Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution

Soonchel Lee[a,1,2], Botao Liu[b,1], Soohyun Lee[c,1,3], Sheng-Xiong Huang[d], Ben Shen[d,e], and Shu-Bing Qian[a,b,4]

[a]Division of Nutritional Sciences and [b]Graduate Field of Genetics, Genomics and Development, Cornell University, Ithaca, NY 14853; [c]Proboco Informatics, Ithaca, NY 14850; [d]Department of Chemistry and [e]Department of Molecular Therapeutics, The Scripps Research Institute, Jupiter, FL 33458

Edited* by Jonathan S. Weissman, University of California, San Francisco, CA, and approved July 24, 2012 (received for review May 9, 2012)

Understanding translational control in gene expression relies on precise and comprehensive determination of translation initiation sites (TIS) across the entire transcriptome. The recently developed ribosome-profiling technique enables global translation analysis, providing a wealth of information about both the position and the density of ribosomes on mRNAs. Here we present an approach, global translation initiation sequencing, applying in parallel the ribosome E-site translation inhibitors lactimidomycin and cycloheximide to achieve simultaneous detection of both initiation and elongation events on a genome-wide scale. This approach provides a view of alternative translation initiation in mammalian cells with single-nucleotide resolution. Systemic analysis of TIS positions supports the ribosome linear-scanning mechanism in TIS selection. The alternative TIS positions and the associated ORFs identified by global translation initiation sequencing are conserved between human and mouse cells, implying physiological significance of alternative translation. Our study establishes a practical platform for uncovering the hidden coding potential of the transcriptome and offers a greater understanding of the complexity of translation initiation.

genome wide | high throughput | leaky scanning | start codon

Protein synthesis is the final step in the flow of genetic information and lies at the heart of cellular metabolism. Translation is regulated principally at the initiation stage, and during the last decade significant progress has been made in dissecting the role of initiation factors (eIFs) in the assembly of elongation-competent 80S ribosomes (1–3). However, mechanisms underlying start codon recognition are not fully understood. Proper selection of the translation initiation site (TIS) on mRNAs is crucial for the production of desired protein products. A fundamental and long-sought goal in understanding translational regulation is the precise determination of TIS codons across the entire transcriptome.

In eukaryotes, ribosomal scanning is a well-accepted model for start codon selection (4). During cap-dependent translation initiation, the small ribosome subunit (40S) is recruited to the 5′ end of mRNA (the m^7G cap) in the form of a 43S preinitiation complex (PIC). The PIC is thought to scan along the message in search of the start codon. It is commonly assumed that the first AUG codon that the scanning PIC encounters serves as the start site for translation. However, many factors influence the start codon selection. For instance, the initiator AUG triplet usually is in an optimal context, with a purine at position −3 and a guanine at position +4 (5). The presence of an mRNA secondary structure at or near the TIS position also influences the efficiency of recognition (6). In addition to these *cis* sequence elements, the stringency of TIS selection also is subject to regulation by *trans- acting* factors such as eIF1 and eIF1A (7, 8). Inefficient recognition of an initiator codon results in a portion of 43S PIC continuing to scan and initiating translation at a downstream site, a process known as "leaky scanning" (4). However, little is known about the frequency of leaky scanning events at the transcriptome level.

Many recent studies have uncovered a surprising variety of potential translation start sites upstream of the annotated coding sequence (CDS) (9, 10). It has been estimated that about 50% of mammalian transcripts contain at least one upstream ORF (uORF) (11, 12). Intriguingly, many non-AUG triplets have been reported to act as alternative start codons for initiating uORF translation (13). Because there is no reliable way to predict non-AUG codons as potential initiators from *in silico* sequence analysis, there is an urgent need to develop experimental approaches for genome-wide TIS identification.

Ribosome profiling, based on deep sequencing of ribosome-protected mRNA fragments (RPF), has proven to be powerful in defining ribosome positions on the entire transcriptome (14, 15). However, the standard ribosome profiling is not suitable for identifying TIS. Elevated ribosome density near the beginning of CDS is not sufficient for unambiguous identification of alternative TIS positions, in particular the TIS positions associated with overlapping ORFs. To overcome this problem, a recent study used an initiation-specific translation inhibitor, harringtonine, to deplete elongating ribosomes from mRNAs (16). This approach uncovered an unexpected abundance of alternative TIS codons, in particular non-AUG codons in the 5′ UTR. However, because the inhibitory mechanism of harringtonine on the initiating ribosome is unclear, whether the harringtonine-marked TIS codons truly represent physiological TIS remains to be confirmed.

We developed a technique, global translation initiation sequencing (GTI-seq), that uses two related but distinct translation inhibitors to differentiate ribosome initiation from elongation effectively. GTI-seq has the potential to reveal a comprehensive and unambiguous set of TIS codons at nearly single-nucleotide resolution. The resulting TIS maps provide a remarkable display of alternative translation initiators that vividly delineates the variation in start codon selection. This technique allows a more complete assessment of the underlying principles that specify start codon use in vivo.

Results

Experimental Design. Cycloheximide (CHX) has been widely used in ribosome profiling of eukaryotic cells because of its potency in stabilizing ribosomes on mRNAs. Both biochemical (17) and structural studies (18) revealed that CHX binds to the exit (E)-

Author contributions: Soonchel Lee, B.L., and S.-B.Q. designed research; Soonchel Lee and B.L. performed research; S.-X.H. and B.S. contributed new reagents/analytic tools; Soonchel Lee, Soohyun Lee, and S.-B.Q. analyzed data; and Soohyun Lee and S.-B.Q. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

Data deposition: The sequences reported in this work have been deposited in the Sequence Read Archive database (accession no. SRA056377).

[1]Soonchel Lee, B.L., and Soohyun Lee contributed equally to this work.

[2]Present address: Cancer Center, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02115.

[3]Present address: Center for Biomedical Informatics, Harvard Medical School, Boston, MA 02115.

[4]To whom correspondence should be addressed. E-mail: sq38@cornell.edu.

See Author Summary on page 14728 (volume 109, number 37).

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1207846109/-/DCSupplemental.

site of the large ribosomal subunit, close to the position where the 3′ hydroxyl group of the deacylated transfer RNA (tRNA) normally binds. CHX thus prevents the release of deacylated tRNA from the E-site and blocks subsequent ribosomal translocation (Fig. 1*A*, *Left*). Recently, a family of CHX-like natural products isolated from *Streptomyces* was characterized, including lactimidomycin (LTM) (19, 20). Acting as a potent protein synthesis inhibitor, LTM uses a mechanism similar but not identical to that used by CHX (17). With its 12-member macrocycle, LTM is significantly larger in size than CHX (Fig. 1*A*). As a result, LTM cannot bind to the E-site when a deacylated tRNA is present. Only during the initiation step, in which the initiator tRNA enters the peptidyl (P)-site directly (21), is the empty E-site accessible to LTM. Thus, LTM acts preferentially on the initiating ribosome but not on the elongating ribosome. We reasoned that ribosome profiling using LTM in a side-by-side comparison with CHX should allow a complete segregation of the ribosome stalled at the start codon from the one in active elongation (Fig. 1*B*).

