol extraction, but rather by using the RNeasy plus mini kit (Qiagen). RNA-seq was carried out using the same protocol as that used for APEX Seq, thereby generating polyA+ libraries for the nuclear and cytoplasmic fraction using Illumina TruSeq kit. All experiments were carried out in biological replicates.

To estimate precision and accuracy of NLS APEX-seq, we used the fractionation data to generate true-positive and false-positive lists. We did so by first obtaining transcripts differentially expressed in the nuclear relative to cytoplasmic fraction ($p_{\text{FDR-adjusted}} < 0.05$), and did the same for nucleus (NLS) APEX-seq vs cytosol (NES) APEX-seq. We categorized transcripts into the following categories (Figure 2G): (1) log₂fold-change NLS > 0 and log₂fold-change (nuclear/cytosolic fractionation) > 0 (true positive; TP), (2) log₂fold-change NLS > 0 and log₂fold-change (nuclear/cytosolic fractionation) < 0 (false positive; FP), (3) log₂fold-change NLS < 0 and log₂fold-change (nuclear/cytosolic fractionation) > 0 (false negative; FN), (4) log₂fold-change NLS < 0 and log₂fold-change (nuclear/cytosolic fractionation) < 0 (true negative; TN). Precision was defined as TP/(TP+FP), accuracy as (TP+TN)/(TP+TN+FP+FN), sensitivity as TP/(TP+FN), and specificity as TN/(TN+FP). Transcripts shorter than 100 nt were excluded from analysis. We calculated precision and accuracy for all transcripts, as well as for ER-transcripts (those enriched by ERM APEX-seq) and non-ER transcripts. Other analysis approaches (using other log₂foldchange cutoffs, making receiver-operator-curves etc.) did not change the main conclusions.

ERM APEX-seq extended analysis

We extensively tested the ERM APEX-seq dataset against the published ER fractionation (Reid and Nicchitta, 2012) and ER-proximal ribosome profiling (Jan et al., 2014) datasets. Briefly the ERM APEX-seq log₂fold-changes were cytosol normalized by subtracting the corresponding cytosol (NES) log₂fold-change. This ratiometric normalization increased specificity for challenging sub-compartments (Figures S2H), and is routinely used for APEX proteomics analysis (Hung et al., 2014). To test the specificity and sensitivity of ERM APEX-seq, we used receiver-operator-curve (ROC) (Linden, 2006). For the true positive list to test against, we considered the ~640 ER-enriched transcripts by ribosome profiling (log₂fold-change > 0.904, as determined by authors), and the false positives were non-secretory RNAs – these were genes not in Phobius (Kall et al., 2004), SignalP (Petersen et al., 2011) and TMHMM (Krogh et al., 2001). In conjunction with our data, we also examined ER ribosome profiling data using ROCs, including transcripts with total RPKM > 10, as recommended by authors. From the ROC analysis, the cutoff was the value that maximized the difference between the true positive rate (TPR) and false positive rate (FPR). For ERM APEX-seq we explored different analysis approaches, including using 2-controls (ERM APEX-seq controls), 4-controls (ERM and cytosol APEX-seq controls) and 18-control samples. We also explored including or excluding ratiometric normalization (i.e. cytosol normalization to reduce background). We finally settled on the 18-

control condition based on high specificity and reasonably-high coverage (> 1000 enriched genes). For this 18-control experiment, DESeq2 log₂fold-change values from cytosol APEX-seq were subtracted from ERM APEX-seq to obtain ratiometric normalization (ERM/cytosol), and a final cutoff of log₂fold-change (ERM/cytosol) > 0.8725 was obtained from the ROC. Unlike the published studies (Jan et al., 2014; Reid and Nicchitta, 2012), we didn't need to use CHX to stabilize transcripts.

Our analysis suggested improved performance (i.e. specificity) when combining replicates, especially when dealing with challenging/open sub-compartments. As few as 4 controls can improve performance. However, combining controls from multiple constructs/experiments is typically unnecessary when dealing with closed sub-compartments such as the nucleus, cytoplasm or mitochondrial matrix.

