Supporting Information

Lee et al. 10.1073/pnas.1207846109

SI Materials and Methods

Cell Culture and Drug Treatment. Human HEK293 cells and mouse embryonic fibroblasts (MEF) were maintained in DMEM with 10% (vol/vol) FBS. Cycloheximide (CHX) was purchased from Sigma and harringtonine from LKT Laboratories. Lactimidomycin (LTM) was described previously (1). All drugs were dissolved in DMSO. Cells were treated with 100 μM CHX, 50 μM LTM, 2 μg/mL (3.8 μM) harringtonine, or an equal volume of DMSO at 37 °C for 30 min.

Polysome Profiling. Sucrose solution was prepared in polysome buffer (pH 7.4, 10 mM Hepes, 100 mM KCl, 5 mM $MgCl_2$). Sucrose density gradients (15–45%, wt/vol) were freshly made in SW41 ultracentrifuge tubes (Backman) using a Gradient Master (BioComp Instruments) according to manufacturer's instructions. Cells were washed using ice-cold PBS containing 100 μg/mL CHX and then were lysed by extensive scraping in polysome lysis buffer (pH 7.4, 10 mM Hepes, 100 mM KCl, 5 mM $MgCl_2$, 100 μg/mL CHX, and 2% Triton X-100). For DMSO control, the CHX was omitted in both PBS and polysome lysis buffer. Cell debris was removed by centrifugation 20,800 × *g* for 10 min at 4 °C. Six hundred microliters of supernatant were loaded onto sucrose gradients followed by centrifugation for 100 min at 178,000 × *g* at 4 °C in a SW41 rotor. Separated samples were fractionated at 0.750 mL/min through a fractionation system (Isco) that continually monitored OD_{254} values. Fractions were collected at 0.5-min intervals.

Purification of Ribosome-Protected mRNA Fragments. The general procedure of ribosome-protected mRNA fragment (RPF) purification was based on the previously reported protocol (2) with some modifications. In brief, polysome-profiling fractions were mixed, and a 140-μL aliquot was digested with 200 U *Escherichia coli* RNase I (Ambion) at 4 °C for 1 h. Then total RNA was extracted by TRIzol reagent (Invitrogen) followed by dephosphorylation with 20 U T4 polynucleotide kinase (New England Biolabs) in the presence of 10 U SUPERase_In (Ambion) at 37 °C for 1 h. The enzyme was heat-inactivated for 20 min at 65 °C. The digested RNA products were separated on a Novex denaturing 15% polyacrylamide Tris-borate-EDTA (TBE)-urea gel (Invitrogen). The gel was stained with SYBR Gold (Invitrogen) to visualize the digested RNA fragments. Gel bands of ~ 28-nt RNA molecules were excised and disrupted physically by centrifugation through the holes of the tube. The gel debris was soaked overnight in the RNA gel elution buffer [300 mM NaOAc (pH 5.5), 1 mM EDTA, 0.1 U/mL SUPERase_In] to recover the RNA fragments. The gel debris was filtered out with a Spin-X column (Corning), and RNA was purified using ethanol precipitation.

cDNA Library Construction and Deep Sequencing. Poly-A tails were added to the purified RNA fragments by *E. coli* poly-(A) polymerase (New England Biolabs) with 1 mM ATP in the presence of 0.75 U/μL SUPERase_In at 37 °C for 45 min. The tailed RNA molecules were reverse transcribed to generate the first-strand cDNA using SuperScript III (Invitrogen) and the following oligos containing barcodes:

SCT01:5′-pCTGATCGTCGGACTGTAGAACTCTCAAGCAGAAGACGGCATACGATTTTTTTTTTTTTTTTTTTVN-3′
MCA02: 5′-pCAGATCGTCGGACTGTAGAACTCTCAAGCAGAAGACGGCATACGATTTTTTTTTTTTTTTTTTTVN-3′
LGT03:5′-pGTGATCGTCGGACTGTAGAACTCTCAAGCAGAAGACGGCATACGATTTTTTTTTTTTTTTTTTTVN-3′
HTC04: 5′-pTCGATCGTCGGACTGTAGAACTCTCAAGCAGAAGACGGCATACGATTTTTTTTTTTTTTTTTTTVN-3′
YAG05:5′-pAGGATCGTCGGACTGTAGAACTCTCAAGCAGAAGACGGCATACGATTTTTTTTTTTTTTTTTTTVN-3′