We designed an integrated GTI-seq approach and performed the ribosome profiling in HEK293 cells pretreated with either LTM or CHX. Although CHX stabilized the polysomes slightly compared with the no-drug treatment (DMSO), 30 min of LTM treatment led to a large increase in monosomes accompanied by a depletion of polysomes (Fig. S1). This result is in agreement with the notion that LTM halts translation initiation while allowing elongating ribosomes to run off (17). After RNase I digestion of the ribosome fractions, the purified RPFs were subjected to deep sequencing. As expected, CHX treatment resulted in an excess of RPFs at the beginning of ORFs in addition to the body of the CDS (Fig. 1*C*). Remarkably, LTM treatment led to a pronounced single peak located at the −12-nt position relative to the annotated start codon. This position corresponds to the ribosome P-site at the AUG codon when an offset of 12 nt is considered (14, 15). LTM treatment also eliminated the excess of ribosomes seen at the stop codon in untreated cells or in the presence of CHX. Therefore, LTM efficiently stalls the 80S ribosome at the start codons.

During the course of our study, Ingolia et al. (16) reported a similar TIS mapping approach using harringtonine, a different translation initiation inhibitor. One key difference between harringtonine and LTM is that the former drug binds to free 60S subunits (22), whereas LTM binds to the 80S complexes already assembled at the start codon (17). We compared the pattern of RPF density surrounding the annotated start codon in the published datasets (16) and the LTM results (Fig. S2). It appears that a considerable amount of harringtonine-associated RPFs are not located exactly at the annotated start codon. To compare the accuracy of TIS mapping accuracy by LTM and harringtonine directly, we performed ribosome profiling in HEK293 cells treated with harringtonine using the same protocol as in LTM treatment. As in the previous study, harringtonine treatment caused a substantial fraction of RPFs to accumulate in regions downstream of the start codon (Fig. 1*D*). The relaxed positioning of harringtonine-associated RPFs after prolonged treatment leaves uncertainty in TIS mapping. In contrast, GTI-seq using LTM largely overcomes this deficiency and offers high precision in global TIS mapping with single-nucleotide resolution (Fig. 1*D*).

Global TIS Identification by GTI-seq. One of the advantages of GTI-seq is its ability to analyze LTM data in parallel with CHX. Because of the structural similarity between these two translation inhibitors, the LTM background reads resembled the pattern of CHX-associated RPFs (Fig. 2*A*). This feature allows us to reduce the background noise of LTM-associated RPFs further by subtracting the normalized density of CHX reads at every nucleotide position from the density of LTM reads at that position. A TIS peak then is called at a position in which the adjusted LTM reads density is well above the background (red asterisk in Fig. 2*A*; see *Materials and Methods* for details). From ∼10,000 transcripts with

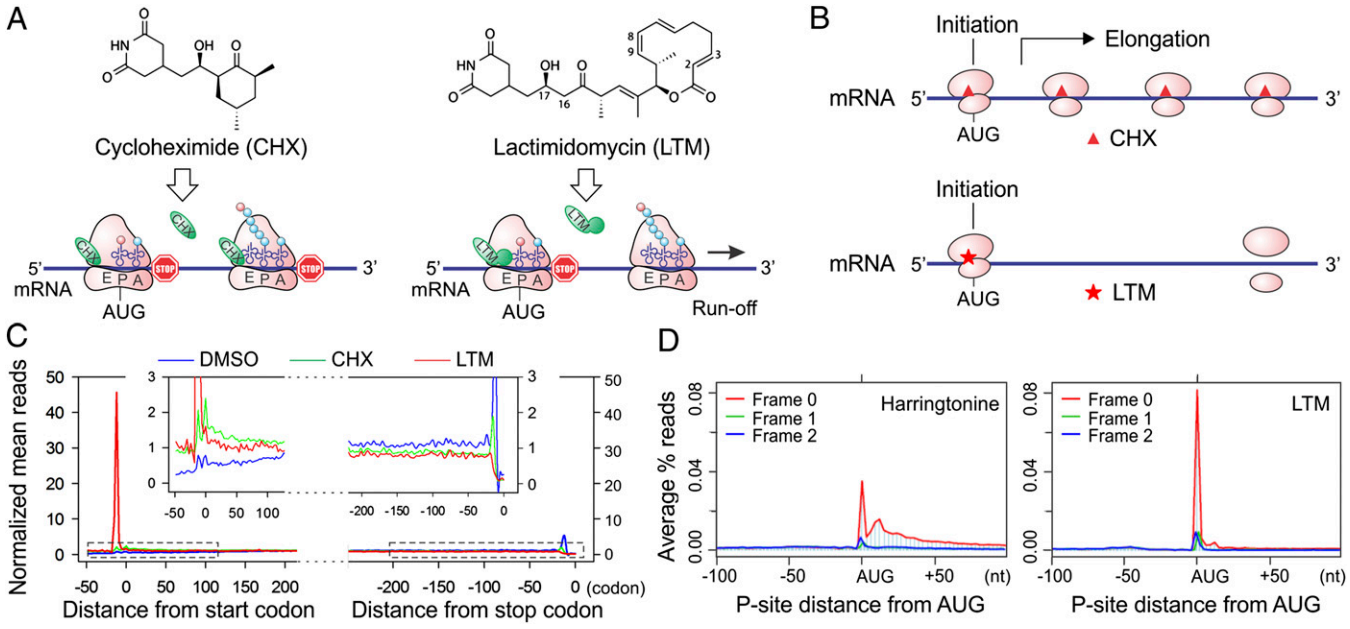


Fig. 1. Experimental strategy of GTI-seq using ribosome E-site translation inhibitors. (*A*) Schematic diagram of the experimental design for GTI-seq. Translation inhibitors CHX and LTM bind to the ribosome E-site, resulting in inhibition of translocation. CHX binds to all translating ribosomes (*Left*), but LTM preferentially incorporates into the initiating ribosomes when the E-site is free of tRNA (*Right*). (*B*) Ribosome profiling using CHX and LTM side by side allows the initiating ribosome to be distinguished from the elongating one. (*C*) HEK293 cells were treated with DMSO, 100 μM CHX, or 50 μM LTM for 30 min before ribosome profiling. Normalized RPF reads are averaged across the entire transcriptome, aligned at either their start site or stop codon from the 5′ end of RPFs. (*D*) Metagene analysis of RPFs obtained from HEK293 cells treated with harringtonine (*Left*) or LTM (*Right*). All mapped reads are aligned at the annotated start codon AUG, and the density of reads at each nucleotide position is averaged using the P-site of RPFs.