To compare the abundance of transcripts recovered by ERM APEX-seq, ER fractionation and ER ribosome profiling, we used a published HEK293 polyA⁺ RNA-seq dataset (Sultan et al., 2014). To estimate the coverage of our ER datasets, we chose a reference gene list comprising of 71 mRNAs that encode ER resident proteins, as previously described (Kaewsapsak et al., 2017). For the comparison of coverage with ERM proteomics, we used the published dataset containing gene names (Hung et al., 2017).

OMM APEX-seq extended analysis

For OMM APEX-seq data in [Figure 6](#) and [7](#), a similar ratiometric normalization to ERM with cytosol was carried out. Based on extensive testing with ERM APEX-seq and nucleus APEX-seq using the corresponding known gene-lists, we found that in general a default log₂fold-change cutoff of 0.75 was suitable for dealing with APEX-seq when prior knowledge was unavailable, or un-assumed. We therefore used log₂fold-change (OMM/cytosol) = 0.75 as a cutoff to identify OMM-enriched transcripts.

For labeling secretory RNAs ([Figure S4](#)), we identify and display secretory RNAs in this order (1) first all Phobius, (2) TMHMM but not Phobius, (3) SignalP but not Phobius or TMHMM (4) Gene ontology cellular component (GOCC) but not Phobius, TMHMM or SignalP. Mitocarta2.0 were excluded from secretory RNA (own category).

Mitochondria drug perturbation

For cycloheximide treatment, APEX labeling in OMM-APEX2 stable cells was initiated by changing the medium to fresh medium containing 500 mM biotin-phenol. This was incubated at 37°C under 5% CO₂ for 15 minutes. Then cycloheximide was added to the medium to a final concentration of 0.1 mg/mL and the cells were further incubated at 37°C under 5% CO₂ for another 15 minutes. H₂O₂ was then added to each sample to a final concentration of 1 mM, and the plate gently agitated for 1 minute. Then the samples were quenched and processed the same way as other APEX-seq samples. For puromycin or CCCP experiment, APEX labeling in OMM-APEX2 stable cells was initiated by changing the medium to fresh medium containing 200 μM puromycin (or 40 μM CCCP) and 500 mM biotin-phenol. This was incubated at 37°C under 5% CO₂ for 30 minutes. H₂O₂ was then added to each sample to a final concentration of 1 mM, and the plate gently agitated for 1 minute. Then the samples were quenched and processed the same way as other APEX-seq samples.

OMM perturbation data analysis

For gene classification, mitochondrial genes were annotated according to MitoCarta 2.0 (Calvo et al., 2015); secretory genes were annotated as in Kaewsapsak *et al.* (Kaewsapsak et al., 2017) according to Phobius, SignalP, TMHMM, GOCC and ER proximity ribosome profiling. For mitochondrial genes, we adapted the original TargetP algorithm prediction score (0 = no targeting sequence, 1-5 from strongest to weakest) to a different scale (0 = no targeting sequence, 1-5 from weakest to strongest) as a metric of the strength of N-terminus mitochondrial targeting sequence. The mitochondrial gene functional classes were annotated according to Gene Ontology and are listed in [Table 2](#). For the gene density plots in [Figure 6A-B](#), all detected transcripts in each condition were plotted by their log₂fold-change normalized against their respective cytosol control using a bin size of 0.2. For the TargetP and functional class analysis in [Figure 7C and 7F](#), the top 100 enriched mitochondrial genes in each experiment were selected based on MitoCarta 2.0 for mitochondrial annotation and log₂fold-change of OMM values normalized against their respective cytosol control.

Correlation and T-distributed stochastic neighbor embedding (t-SNE) analysis

For the 9-location correlation and t-SNE(van der Maaten and Hinton, 2008) analysis we only included genes with total counts greater than 3900 across all 36 samples (2 labeled targets and 2 unlabeled controls per location, average $\sim 10^2$ counts/sample). We excluded genes with transcript length less than 100 nt. The raw counts from HTSEQ were rlog-transformed using DESeq2-normalized counts. Pearson correlation on the transformed counts was carried out using the package corrplot in R, using clustering method "centroid" and order "hclust". t-SNE analysis was also performed in R.