Reverse-transcription products were resolved on a 10% polyacrylamide TBE-urea gel as described above. The expected 92-nt band of the first-strand cDNA was excised and recovered using DNA gel elution buffer (300 mM NaCl, 1 mM EDTA). The purified first-strand cDNA then was circularized by 100 U CircLigase II (Epicentre) following the manufacturer's instructions. The circular single-strand DNA was purified using ethanol precipitation and was relinearized by 7.5 U apurinic/apyrimidinic endonuclease (APE1) in 1× buffer 4 (New England Biolabs) at 37 °C for 1 h. The linearized products were resolved on a Novex 10% polyacrylamide TBE-urea gel (Invitrogen). The expected 92-nt band then was excised and recovered.

The single-stranded template then was amplified by PCR using the Phusion High-Fidelity enzyme (New England Biolabs) according to the manufacturer's instructions. The primers qNTI200 (5′-CAAGCAGAAGACGGCATA-3′) and qNTI201 (5′-AATGATACGGCGACCACCG ACAGGTTCAGAGTTCTACAGTCCGACG-3′) were used to create a DNA library suitable for sequencing. The PCR contains 1× HF buffer, 0.2 mM dNTP, 0.5 μM primers, and 0.5 U Phusion polymerase. PCR was carried out with an initial 30-s denaturation at 98 °C, followed by 12 cycles of denaturation for 10 s at 98 °C, annealing for 20 s at 60 °C, and extension for 10 s at 72 °C. PCR products were separated on a nondenaturing 8% polyacrylamide TBE gel as described above. The expected 120-bp band was excised and recovered as described above. After quantification by Agilent BioAnalyzer DNA 1000 assay, equal amounts of barcoded samples were pooled into one sample. Mixed DNA samples (~3–5 pmol) typically were used for cluster generation followed by sequencing using the sequencing primer 5′-CGACAGGTTCAGAGTTC TACAGTCCGACGATC-3′ (Illumina HiSEQ system, Cornell University Life Sciences Core Laboratories Center, Ithaca, NY).

Mapping RPF to RefSeq Transcripts. To remove adaptor sequences, seven nucleotides were cut from the 3′ end of each 50-nt-long Illumina sequence read, and a stretch of A's were removed from the 3′ end, allowing one mismatch. The remaining insert sequence was separated according to the 2-nt barcode at the 5′ end after the barcode was removed. Reads between 26 and 29 nt in length were mapped to the sense strand of the entire human or mouse RefSeq transcript sequence library (release 49), using Bowtie-0.12.7 (3). One mismatch was allowed in all mappings; in cases of multiple mapping, mismatched positions were not used if a perfect match existed. Reads mapped more than 100 times were discarded to remove poly-A–derived reads. Finally, reads were counted at every position of individual transcripts by using the 13th nucleotide of the read for the P-site position. Two HEK293 technical replicates were pooled for most analyses.

Coding Sequence Annotation. The most recent freezes of data from the Consensus Coding DNA Sequence (CCDS) database (4) were downloaded from the National Center for Biotechnology Information FTP site (January 24, 2011 for mouse, September 7,

2011 for human) to find annotated translational start and end positions on each mRNA. Each of the CCDS nucleotide sequences was mapped to the associated RefSeq mRNA sequences based on following conditions: (*i*) the first three nucleotides must match perfectly; (*ii*) up to two mismatches are allowed in the first 10 nucleotides; (*iii*) up to 20 mismatches are allowed in the full length, with no gaps allowed. The maximum number of mismatches in an accepted alignment was 10.

Read Aggregation Plots. The number of RPF reads aligned to each position of individual transcripts was first normalized by the total number of reads recovered on the same mRNA. The reads counts then were averaged across all mRNAs for each position relative to the annotated start codon. To avoid multiple counting of the same reads mapped to multiple isoforms of the same gene, redundant mRNAs were removed based on the sequence context of −100 nt to +100 nt relative to the annotated translation initiation site (aTIS). The same approach was used to obtain average read aggregation relative to downstream TIS (dTIS) or upstream TIS (uTIS) positions.