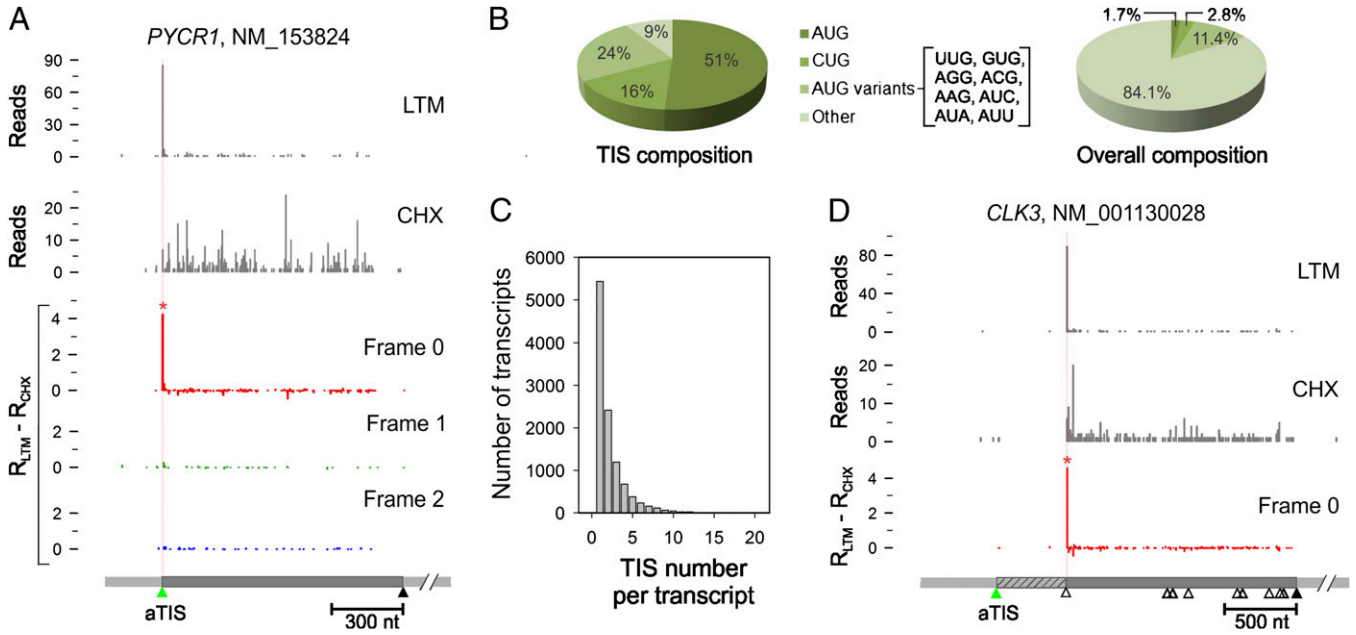


Fig. 2. Global identification of TIS by GTI-seq. (*A*) TIS identification on the *PYCR1* transcript. LTM and CHX reads are plotted as gray bar graphs. TIS identification is based on normalized density of LTM reads minus the density of CHX reads. The three reading frames are separated and presented as distinct colors. The identified TIS position is marked by a red asterisk and highlighted by a vertical line color-coded by the corresponding reading frame. The annotated coding region is indicated by a green triangle (start codon) and a black triangle (stop codon). (*B*) Codon composition of all TIS codons identified by GTI-seq (*Left*) is shown in comparison with the overall codon distribution over the entire transcriptome (*Right*). (*C*) Histogram showing the overall distribution of TIS numbers identified on each transcript. (*D*) Misannotation of the start codon on the *CLK3* transcript. The annotated coding region is indicated by the green (start codon) and black (stop codon) triangles. AUG codons on the body of the coding region are also shown as open triangles. For clarity, only one reading frame is shown.

detectable TIS peaks, we identified a total of 16,863 TIS sites (Dataset S1). Codon composition analysis revealed that more than half the TIS codons used AUG as the translation initiator (Fig. 2*B*). GTI-seq also identified a significant proportion of TIS codons using near-cognate codons that differ from AUG by a single nucleotide, in particular CUG (16%). Remarkably, nearly half the transcripts (49.6%) contained multiple TIS sites (Fig. 2*C*), suggesting that alternative translation prevails even under physiological conditions. Surprisingly, over a third of the transcripts (42.3%) showed no TIS peaks at the annotated TIS position (aTIS) despite clear evidence of translation (Dataset S1). Although some could be false negatives resulting from the stringent threshold cutoff for TIS identification (Fig. S3), others were attributed to alternative translation initiation (see below). However, it is possible that some cases represent misannotation. For instance, the translation of *CLK3* clearly starts from the second AUG, although the first AUG was annotated as the initiator in the current database (Fig. 2*D*). We found 50 transcripts that have possible misannotation in their start codons (Dataset S2). However, some mRNAs might have alternative transcript processing. In addition, we could not exclude the possibility that some of these genes might have tissue-specific TIS.

Characterization of Downstream Initiators. In addition to validating initiation at the annotated start codon, GTI-seq revealed clear evidence of downstream initiation on 27% of the analyzed transcripts with TIS peaks (Dataset S1). As a typical example, *AIMP1* showed three TIS peaks exactly at the first three AUG codons in the same reading frame (Fig. 3*A*). Thus, the same transcript generates three isoforms of AIMP1 with varied NH_2 termini, a finding that is consistent with the previous report (23). Of the total TIS positions identified by GTI-seq, 22% (3,741/16,863) were located downstream of aTIS codons; we termed these positions "dTIS." Nearly half of the identified dTIS codons used AUG as the initiator (Fig. 3*B*).

What are the possible factors influencing downstream start codon selection? We classified genes with multiple TIS codons into three groups based on the Kozak consensus sequence of the first AUG. The relative leakiness of the first AUG codon was estimated by measuring the fraction of LTM reads at the first AUG over the total reads recovered on and after this position. The AUG codon with a strong Kozak sequence context showed higher initiation efficiency (or lower leakiness) than a codon with a weak or no consensus sequence ($P = 1.12 \times 10^{-142}$) (Fig. 3*C*). These results indicate the critical role of sequence context in start codon recognition. To substantiate this conclusion further, we performed a reciprocal analysis by grouping genes according to whether an initiation peak was identified at the aTIS or dTIS positions on their transcripts (Fig. 3*D*). A survey of the sequences flanking the aTIS revealed a clear preference of Kozak sequence context for different gene groups. We observed the strongest Kozak consensus sequence in the gene group with aTIS initiation but no detectable dTIS, (Fig. 3*D*, *Bottom*). This sequence context was largely absent in the group of genes lacking detectable translation initiation at the aTIS (Fig. 3*D*, *Top*). Thus, ribosome leaky scanning tends to occur when the context for an aTIS is suboptimal.