For the OMM drug-treatment experiments, we only included genes with total counts greater than 3500 across 32 samples (2 labeled targets and 2 labeled controls for the following: OMM_basal, OMM_cycloheximide, OMM_puromycin, OMM_cccp, NES_basal, NES_cycloheximide, NES_puromycin, NES_cccp). Therefore, average $\sim 10^2$ counts/sample). All other analyses were carried out identically to the 9-location analysis.

Data analysis using DESeq2

Differentially-expressed genes were determined using DESeq2(Love et al., 2014), using FDR < 0.05 and 18 controls. We tested the effect of imposing other FDR (false discovery rate) cutoffs, and found no appreciable increasing in precision with further decreasing FDR to 0.01. However, increasing FDR beyond 0.05 dramatically decreased precision. All P-values from DESeq2 are FDR adjusted for multiple-hypothesis testing.

For the OMM perturbation experiments, we used the same 18-control strategy, but replacing the unperturbed OMM APEX-seq and cytosol APEX-seq values with the corresponding drug-perturbed values. Using a strategy where we only had 4 controls per perturbation experiments (2 OMM unlabeled and 2 cytosol unlabeled) did not appreciably change the conclusions.

There were two exceptions to the 18-control approach in all our analysis. For the LMA gene in nuclear lamina, and the SENP2 gene in the nuclear pore construct, the unlabeled control had high counts for these genes in the RNA-seq data, as these genes were used to target APEX2 to the corresponding location. Therefore, to compensate, we replaced the LMA 18-control DESeq2 \log_2 fold-change with the 2-control \log_2 fold-change (using LMA controls) in the LMA data; we did the same for SENP2 in the nuclear pore dataset. Otherwise, all other data was as is. As the default DESeq2 approach(Love et al., 2014) replaces outliers when there are > 7 replicates, with our 18-control experiment we didn't need to change the corresponding DESeq2 values of sub-compartments other than nuclear pore and nuclear lamina.

FPKM data sources

To obtain FPKM (reads per kilobase per million reads) of genes, we used two sources. First, we used published polyA+ data(Sultan et al., 2014) by averaging data from two protocols (Qiaquick and Trizol). The corresponding fastq files were downloaded from the European Nucleotide Archive (accession number PRJEB4197), and processed identically to the APEX-seq fastq files. Secondly, we estimated FPKM from the raw counts from our 18-control samples. Briefly we calculated the FPKM for each control sample and averaged the FPKM values from all 18 controls. As the RNA-seq reads from unlabeled samples were obtained after treatment with streptavidin beads, these FPKM values might suffer from sequence- or length-dependent biases. However, genome-wide we found the data to be highly-correlated ($r = 0.95$) with the published data(Sultan et al., 2014). For the correlation analysis, we considered FPKM > 1. For all analyses that used FPKM values, we also only considered genes with FPKM >1.

Heatmap and gene-ontology (GO)-term analysis

For the integrated analysis of locations, we excluded MITO APEX-seq as the labeled targets perturbed the entire analysis; we believe this is due to the large enrichment of ~ 13 mitochondrial mRNAs and 2 rRNAs, which constitute over 50% reads in targets; combined with the relative spatial isolation of the matrix.

To generate reliable gene data for the integrated analysis, we took our APEX-seq enrichment data and imposed the following filtering criteria: (1) excluding transcripts shorter than 100 nt that were not recovered efficiently. (2) Considering transcripts with common gene names (typically HUGO gene nomenclature committee (HGNC) approved names(Maglott et al., 2011; Pruitt et al., 2007)) (3) only including genes that were enriched in at least 1 location ($p_{\text{FDR-adjusted}} < 0.05$ and \log_2 fold-change > 0.75 for at least 1 location); and (4) Genes had \log_2 fold-change data estimated from DESeq2 for all locations (i.e. excluding any genes with NA values in any location). This last filtering step typically excluded low-abundant transcripts that might occasionally show up as enriched in a location but didn't have sufficient counts in

other locations. Such low-abundant transcript-data was typically less reliable. We did not impose any FPKM or counts cutoff in our analysis. Our analysis yielded 3262 genes. As the fold changes in this study were calculated relative to unlabeled controls, enrichment by APEX-seq is a proxy for transcript concentration. Thus, the cytosol, which constitutes most of cell, recovers fewer transcripts than expected, as it is not possible to highly concentrate transcripts there.