Identification of TIS Positions. A peak is defined at the nucleotide level on a transcript. A peak position satisfies the following conditions: (*i*) The transcript must have both LTM and CHX reads. (*ii*) The position must have at least 10 reads from the LTM data. (*iii*) The position must be a local maximum within seven nucleotides (4). The position must have $R_{LTM} - R_{CHX}$ of at least 0.05, where $R_k = (X_k/N_k) \times 10$ (k = LTM, CHX), X_k is the number of reads on that position in data k, and N_k is the total number of reads on that transcript in data k. Generally, a peak position is also designated a TIS. However, if a peak was not detected on the first position of any AUG or near-cognate start codon but was present at the first position of a codon immediately preceding or succeeding one of these codons, the position was designated a TIS.

Identification of Potentially Misannotated TIS. Among mRNAs with at least one identified dTIS position, those with no aTIS or uTIS peak were selected. Then, the first dTIS in frame 0 was identified as the potentially correct aTIS (pcaTIS). If this dTIS was not associated with an AUG or near-cognate start codon, it was discarded. Any mRNA with a 5′ UTR shorter than 12 nt was excluded, because our method requires at least a 12-nt 5′ UTR to detect the aTIS that would be at the 13th position on a read. To reduce possible false positives, we ensured that (*i*) the total CHX reads in the region from position 1 to pcaTIS position −2 on an mRNA must be less than 10; (*ii*) the maximum CHX reads in this region must be less than 2; (*iii*) total LTM reads from position aTIS−1 to aTIS+1 must be 0; (*iv*) the average CHX read density between pcaTIS−1 and pcaTIS+11 must be higher than 0.1 reads per nucleotide.

Codon Composition Analysis. The number of TIS positions associated with each codon type was counted. The enumeration was done after filtering redundant TIS positions based on its flanking sequence context from −30 to +122 nt relative to the TIS position to avoid double counting of the TIS on the common regions of transcript isoforms. The same redundancy filtering was applied in most other analyses and counting described below. Background codon composition was based on all codons in the annotated coding sequences (CDS) and 5′ UTR of all mRNAs, regardless of reading frame. Redundancy filtering was not performed for background counting.

Measuring False-Positive and False-Negative Rates. To assess false-negative rates under the current $R_{LTM} - R_{CHX}$ threshold of 0.05, we used annotated TIS sites in which the number of CHX reads within five codons downstream of the aTIS was in the top 10th percentile. Of the 2,947 mRNAs, 83.5% have a peak called at the aTIS after the ±1-nt correction. The other 484 mRNAs include 39 mRNAs with 5′ UTR shorter than 12 nt, 102 mRNAs with a dTIS peak within five codons, and 117 mRNAs with a uTIS peak whose associated ORF overlaps the aTIS. Because the last two cases may represent true TIS sites, we computed the lower bound of the false-negative rate as (484 − 102 − 117)/(2,947 − 102 − 117) = 9.7%. We regarded 1 − 83.5% = 16.5% as the upper bound. The upper and lower bounds of false-negative rates are computed for various threshold values in the same manner. To assess false-positive rates, we used the 15,450 mRNAs with no CHX reads within five codons downstream of the aTIS as the set of strictly untranslated aTIS sites. Additionally we considered 21,873 mRNAs with fewer than five CHX reads within the same window. A total of 90 (0.6%) and 1,146 (5.2%) mRNAs, respectively, have a detected peak at each aTIS. The same calculation is applied to other threshold values.

Ribosomal Leaky Scanning Analysis. Three subsets of aTIS positions were collected based on whether the aTIS has the initiation peak and whether the mRNA has any detectable AUG-associated dTIS (Fig. 3*D*). Sequence logos were drawn using Berkeley Weblogo (5). The uTIS positions with the maximum peak height on an mRNA were grouped according to whether the aTIS has a peak [aTIS(Y)] or does not [aTIS(N)], and their Kozak sequence context was analyzed (Fig. 5*A*). For counting the types of uTIS-associated uORFs (Fig. 5*C*), the most downstream uTIS on each mRNA was assigned to one of two groups according to whether the aTIS has a peak [aTIS(Y)] or does not [aTIS(N)]. The same uTIS sets collected for the Kozak sequence context analysis were used to measure the stability of downstream RNA secondary structures. Each of these subsets was divided into three groups according to the initiation context: AUG (Kozak), AUG (non-Kozak) + CUG, and AUG variants + others. The AUG (Kozak) group includes an AUG with either −3A/G or +4G, or both. The AUG (non-Kozak) group is an AUG with neither −3A/G nor +4G. For each TIS position, a window length of 22 nt was moved at step size of 1 nt, starting from −12 nt relative to each uTIS to +100 nt, and the Gibbs free energy (ΔG) was calculated for each window using the RNAfold program (6). The ΔG values were averaged for each position relative to the uTIS across all uTIS positions in each set.