Cells use the leaky scanning mechanism to generate protein isoforms with changed subcellular localizations or altered functionality from the same transcript (24). GTI-seq revealed many more genes that produce protein isoforms via leaky scanning than had been previously reported (Dataset S1). For independent validation of the dTIS positions identified by GTI-seq, we cloned the gene *CCDC124* whose transcript showed several initiation peaks above the background (Fig. 3*E*). One dTIS is in the same reading frame as the aTIS, allowing us to use a COOH-terminal tag to detect different translational products in transfected cells. Immunoblotting of transfected HEK293 cells showed two clear bands

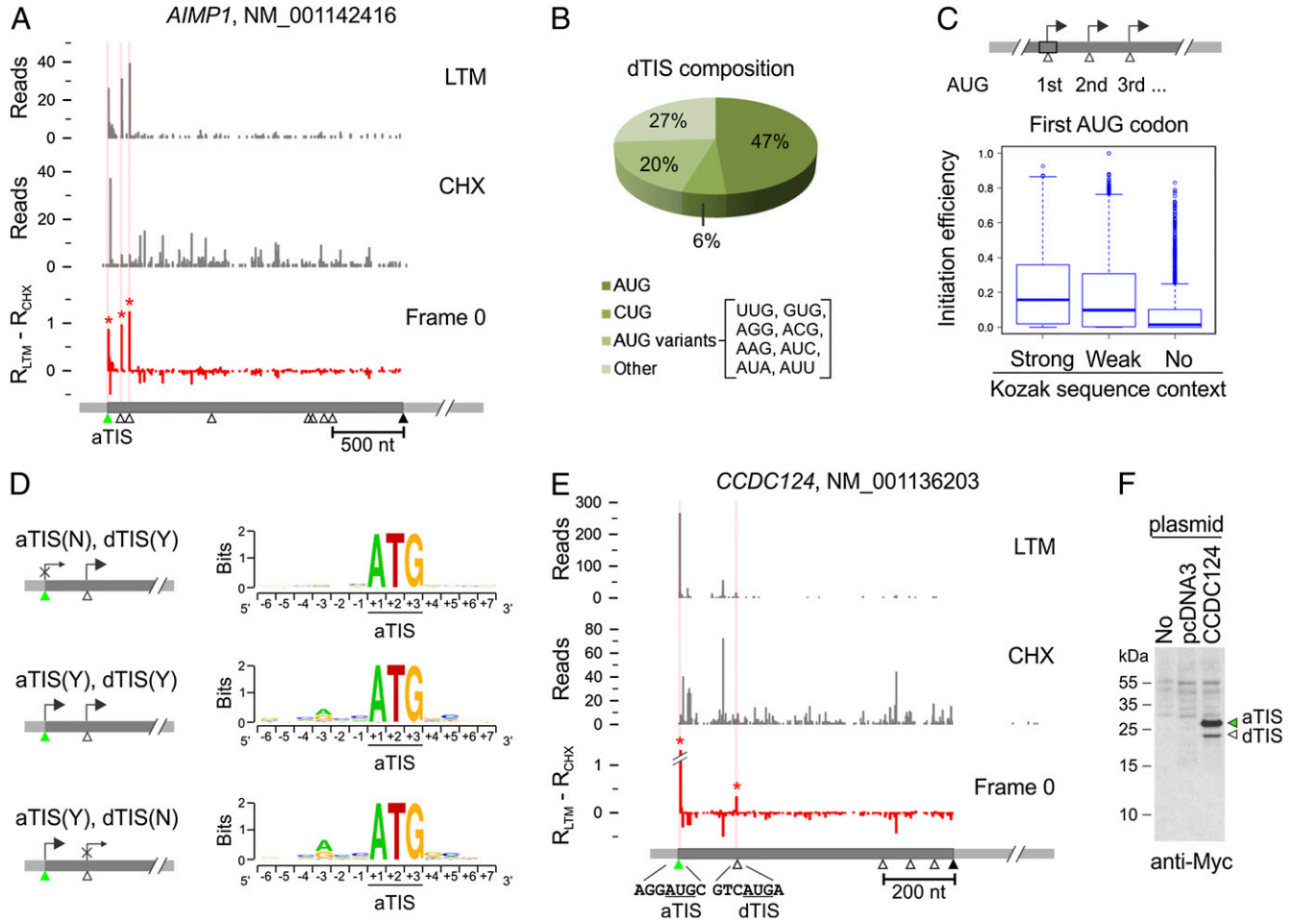


Fig. 3. Characterization of dTIS. (*A*) Identification of multiple TIS codons on the *AIMP1* transcript. For clarity, only one reading frame is shown. (*B*) Codon composition of total dTIS codons identified by GTI-seq. (*C*) Relative efficiency of initiation at the first AUG codon with different Kozak sequence contexts (one-tailed Wilcoxon rank sum test: strong vs. weak: $P = 7.92 \times 10^{-24}$; weak vs. no Kozak context: $P = 1.34 \times 10^{-75}$). (*D*) Genes are grouped according to the identified initiation at an aTIS, at a dTIS, or at both. The sequence context surrounding the aTIS is shown as sequence logos. χ^2 test, $P = 2.57 \times 10^{-100}$ for the −3 position and $P = 3.95 \times 10^{-18}$ for the +4 position. (*E*) Identification of multiple TIS codons on the *CCDC124* transcript. (*F*) Validation of *CCDC124* TIS codons by immunoblotting. The DNA fragment encompassing both the 5′ UTR and the CDS of *CCDC124* was cloned and transfected into HEK 293 cells. Whole-cell lysates were immunoblotted using *c-myc* antibody.

whose molecular masses correspond to full-length CCDC124 (28.9 kDa) and the NH_2-terminally truncated isoform (23.7 kDa), respectively. Intriguingly, the relative abundance of both isoforms matched well to the density of corresponding LTM reads, suggesting that GTI-seq might provide quantitative assessment of translation initiation.

Characterization of Upstream Initiators. Sequence-based computational analyses predicted that about 50% of mammalian transcripts contain at least one uORF (11, 12). In agreement with this notion, GTI-seq revealed that 54% of transcripts bear one or more TIS positions upstream of the annotated start codon (Dataset S1). These upstream TIS (uTIS) codons, when outside the aTIS reading frame, often are associated with short ORFs. A classic example is *ATF4*, whose translation is controlled predominantly by several uORFs (25–27). This feature was clearly captured by GTI-seq (Fig. 4*A*). As expected, the presence of these uORFs efficiently repressed the initiation at the aTIS, as evidenced by few CHX reads along the CDS of *ATF4*.

Nearly half of the total TIS positions identified by GTI-seq were uTIS (7,936/16,863). In contrast to the dTIS, which used AUG as the primary start codon (Fig. 3*B*), the majority of uTIS (74.4%) were non-AUG codons (Fig. 4*B*). CUG was the most prominent of these AUG variants, with a frequency even higher than that of AUG (30.3% vs. 25.6%). In a few well-documented examples, the CUG triplet was reported to serve as an alternative initiator (13). To confirm experimentally the alternative initiators identified by GTI-seq, we cloned the gene *RND3* that showed a clear initiation peak at a CUG codon in addition to the aTIS (Fig. 4*C*). The two initiators are in the same reading frame without a stop codon between them, thus permitting us to detect different translational products using an antibody against the fused COOH-terminal tag. Immunoblotting of transfected HEK293 cells showed two protein bands corresponding to the CUG-initiated long isoform (34 kDa) and the main product (31 kDa) (Fig. 4*C*). Once again, the levels of both isoforms were in accordance with the relative densities of LTM reads, further supporting the quantitative feature of GTI-seq in TIS mapping.