Heatmaps of this data were generated using pheatmap2 in R. From the heatmap, clusters were estimated using hierarchical clustering. The cluster number was checked using a number of approaches including gap statistics (Tibshirani et al., 2001), and the combined cluster range was then explored. The genes belonging to each cluster was then compared to all enriched genes to estimate cellular GO-terms. GO-term analysis was carried out using PANTHER (Mi et al., 2013; Thomas et al., 2003) (<http://pantherdb.org/about.jsp>), using Fischer's exact test with FDR multiple test correction. We only consider GO-terms with FDR < 0.005. To construct the heatmaps of most variable mRNAs (Figure 3C), we considered genes with average counts > 1000. For the lncRNA heatmap (Figure 3D), we excluded all genes that were not lncRNAs, processed transcripts and pseudogenes.

We took a similar approach for making heatmaps was used for the OMM analysis with the drug perturbations (puromycin, CCCP, cycloheximide). However, for clustering we did not use CCCP data. For the subsequent GO-term analysis we used the Reactome pathway from PANTHER, with the control set comprising all genes enriched in at least one of the 4 OMM APEX-seq conditions (~1900 genes).

Nuclear-locations m⁶A modification and length analysis

To examine transcript-length differences across the nuclear locations (nucleus, nucleolus, nuclear pore, nuclear lamina), we filtered genes as described above for the heatmaps, but we excluded the filtering step by common gene names. Our analysis yielded 3290 genes. We obtained transcript lengths from two sources: (1) The longest stable isoform, as obtained from Biomart, Ensembl (Durinck et al., 2005), and (2) the most-abundant isoform across all compartments in our APEX-seq data, as determined by rMATs (Shen et al., 2014). Using the transcript length from either of these databases yielded similar trends and conclusions.

To determine the contribution of 5' UTR, CDS (coding sequence) and 3' UTR to the overall transcript-length difference between nucleus- and nuclear-pore enriched transcripts, we considered the most-abundant isoform, but excluded non-coding transcripts (i.e. transcripts with a CDS length = 0).

To calculate the number of m⁶A sites per transcript, we used a published dataset to obtain high-confidence m⁶A sites in HEK293 (Meyer et al., 2012).

Network analysis

RNA interactions for the 3262 genes were investigated in multiple ways. These include (1) tabulating the overlapping genes using UpSet (Lex et al., 2014) in R; and (2) using circlize (Gu et al., 2014) to investigate the data by comparing which transcripts in each cellular sub-compartment most often had residency in other locations. In all instances the fold change for all genes for all transcripts was binarized to either 1 (log₂fold-change > 0.75), else 0.

Lamin-associated domains (LADs) and nucleolus-associated domains analysis

For the LADs and NADs analysis, we aggregated data from the following sources to obtain the relevant associated genomic regions: Guelen *et al.* (Guelen et al., 2008), Dillinger *et al.* (Dillinger et al., 2017), Nemeth *et al.* (Nemeth et al., 2010). We used library ((TxDb.Hsapiens.UCSC.hg18.knownGene) in R, obtained from Bioconductor (Durinck et al., 2005; Gentleman et al., 2004) to get the genes contained within these regions.

Quantification of intron retention and intron switching

rMATs (Shen et al., 2014) was used to quantify intron retention in each location. rMATs was run by comparing the APEX-seq BAM files against all controls using the arguments -t paired, and a GTF file downloaded from Ensembl (Genecode v26, Ensembl 88). Only retained intron events with FDR ≤ 0.05 were considered using the "RI.MATS.JCEC.txt" output file (using both junction and exon counts).

rMATs (v 4.0.1) was used to quantify the number of isoform-switching genes (Figures S5F-G) in the APEX-seq labeled samples relative to all unlabeled controls. The number of significant differential-splicing events (FDR < 0.05) for each compartment was read from the rMATs JCEC files. To remove noise from low-abundance transcripts, the KDEL labeled sample was used as a filter. Any differential splicing

events identified in the KDEL labeled samples were ignored, as previous analysis using DESEQ2 found no significantly-enriched genes at that location. The number of genes containing at least one differential splicing event between a labeled compartment and unlabeled controls are reported for the respective alternative splicing event.