TIS Conservation Between Human and Mouse. Human and mouse RefSeq protein accessions were extracted from HomoloGene (release 65) (7). Each RefSeq protein accession was matched to the associated mRNA accession, CCDS ID, and CCDS amino acid sequence. The amino acid sequences of each homologous protein pair were aligned to each other using Clustalw 2.1 (8) to calculate the alignment score and to filter one-to-one orthologous relationships. If two or more proteins from the same species were in the same HomoloGene group, only the single reciprocally best-matched pair was used. Likewise, if an orthologous gene had mRNA isoforms, the reciprocally best-matched isoform pair was chosen. Any tied matches were removed. The alignment score was computed as [1 − (the number of mismatches and gaps)/(length of human protein)] * 100. Any alignment with an alignment score less than 50 was discarded. The 5′ UTR of an orthologous mRNA was considered as an orthologous 5′ UTR.

Among the human mRNAs that have a mouse ortholog, 5′ UTRs and CDSs were grouped independently into well-aligned and poorly aligned categories. A 5′ UTR with an alignment score less than 50 or with a 3′ end gap of 30 nt or longer was considered poorly aligned. Likewise, a CDS with an initial gap of 30 nt or longer was considered poorly aligned. Note that a CDS with an alignment score less than 50 was discarded beforehand. Within each category, human uTIS or dTIS were classified into five groups, according to sequence conservation (S0 vs. S1) and subtype conservation (T0 vs. T1).

A TIS is conserved in sequence (S1) if there is a mouse TIS peak at the same position on the aligned orthologous mouse sequence or if there is a mouse TIS peak with a similar surrounding sequence. The surrounding sequence is taken from −6 to +24 nt relative to each uTIS. The sequence similarity must be at least 75% identity with no gaps. If a mouse TIS exists in the orthologous 5′ UTR or CDS but is not conserved in sequence, it is assigned to the S0 category. If no mouse TIS exists, it is classified as "N." If the mouse ortholog has no detectable TIS at all, the pair was removed from the analysis.

A TIS is conserved in subtype (T1) if the corresponding mouse uTIS or dTIS is of the same type. For a uTIS, two subtypes, "N-terminal extended" versus "overlapped" and "separated'" were considered. For a dTIS, frame 0 versus frame 1 and frame 2 were used as two subtypes. The priority is set in the order of T1S1, T1S0, T0S1, T0S0, and N, in case aTIS belongs to two or more classes.

Identification of Translated ORFs in Noncoding RNA and Conservation Analysis. Human and mouse noncoding RNAs (ncRNAs) were collected from the RefSeq (release 49) by extracting the RNAs with an accession beginning with "NR" and with no mRNA isoforms. To avoid false detection of TIS positions resulting from spurious mapping of reads sourced from mRNA transcripts, only reads unique to a single ncRNA were used. From the human ncRNAs with at least one identified TIS, the PhastCons score for every nucleotide position within either ORF or non-ORF regions was collected. The PhastCons scores were obtained from the primate subsets of the 46-way vertebrate genomic alignment using the University of California at Santa Cruz Table Browser (http://genome.ucsc.edu) (9, 10). ncRNAs whose genomic positions were ambiguous (e.g., the ncRNA is not included in the refGene table of the UCSC database or for which the length of the RNA is different from the refGene record) were excluded from the analysis.

Plasmid Construction and Immunoblotting. cDNA was synthesized by SuperScript III RT (Invitrogen) using 1 μg of total RNA extracted from HEK293 cells. *CCDC124* and *RND3* genes encompassing both the 5′ UTR and the CDS were amplified by PCR using the following oligo pairs:

ccdc124F: 5′-GGCGCCAAGCTTGGAGGCGCGACCGGGCCGGCGCTGG-3′
ccdc124R: 5′-GGCGCCCTCGAGTTGGGGGCATTGAAGGGCACGGCCC-3′
rnd3F: 5′-GGCGCCAAGCTTCAGTCGGCTCGGAATTGGACTTGGG -3′
rnd3R: 5′-GGCGCCCTCGAGCTATTCTGCACCCTGGAGGCGTAGC-3′