Global Impacts of uORFs on Translational Efficiency. Initiation from an uTIS and the subsequent translation of the short uORF negatively influence the main ORF translation (10, 11). To find possible factors governing the alternative TIS selection in the 5′ UTR, we categorized uTIS-bearing transcripts into two groups according to whether initiation occurs at the aTIS and compared the sequence context of uTIS codons (Fig. 5*A*). For transcripts with initiation at

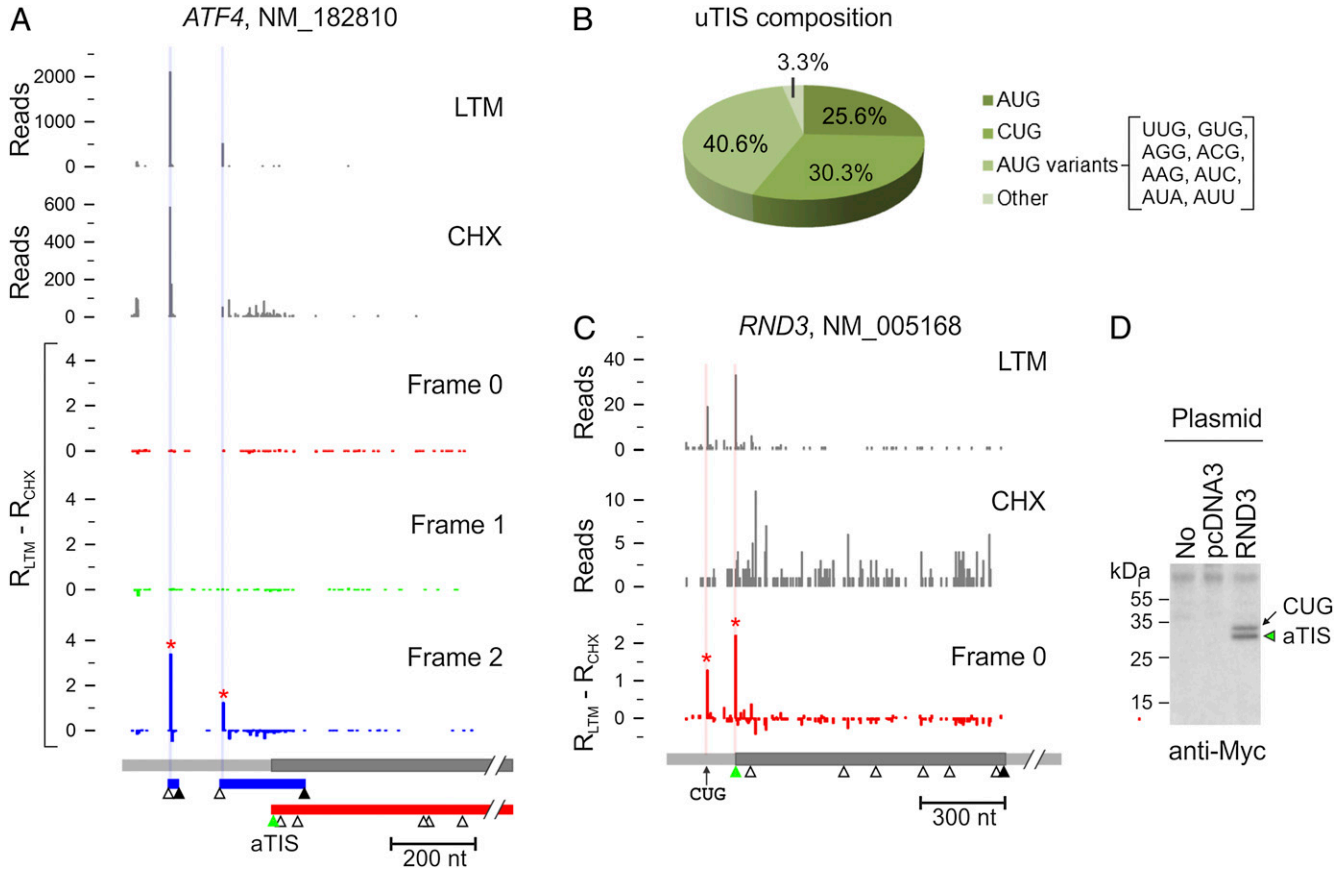


Fig. 4. Characterization of uTIS. (*A*) Identification of multiple TIS codons on the *ATF4* transcript. *Inset* shows a region of frame 0 with the *y* axis enlarged 10-fold, showing the LTM peak at the annotated start codon AUG. Different ORFs are shown in boxes color-coded for the different reading frames. (*B*) Codon composition of total uTIS codons identified by GTI-seq. (*C*) Identification of multiple TIS codons on the *RND3* transcript. (*D*) Validation of *RND3* TIS codons by immunoblotting. The DNA fragment encompassing both the 5′ UTR and the CDS of *RND3* was cloned and transfected into HEK 293 cells. Whole-cell lysates were immunoblotted using c-*myc* antibody.

both uTIS and aTIS positions [aTIS(Y)], the uTIS codons were preferentially composed of nonoptimal AUG variants. In contrast, the uTIS codons identified on transcripts with repressed aTIS initiation [aTIS(N)] showed a higher percentage of AUG with Kozak consensus sequences ($P = 1.74 \times 10^{-80}$). These results are in agreement with the notion that the accessibility of an aTIS to the ribosome for initiation depends on the context of uTIS codons.

Recent work showed a correlation between secondary structure stability of local mRNA sequences near the start codon and the efficiency of mRNA translation (28–30). To examine whether the uTIS initiation also is influenced by local mRNA structures, we computed the free energy associated with secondary structures from regions surrounding the uTIS position (Fig. 5*B*). We observed an increased folding stability of the region shortly after the uTIS in transcripts with repressed aTIS initiation (Fig. 5*B*, blue line). In particular, more stable mRNA secondary structures were present on transcripts with less optimal uTIS codons (Fig. 5*B*, *Center* and *Right*). Therefore, when the consensus sequence is absent from the start codon, the local mRNA secondary structure has a stronger correlation with the TIS selection.

Depending on the uTIS positions, the associated uORF can be separated from or overlap the main ORF. These different types of uORF could use different mechanisms to control the main ORF translation. For instance, when the uORF is short and separated from the main ORF, the 40S subunit can remain associated with the mRNA after termination at the uORF stop codon and can resume scanning, a process called "reinitiation" (2). When the uORF overlaps the main ORF, the aTIS initiation relies solely on the leaky scanning mechanism. We sought to dissect the respective contributions of reinitiation and leaky scanning to the regulation of aTIS initiation. Interestingly, we found a higher percentage of separated uORFs in aTIS(N) transcripts (Fig. 5*C*, $P = 3.52 \times 10^{-41}$). This result suggests that the reinitiation generally is less efficient than leaky scanning and is consistent with the negative role of uORFs in translation of main ORFs.