Isoform analysis, including isoform switching

Sleuth(Pimentel et al., 2017) was used to perform differential transcript expression analysis between locations, which were compared against all control samples. For the analyses in identifying isoform switching, to generate gene-level abundances the Bioconductor(Gentleman et al., 2004) tximport package was used to import kallisto(Bray et al., 2016) abundances and aggregate to the gene-level. DESeq2(Love et al., 2014) was subsequently used to perform differential gene expression analysis. Genes that displayed isoform switching were identified as follows: First, using the differential gene expression output from DESeq2, genes displaying no significant differential expression between the nucleus (NLS) and cytosol (NES) samples were identified. For each of these genes, we then determined if there were any transcripts that were significantly enriched in either the nucleus or cytosol samples (as determined by Sleuth(Pimentel et al., 2017)). Genes displaying no differential expression between the nucleus and cytosol, but with at least one transcript enriched in the nucleus and a different transcript enriched in the cytosol, were called as displaying isoform switching. To select genes to display in [Figure 4E](#), an expression cutoff (at the gene level) of $\log_2\text{counts} \geq 12$ and an isoform difference metric ≥ 10 was set. The isoform difference metric was computed by taking the sum of the absolute values of the $\log_2\text{fold-change}$ enrichments for the most cytosol-biased and most nuclear-biased transcript.

Repeat analysis:

A list of all annotated repeat elements was downloaded from UCSC Table Browser(Karolchik et al., 2004). To determine the relative enrichment of repeat elements in the genes enriched in each location, a set of enriched genes in each location was determined as described previously ($\log_2\text{fold-change} \geq 0.75$, $p_{\text{FDR-adjusted}} < 0.05$). A unique set of genes for each location was determined by removing genes that were enriched in more than one APEX-seq location.

For each set of genes, a corresponding list of genomic coordinates comprising only exonic sequences was generated. The most abundant isoform of each gene was used for determining the coordinates of the exons. This list was then intersected with the repeat annotation using bedtools(Quinlan and Hall, 2010) intersect with the -F .51 command to require that at least half of the repeat annotation was present in an exonic sequence. This was then aggregated by gene to generate a "repeat count" by gene table (with all repeat families as rows, and all genes uniquely enriched in a given location as columns). This table was then binarized to result in a table reporting the presence or absence of a repeat element in each gene. The proportion of genes in each location that contained a given repeat family was then determined. To perform FDR calculations, the gene-location pairings were randomly permuted 1000 times, and the number of permutations in which the resulting enrichment value was at least as great as the observed enrichment value was divided by the total number of permutations.

To quantify the abundance of rRNA repeat elements, the same list of all annotated repeat elements from UCSC(Karolchik et al., 2004) was used. The number of reads mapping to any rRNA element for each location was determined using bedtools intersect, essentially using the rRNA repeat annotations as a "gene" or "feature" to quantify overall abundance. These values were depth normalized to the total number of aligned reads in each BAM file, and averaged across replicates.

Sequential error-robust fluorescence *in situ* hybridization (FISH) design and analysis

Sequential oligo library for 56 selected genes was designed according to Moffitt *et al.*(Moffitt and Zhuang, 2016) and synthesized by CustomArray. The library was then PCR amplified using Phusion Hot-start Master Mix (NEB, M05365) and then cleaned up using DNA clean and concentrator -5 columns (DCC-5, Zymo Research, D4013). The library was then in-vitro transcribed with T7 polymerase using the HiScribe kit (NEB, E20505) at 37°C overnight. The resulting RNA was then reverse transcribed with Maxima H Minus RT enzyme (ThermoFisher, EP0751) at 50°C for 1 hour and the remaining RNA digested with 0.25 M EDTA and 0.5 M NaOH at 95°C for 10 minutes. The correct size of the reverse transcribed RNA was confirmed by running the product on a 15% urea TBE gel. The probes were further cleaned up using DCC-25 columns (Zymo Research, D4005).