The PCR fragments were cloned to Hind III and Xho I sites of pcDNA3.1/myc-His B. Plasmid transfection was performed using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. After 48 h transfection, cells were lysed by the lysis buffer [Tris-buffered saline (pH 7.4), 2% Triton X-100]. The whole-cell lysates were heat-denatured for 10 min in NuPAGE LDS Sample Buffer (Invitrogen). The protein samples were resolved on 12% NuPAGE gel (Invitrogen) and then were transferred to Immobilon-P membranes (Millipore). After blocking for 1 h in TBS containing 5% blotting milk, membranes were incubated with c-*myc* antibodies (Santa Cruz Biotechnology) at 4 °C overnight. After incubation with HRP-coupled secondary antibodies (Sigma), immunoblots were developed using enhanced chemiluminescence (GE Healthcare).

1. Ju J, et al. (2009) Lactimidomycin, iso-migrastatin and related glutarimide-containing 12-membered macrolides are extremely potent inhibitors of cell migration. *J Am Chem Soc* 131:1370–1371.
2. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324: 218–223.
3. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25.
4. Pruitt KD, et al. (2009) The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res* 19:1316–1323.
5. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: A sequence logo generator. *Genome Res* 14:1188–1190.
6. Gruber AR, Lorenz R, Bernhart SH, Neubock R, Hofacker IL (2008) The Vienna RNA websuite. *Nucleic Acids Res* 36(Web Server issue):W70– W 74.
7. Sayers EW, et al. (2011) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 39(Database issue):D38–D51.
8. Larkin MA, et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23: 2947–2948.
9. Karolchik D, et al. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 32(Database issue):D493–D496.
10. Kent WJ, et al. (2002) The human genome browser at UCSC. *Genome Res* 12: 996–1006.

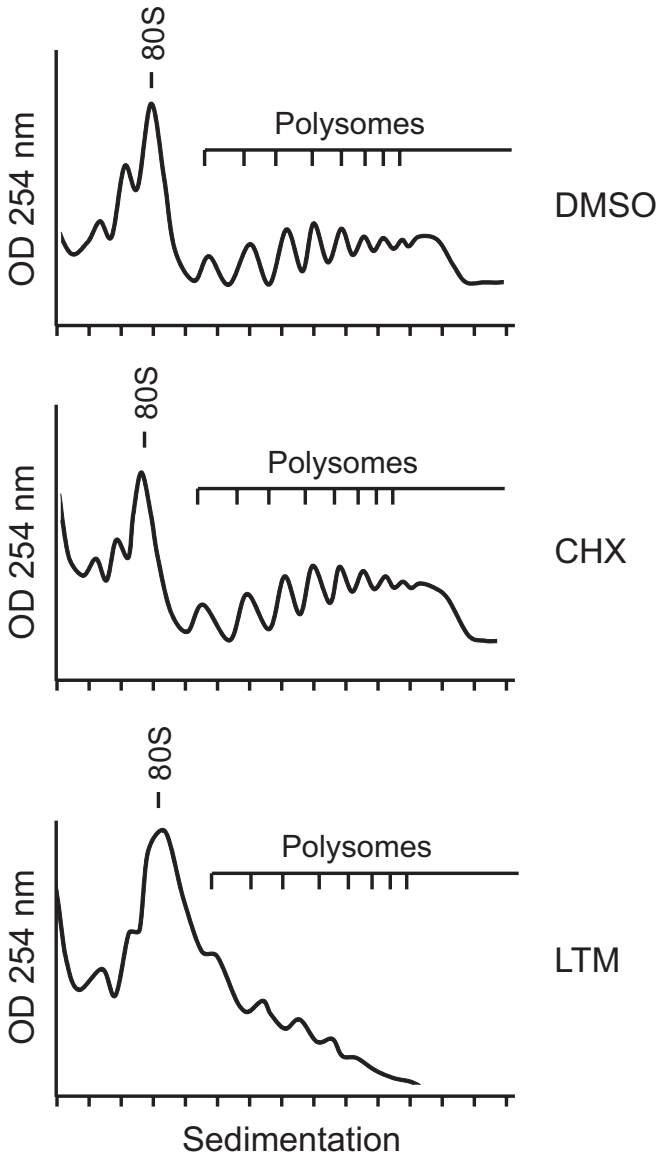


Fig. S1. Polysome profile analysis in cells treated with ribosome exit-site translation inhibitors. HEK293 cells were pretreated with equal volume of DMSO, 100 μM CHX, or 50 μM LTM for 30 min followed by sucrose gradient sedimentation. Both 80S monosome and polysome peaks are indicated.