Cross-Species Conservation of Alternative Translation Initiators. The prevalence of alternative translation reshapes the proteome landscape by increasing the protein diversity or by modulating translation efficiency. The biological significance of alternative initiators could be preserved across species if they are of potential fitness benefit. We applied GTI-seq to a mouse embryonic fibroblast (MEF) cell line and identified TIS positions, including uTIS and dTIS, across the mouse transcriptome (Dataset S3). MEF cells showed remarkable similarity to HEK293 cells in overall TIS features (Fig. S4). For example, uTIS codons used non-AUG, especially CUG, as the dominant initiator. Additionally, about half the transcripts in MEF cells exhibited multiple initiators. Thus, the general features of alternative translation are well conserved between human and mouse cells.

To analyze the conservation of individual alternative TIS position on each transcript, we chose a total of 12,949 human/mouse orthologous mRNA pairs. We analyzed the 5′ UTR and CDS regions separately to measure the conservation of uTIS and dTIS positions, respectively (Fig. 6*A*). Each group was classified into two subgroups based on their sequence similarity. For genes with high

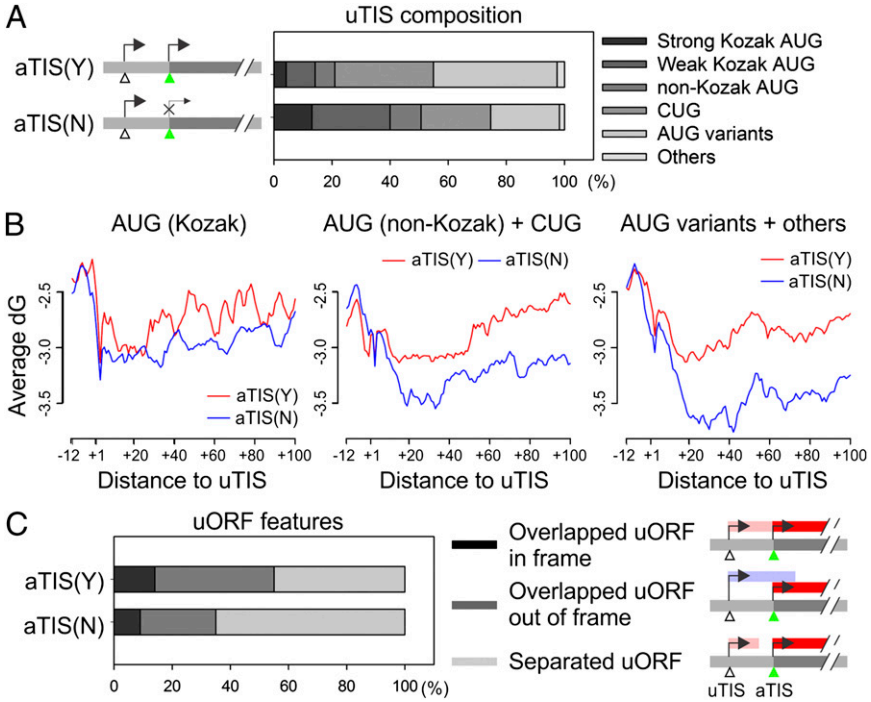


Fig. 5. Impact of uORF features on translational regulation. (*A*) The sequence composition of uTIS codons for genes with [aTIS(Y)] or without [aTIS(N)] aTIS initiation. Genes are classified into two groups based on aTIS initiation, and the uTIS sequence composition is categorized based on the consensus features shown on the right. (*B*) The contribution of mRNA secondary structure to TIS selection. Genes are grouped based on uTIS codon features listed in *A*. For each group, the transcripts with (red line) or without (blue line) aTIS initiation are analyzed for the averaged Gibbs free energy (ΔG) value in regions surrounding the identified uTIS codons. (*C*) The composition of uORFs in gene groups with or without aTIS initiation on their transcripts. Different ORF features are shown on the right.

sequence similarity, 85% of the uTIS and 60% of dTIS positions were conserved between human and mouse cells. Some of these alternative TIS codons were located at the same positions on the aligned sequences (Fig. S5). For example, *RNF10* in HEK293 cells showed three uTIS positions, which also were found at the identical positions on the aligned 5′ UTR sequence of the mouse homolog in MEF cells (Fig. 6*B*). Remarkably, genes with low sequence similarity also displayed high TIS conservation across the two species (Fig. 6*A*). For instance, the 5′ UTR of the *CTTN* gene has low sequence identity between human and mouse homologs (alignment score = 40.3) (Fig. 6*C*). However, a clear uTIS was identified at the same position on the aligned region in both cells. Notably, the majority of alternative ORFs conserved between human and mouse cells were of the same type, i.e., either separated from or overlapping the main ORF (Fig. 6*A* and Fig. S5). The evolutionary conservation of those TIS positions and the associated ORFs is a strong indication of the functional significance of alternative translation in regulating gene expression.

Characterization of Non-Protein Coding RNA Translation. The mammalian transcriptome contains many non–protein-coding RNAs (ncRNAs) (31). ncRNAs have gained much attention recently because of increasing recognition of their role in a variety of cellular processes, including embryogenesis and development (32). Motivated by the recent report of the possible translation of large intergenic ncRNAs (16), we sought to explore the possible translation, or at least ribosome association, of ncRNAs in HEK293 cells. We selected RPFs uniquely mapped to ncRNA sequences to exclude the possibility of spurious mapping of reads originated from mRNAs. Of 5,763 ncRNAs annotated in RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/), we identified 228 ncRNAs (about 4%) that were associated with RPFs marked by both CHX and LTM (Fig. 6*D* and Dataset S4). Compared with protein-coding mRNAs, most ORFs recovered from ncRNAs were very short, with a median length of 54 nt (Fig. 6*E*). Several ncRNAs also showed alternative initiation at non-AUG start codons, as exemplified by *LOC100506233* (Fig. 6*F*).

Comparative genomics reveals that the coding regions often are evolutionarily conserved elements (33). We retrieved the PhastCons scores (http://genome.ucsc.edu) for both coding and noncoding regions of ncRNAs and found that the ORF regions identified by GTI-seq indeed showed a higher conservation (Fig. 6*G*). Some ncRNAs showed a clear enrichment of highly conserved bases within the ORFs marked by both LTM and CHX reads (Fig. S6). Despite the apparent engagement by the protein synthesis machinery, the physiological functions of the coding capacity of these ncRNAs remain to be determined.

Discussion

The mechanisms of eukaryotic translation initiation have received increasing attention because of their central importance in diverse biological processes (1). The use of multiple initiation codons in a single mRNA contributes to protein diversity by expressing several protein isoforms from a single transcript. Distinct ORFs defined by alternative TIS codons also could serve as regulatory elements in controlling the translation of the main ORF (10, 11). Although we have some understanding of how ribosomes determine where and when to start initiation, our knowledge is far from complete. GTI-seq provides a comprehensive and high-resolution view of TIS positions across the entire transcriptome. The precise TIS mapping offers insights into the mechanisms of start codon recognition.