To hybridize FISH probes on to cells, HEK 293T cells were fixed with 4% paraformaldehyde (PFA) in PBS for 10 minutes before permeabilized with 0.5% Triton-X in PBS for 10 minutes. The sample was then incubated in 50% formamide and 0.1% TWEEN 20 in 2X SSC solution for 35 minutes. 500-800 ng/ μ L of the synthesized probe were added onto the cells using a coverslip and the slides were denatured at 90°C for 10 minutes. Probes were hybridized overnight at 42°C in a humidified chamber. The cells were then washed twice with prewarmed 42°C 2XSSC solution for 10 minutes the next day before imaging on a confocal microscope. A total of 20 fields of view, each with > 20 cells, were imaged for all 56 genes and then the data were processed using MATLAB. For imaging quantitation and analysis of each field of view, we generated a mask for each individual gene of interests using a uniform threshold cutoff of 0.5 – 0.998 after removing all the non-cell pixels. The colocalization with *MTND3* was calculated by intersecting the mask of a particular gene of interest (for example, “*XIST*” mask) with *MTND3* mask and then divided sum intensity of the intersected mask by the sum intensity of the gene mask of interests. The colocalization results for all 20 fields of view were then calculated to obtain the average and standard deviation.

Sequential FISH quality control

For the sequential FISH experiment we designed probes against 56 transcripts, which include a combination of known and unknown-location genes. Using the FISH images generated for each demultiplexed transcript, we subsequently excluded transcripts that (1) could not be decoded based on the barcode, or (2) that didn't show any localization (typically low-abundant ones) based on the images being hazy and lacking punctate spots. To carrying out this exclusion in a relatively unbiased manner, we had 3 people (F.M.F., S.H., K.R.P.) independently examine the images for all genes, and rate transcript-localization information as (1) high confidence (2) medium confidence and (3) low/no confidence. We then tallied all these ratings and subsequently excluded transcripts that were assigned a “no confidence” value by any of the 3 people. This filtering exercise resulted in 29 transcripts being retained. In general, the 3 people strongly agreed on the confidence ratings. We separately also confirmed that for transcripts with known localization, the discarded genes did not correlate well with known localizations. Finally, the included genes were plotted as $\log_2(\text{colocalization-score})$ versus APEX-seq $\log_2\text{fold-change}$ values (Figure 2H, S3B).

MITO APEX-seq extended analysis

For the MITO APEX-seq, we obtained strong enrichment ($\log_2\text{fold-change} > 2.9$) of the 13 MT mRNAs and 2 MT rRNAs in the targets (i.e. labeled libraries) relative to unlabeled controls. These 15 genes made up > 50% of reads in the MITO APEX-seq labeled samples.

NES APEX-seq extended analysis

We did not find many transcripts highly enriched ($\log_2\text{fold-change} > 0.75$) by cytosol (NES) APEX-seq. We believe fewer transcripts are recovered because APEX-seq enrichment is a proxy for RNA concentration rather than RNA amount. Thus, the cytosol, which contains a majority of transcripts, recovers fewer highly-enriched transcripts relative to the unlabeled (i.e. whole-cell) controls. Nonetheless we compared the transcripts enriched by cytosol APEX-seq ($\log_2\text{foldchange} > 0$, $p_{\text{FDR-adjusted}} < 0.05$) and found them to have higher enrichment in the cytosol fractionation fraction relative to transcripts with cytosol APEX-seq $\log_2\text{foldchange} < 0$.

KDEL APEX-seq extended analysis

For the KDEL construct, DESeq2 did not show any significantly enriched transcripts relative to unlabeled controls, when FDR adjusted. If there are any significantly-enriched transcripts larger than 100 nt, they are few in number. We therefore used the KDEL APEX-seq data as a “negative control” our 8-location integrated analysis by rejecting analysis strategies that yield a large number of enriched KDEL transcripts, as such approaches likely have large false positives. We found that using a $\log_2\text{fold-change}$ between 0.6 and 0.9 was a sufficient comprise to obtain highly-specific gene lists without further comprising coverage (Figure S3C).