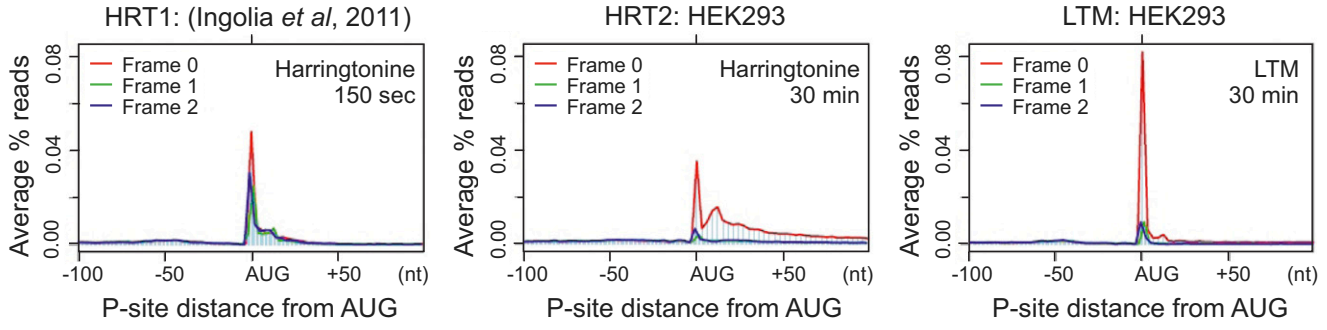


Fig. S2. Metagene analysis of RPFs obtained using different approaches. RPF reads reported by Ingolia et al. (1) using harringtonine in mouse embryonic stem cells were replotted after peptidyl (P)-site adjustment based on the original report (HRT1, *Left*). RPF reads obtained from HEK293 cells treated with either harringtonine (HRT2, *Center*) or LTM (*Right*) were plotted by applying a 12-nt offset to reads with a length range of 26–29 nt. All mapped reads are aligned at the annotated start codon AUG, and the reads density at each nucleotide position is averaged using the P-site of RPFs.

1. Ingolia NT, Lareau LF, Weissman JS (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 147:789–802.

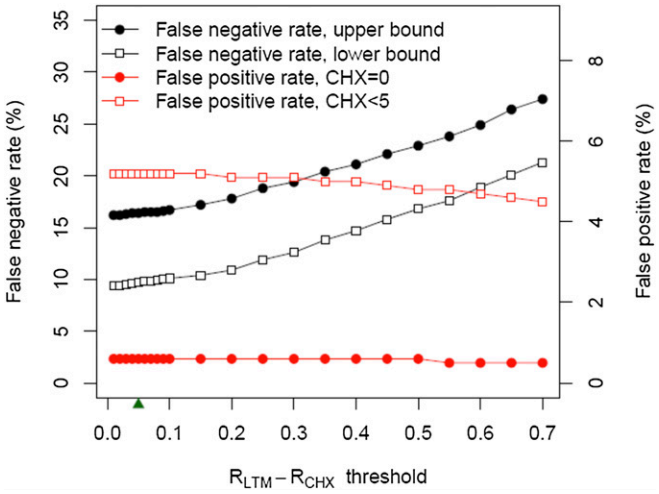


Fig. S3. False-positive and false-negative rates at various $R_{LTM} - R_{CHX}$ thresholds. The false-negative rate is computed as the percentage of undetected aTIS among the top 10% translated aTIS codons based on CHX reads within five codons downstream of the aTIS. The lower and upper bounds of false-negative rate are determined by either including or excluding the cases having a dTIS within five codons and/or a uORF overlapping aTIS. The false-positive rate is computed as the percentage detected among strictly untranslated aTIS codons with either no CHX reads (CHX = 0) or fewer than five CHX reads (CHX < 5) within five codons downstream of the aTIS.