Global TIS Mapping at Single-Nucleotide Resolution by GTI-seq. Traditional toeprinting analysis showed heavy ribosome pausing at both the initiation and the termination codons of mRNAs (34,

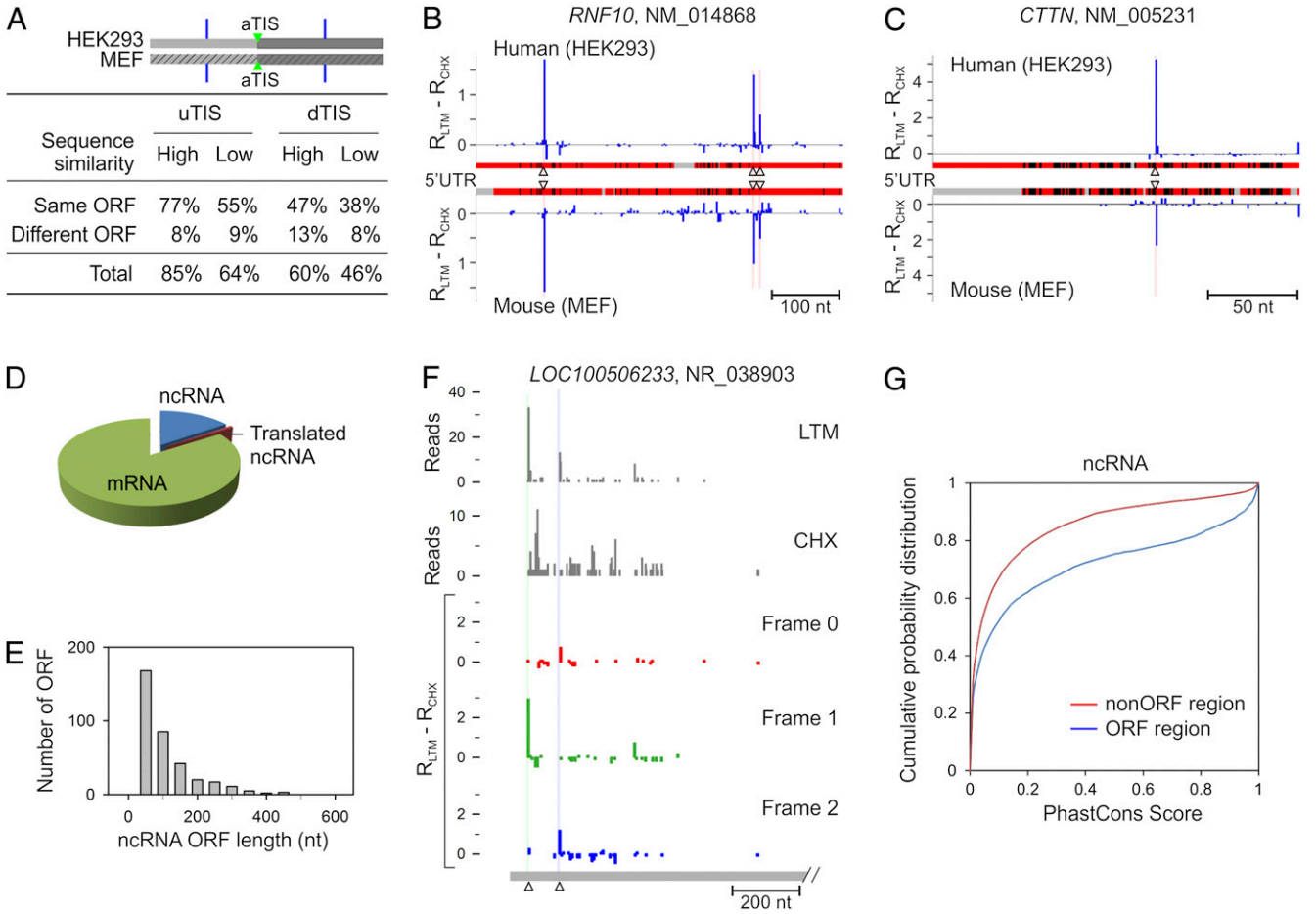


Fig. 6. Cross-species conservation of alternative TIS positions and identification of translated ncRNA. (*A*) Evolutionary conservation of alternative TIS positions identified by GTI-seq in HEK293 and MEF cells. Alternative uTIS and dTIS positions identified on human-mouse ortholog mRNA pairs are each classified into two subsets according to the alignment score of relevant sequences (5′ UTR for uTIS and CDS for dTIS). Each subset is divided further based on types of alternative ORFs. Percentage values are presented in the table. (*B*) Conservation of uTIS positions on the *RNF10* transcript with high 5′ UTR sequence similarity between HEK293 and MEF cells. Red regions indicate matched sequences, black regions indicate mismatched sequences, and gray regions indicate sequence gaps. Identified uTIS positions are indicated by triangles. (*C*) Conservation of uTIS positions on the *CTTN* transcript with low sequence similarity of 5′ UTR between HEK293 and MEF cells. (*D*) Pie chart showing the relative percentage of mRNA, ncRNA and translated ncRNA identified by GTI-seq. (*E*) Histogram showing the overall length distribution of ORFs identified in ncRNAs. (*F*) Identification of multiple TIS positions on the ncRNA *LOC100506233*. (*G*) Evolutionary conservation of the ORF region on ncRNAs identified by GTI-seq. PhastCons scores are retrieved from the primate genome sequence alignment.

35). Consistently, deep sequencing-based ribosome profiling also revealed higher RPF density at both the start and the stop codons (14, 15). Although this feature enables approximate determination of decoded mRNA regions, it does not allow unambiguous identification of TIS positions, especially when multiple initiators are used. Translation inhibitors acting specifically on the first round of peptide bond formation allow the run-off of elongating ribosomes, thereby specifically halting ribosomes at the initiation codon. Indeed, harringtonine treatment caused a profound accumulation of RPFs in the beginning of CDS (16). A caveat regarding the use of harringtonine is that this drug binds to free 60S subunits, and the inhibitory mechanism is unclear. In particular, it is not known whether harringtonine completely blocks the initiation step. We observed that a significant fraction of ribosomes still passed over the start codon in the presence of harringtonine.

The translation inhibitor LTM has several features that contribute to the high resolution of global TIS identification. First, LTM binds to the 80S ribosome already assembled at the initiation codon and permits the formation of the first peptide bond (17). Thus, the LTM-associated RPF more likely represents physiological TIS positions. Second, LTM occupies the empty E-site of initiating ribosomes and thus completely blocks the translocation. This feature allows TIS identification at single-nucleotide resolution. With this precision, different reading frames become unambiguous, thereby revealing different types of ORFs within each transcript. Third, because of their similar structure and the use of the same binding site in the ribosome, LTM and CHX can be applied side by side to achieve simultaneous assessment of both initiation and elongation for the same transcript. With the high signal/noise ratio, GTI-seq offers a direct approach to TIS identification with minimal computational aid. From our analysis, the uncovering of alternative initiators allows us to explore the mechanisms of TIS selection. We also experimentally validated different translational products initiated from alternative start codons, including non-AUG codons. Further confirming the accuracy of GTI-seq, a sizable fraction of alternative start codons identified by GTI-seq exhibited high conservation across species. The evolutionary conservation strongly suggests a physiological significance of alternative translation in gene expression.