Proteomic analysis

For the proteomic analysis, we used subcellular protein localization data from Thul *et al.* (Thul *et al.*, 2017), using the main location for genes that were imaged using validated and supported antibodies. We filtered the data to exclude duplicate protein localization entries from multiple cell-lines. The majority of the proteins

in that data set are nucleoplasmic or cytosolic. We calculated an “enrichment” score for each protein type in each location by carrying out a two-step normalization: (1) obtaining the enrichment of that protein type relative to all proteins, and (2) enrichment of that protein in each location relative to all locations.

Other analysis and data availability

Where appropriate, the following tests were employed (1) Student’s t-test, (2) Mann-Whitney U test (Wilcoxon rank-sum test) (3) Kolmogorov–Smirnov (KS) test, (4) Fischer’s exact test, (5) hypergeometric distribution test. All tests were carried out in R. All analysis was carried out using R (most plots using ggplot2(Wickham, 2009)), python and Microsoft Excel. All sequencing data available through GEO.

REFERENCES

- Anders, S., Pyl, P.T., and Huber, W. (2014). HTSeq - a Python framework to work with high-throughput sequencing data (Cold Spring Harbor Laboratory).
- Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114-2120.
- Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat Biotech* 34, 525-527.
- Calvo, S.E., Clauser, K.R., and Mootha, V.K. (2015). MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. *Nuc Acids Res* 44, D1251-D1257.
- Dillinger, S., Straub, T., and Nemeth, A. (2017). Nucleolus association of chromosomal domains is largely maintained in cellular senescence despite massive nuclear reorganisation. *PLoS One* 12, e0178821.
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15-21.
- Dobin, A., and Gingeras, T.R. (2015). Mapping RNA-seq reads with STAR. In *Curr Prot in Bioinform* (John Wiley & Sons, Inc.), pp. 11.14.11-11.14.19.
- Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 21, 3439-3440.
- Ewels, P., Magnusson, M., Lundin, S., and Kaller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32, 3047-3048.
- Flynn, R.A., Zhang, Q.C., Spitale, R.C., Lee, B., Mumbach, M.R., and Chang, H.Y. (2016). Transcriptome-wide interrogation of RNA secondary structure in living cells with icSHAPE. *Nat Protoc* 11, 273-290.
- Gagnon, K.T., Li, L., Janowski, B.A., and Corey, D.R. (2014). Analysis of nuclear RNA interference in human cells by subcellular fractionation and Argonaute loading. *Nat Protoc* 9, 2045-2060.
- Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., *et al.* (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5, R80.
- Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). circlize Implements and enhances circular visualization in R. *Bioinformatics* 30, 2811-2812.
- Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., de Laat, W., *et al.* (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* 453, 948-951.
- Hung, V., Lam, S.S., Udeshi, N.D., Svinkina, T., Guzman, G., Mootha, V.K., Carr, S.A., and Ting, A.Y. (2017). Proteomic mapping of cytosol-facing outer mitochondrial and ER membranes in living human cells by proximity biotinylation. *ELife* 6.
- Hung, V., Udeshi, N.D., Lam, S.S., Loh, K.H., Cox, K.J., Pedram, K., Carr, S.A., and Ting, A.Y. (2016). Spatially resolved proteomic mapping in living cells with the engineered peroxidase APEX2. *Nat Protoc* 11, 456-475.
- Hung, V., Zou, P., Rhee, H.-W., Udeshi, Namrata D., Cracan, V., Svinkina, T., Carr, Steven A., Mootha, Vamsi K., and Ting, Alice Y. (2014). Proteomic mapping of the human mitochondrial intermembrane space in live cells via ratiometric APEX tagging. *Mol Cell* 55, 332-341.
- Jan, C.H., Williams, C.C., and Weissman, J.S. (2014). Principles of ER cotranslational translocation revealed by proximity-specific ribosome profiling. *Science* 346, 1257521-1257521.