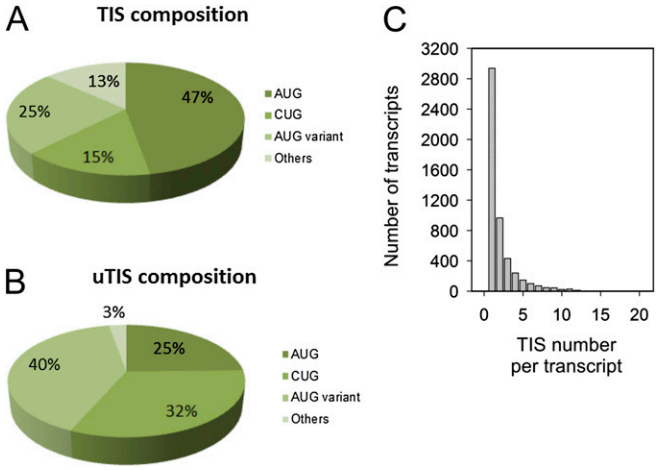
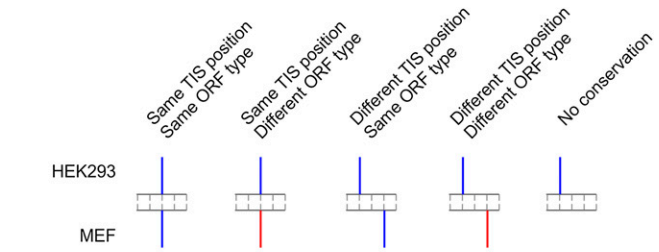


Fig. S4. Global TIS identification in MEF cells. (*A*) Codon composition of all TIS codons identified by global translation initiation sequencing (GTI-seq) in MEF cells. (*B*) Codon composition of uTIS codons identified by GTI-seq in MEF cells. (*C*) Histogram showing the overall distribution of the number of TIS positions identified on each transcript from MEF cells.



	Sequence similarity					
uTIS	H	19%	1%	58%	7%	15%
	L	8%	0%	47%	9%	36%
dTIS	H	15%	0%	32%	13%	39%
	L	9%	0%	29%	8%	54%

Fig. S5. Conservation of alternative TIS positions between human and mouse cells. Alternative TIS positions identified on human mRNAs are classified based on whether the position, sequence context, or ORF type is conserved in the mouse orthologous mRNAs (same color represents same type). TIS sites with a mouse counterpart at the identical position or with a similar local sequence context on the aligned orthologous sequences are merged. uTIS and dTIS positions are classified into two subsets each according to the global alignment score of sequences (5′ UTR for uTIS and CDS for dTIS). Percentage values are presented in the table.

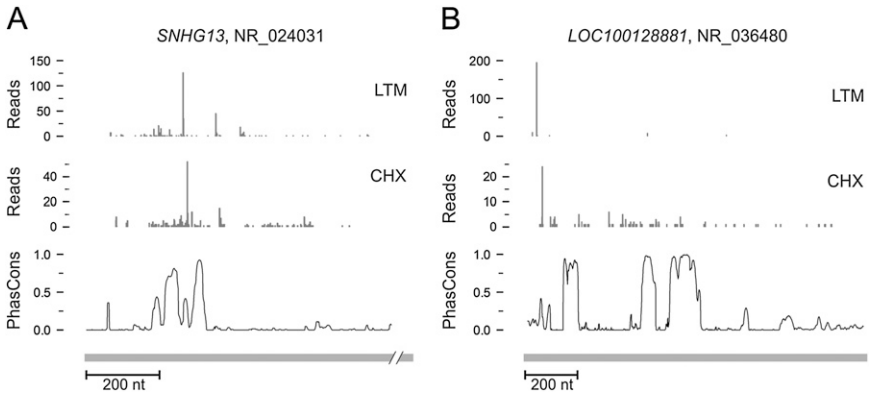


Fig. S6. ORF conservation in ncRNAs. (*A*) Translation in ncRNA *SNHG13* is illustrated by LTM- and CHX-associated RPF reads. PhastCons scores retrieved from the primate genome sequence alignment are plotted also (*B*) Translation in ncRNA *LOC100128881* is illustrated by LTM- and CHX-associated RPF reads. PhastCons scores retrieved from the primate genome sequence alignment are plotted also.

Dataset S1. TIS positions identified in HEK293 cells

Dataset S1

Dataset S2. Genes with possible misannotation

Dataset S2

Dataset S3. TIS positions identified in MEF cells

Dataset S3

Dataset S4. TIS positions identified in ncRNAs

Dataset S4