Kaewsapsak, P., Shechner, D.M., Mallard, W., Rinn, J.L., and Ting, A.Y. (2017). Live-cell mapping of organelle-associated RNAs via proximity biotinylation combined with protein-RNA crosslinking. *Elife* 6.

Kall, L., Krogh, A., and Sonnhammer, E.L. (2004). A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* 338, 1027-1036.

Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. (2004). The UCSC Table Browser data retrieval tool. *Nuc Acids Res* 32, D493-496.

Kent, W.J., S., Z.A., G., B., S., H.A., and D., K. (2010). BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* 26, 2204-2207.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome Res* 12, 996-1006.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 305, 567-580.

Lee, B., Flynn, R.A., Kadina, A., Guo, J.K., Kool, E.T., and Chang, H.Y. (2017). Comparison of SHAPE reagents for mapping RNA structures inside living cells. *RNA* 23, 169-174.

Lex, A., Gehlenborg, N., Strobel, H., Vuilleumot, R., and Pfister, H. (2014). UpSet: visualization of intersecting sets. *IEEE Trans Vis Comp Graph* 20, 1983-1992.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078-2079.

Linden, A. (2006). Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. *J Eval Clin Pract* 12, 132-139.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15.

Maglott, D., Ostell, J., Pruitt, K.D., and Tatusova, T. (2011). Entrez gene: gene-centered information at NCBI. *Nuc Acids Res* 39, D52-57.

Mercer, Tim R., Neph, S., Dinger, Marcel E., Crawford, J., Smith, Martin A., Shearwood, A.-Marie J., Haugen, E., Bracken, Cameron P., Rackham, O., Stamatoyannopoulos, John A., *et al.* (2011). The human mitochondrial transcriptome. *Cell* 146, 645-658.

Meyer, K.D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C.E., and Jaffrey, S.R. (2012). Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell* 149, 1635-1646.

Mi, H., Muruganujan, A., Casagrande, J.T., and Thomas, P.D. (2013). Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc* 8, 1551-1566.

Moffitt, J.R., and Zhuang, X. (2016). RNA Imaging with Multiplexed Error-Robust Fluorescence In Situ Hybridization (MERFISH). *Meth Enzy* 572, 1-49.

Nemeth, A., Conesa, A., Santoyo-Lopez, J., Medina, I., Montaner, D., Peterfia, B., Solovei, I., Cremer, T., Dopazo, J., and Langst, G. (2010). Initial genomics of the human nucleolus. *PLoS Genet* 6, e1000889.

Petersen, T.N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* 8, 785-786.

Pimentel, H., Bray, N.L., Puente, S., Melsted, P., and Pachter, L. (2017). Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat Methods* 14, 687-690.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nuc Acids Res* 35, D61-65.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841-842.

Reid, D.W., and Nicchitta, C.V. (2012). Primary role for endoplasmic reticulum-bound ribosomes in cellular translation identified by ribosome profiling. *J Biol Chem* 287, 5518-5527.

Shen, S., Park, J.W., Lu, Z.X., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q., and Xing, Y. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci USA* 111, E5593-5601.

Spitale, R.C., Flynn, R.A., Zhang, Q.C., Crisalli, P., Lee, B., Jung, J.-W., Kuchelmeister, H.Y., Batista, P.J., Torre, E.A., Kool, E.T., *et al.* (2015). Structural imprints in vivo decode RNA regulatory mechanisms. *Nature* 519, 486-490.

Sultan, M., Amstislavskiy, V., Risch, T., Schuette, M., Dokel, S., Ralser, M., Balzereit, D., Lehrach, H., and Yaspo, M.L. (2014). Influence of RNA extraction methods and library selection schemes on RNA-seq data. *BMC Genomics* 15, 675.

Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* 13, 2129-2141.

Thul, P.J., Akesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., Alm, T., Asplund, A., Bjork, L., Breckels, L.M., *et al.* (2017). A subcellular map of the human proteome. *Science* 356.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J Royal Stat Soc: Series B (Stat Meth)* 63, 411-423.

van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J Machine Learn Res* 9, 2579-2605.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer, 1-+.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nuc Acids Res* 31, 3406-3415.