Diversity and Complexity of Alternative Start Codons. GTI-seq revealed that the majority of identified TIS positions belong to alternative start codons. The prevailing alternative translation was corroborated by the finding that nearly half the transcripts

contained multiple TIS codons. Although dTIS codons use the conventional AUG as the main initiator, a significant fraction of uTIS codons are non-AUG, with CUG being the most frequent one. In a few well-documented cases, including *FGF2* (36), *VEGF* (37), and *Myc* (38), the CUG triplet was reported to serve as the non-AUG start codon. With the high-resolution TIS map across the entire transcriptome, GTI-seq greatly expanded the list of mRNAs with hidden coding potential not visible by sequence-based *in silico* analysis.

By what mechanisms are alternative start codons selected? GTI-seq revealed several lines of evidence supporting the linear-scanning mechanism for start codon selection. First, the uTIS context, such as the Kozak consensus sequence and the secondary structure, largely influenced the frequency of aTIS initiation. Second, the stringency of an aTIS codon negatively regulated the dTIS efficiency. Third, the leaky potential at the first AUG was inversely correlated with the strength of its sequence context. Because it is less likely that a preinitiation complex will bypass a strong initiator to select a suboptimal one downstream, it is not surprising that most uTIS codons are not canonical, whereas the dTIS codons are mostly conventional AUG. In addition to the leaky scanning mechanism for alternative translation initiation, ribosomes could translate a short uORF and reinitiate at downstream ORFs (2). After termination of a uORF is completed, it was assumed that some translation factors remain associated with the ribosome, facilitating the reinitiation process (39). However, this mechanism is widely considered to be inefficient. From the GTI-seq data set, about half the uORFs were separated from the main ORFs. Compared with transcripts with overlapping uORFs that must rely on leaky scanning to mediate the downstream translation, we observed repressed aTIS initiation in transcripts containing separated uORFs. It is likely that the ribosome reinitiation mechanism plays a more important role in selective translation under stress conditions (27).

Biological Impacts of Alternative Translation Initiation. One expected consequence of alternative translation initiation is an expanded proteome diversity that has not been and could not be predicted by *in silico* analysis of AUG-mediated main ORFs. Indeed, many eukaryotic proteins exhibit a feature of NH_2-terminal heterogeneity presumably caused by alternative translation. Protein isoforms localized in different cellular compartments are typical examples, because most localization signals are within the NH_2-terminal segment (40, 41). Alternative TIS selection also could produce functionally distinct protein isoforms. One well-established example is C/EBP, a family of transcription factors that regulate the expression of tissue-specific genes during differentiation (42).

When an alternative TIS codon is not in the same frame as the aTIS, it is conceivable that the same mRNA will generate unrelated proteins. This production could be particularly important for the function of uORFs, which often are separated from the main ORF and encode short polypeptides. Some of these uORF peptide products control ribosome behavior directly, thereby regulating the translation of the main ORF. For instance, the translation of *S*-adenosylmethionine decarboxylase is subject to regulation by the six-amino acid product of its uORF (43). The alternative translational products also could function as biologically active peptides. A striking example is the discovery of short ORFs in noncoding RNAs of *Drosophila* that produce functional small peptides during development (44). However, both computational prediction and experimental validation of peptide-encoding short ORFs within the genome are challenging. Our study using GTI-seq represents a potential addition to the expanding ORF catalog by including ORFs from ncRNAs.

Perspective. The enormous biological breadth of translational regulation has led to an enhanced appreciation of its complexities. However, current endeavors aiming to understand protein translation have been hindered by technological limitations. Comprehensive cataloging of global TIS and the associated ORFs is just the beginning step in unveiling the role of translational control in gene expression. More focused studies will be needed to decipher the function and regulatory mechanism of novel ORFs individually. A systematic, high-throughput method like GTI-seq offers a top-down approach, in which one can identify a set of candidate genes for intensive study. GTI-seq is readily applicable to broad fields of fundamental biology. For instance, applications of GTI-seq in different tissues will facilitate the elucidation of the tissue-specific translational control. The illustration of altered TIS selection under different growth conditions will set the stage for future investigation of translational reprogramming during organismal development as well as in human diseases.

Materials and Methods

HEK293 or MEF cells were treated with 100 μM CHX, 50 μM LTM, 2 μg/mL harringtonine, or DMSO at 37 °C for 30 min. Cells were lysed in polysome buffer, and cleared lysates were separated by sedimentation through sucrose gradients. Collected polysome fractions were digested with RNase I, and the RPF fragments were size selected and purified by gel extraction. After the construction of the sequencing library from these fragments, deep sequencing was performed using Illumina HiSEQ. The trimmed RPF reads with final lengths of 26–29 nt were aligned to the RefSeq transcript sequences by Bowtie-0.12.7, allowing one mismatch. A TIS position on an individual transcript was called if the normalized density of LTM reads at the every nucleotide position minus the density of CHX reads at that position was well above the background. In the analysis of noncoding RNA, only reads unique to single ncRNA were used. To validate the identified TIS codons experimentally, specific genes encompassing both the 5′ UTR and the CDS were amplified by RT-PCR from total cellular RNAs extracted from HEK293 cells. The resultant cDNAs were cloned into pcDNA3.1 containing a c-*myc* tag at the COOH terminus. After transfection into HEK293 cells, whole-cell lysates were used for immunoblotting using anti-*myc* antibody. Full methods are available in *SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank S.-B.Q. laboratory members for helpful discussions during the course of this study; Drs. Chaolin Zhang (Rockefeller University) and Adam Siepel (Cornell University) for critical reading of the manuscript; and the Cornell University Life Sciences Core Laboratory Center for performing deep sequencing. This work was supported by National Institutes of Health (NIH) Grants CA106150 (to B.S.) and 1 DP2 OD006449-01, Ellison Medical Foundation Grant AG-NS-0605-09, and Department of Defense Exploration-Hypothesis Development Award W81XWH-11-1-02368 (to S.-B.Q.).

1. Sonenberg N, Hinnebusch AG (2009) Regulation of translation initiation in eukaryotes: Mechanisms and biological targets. *Cell* 136:731–745.
2. Jackson RJ, Hellen CU, Pestova TV (2010) The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat Rev Mol Cell Biol* 11:113–127.
3. Gray NK, Wickens M (1998) Control of translation initiation in animals. *Annu Rev Cell Dev Biol* 14:399–458.
4. Kozak M (2002) Pushing the limits of the scanning mechanism for initiation of translation. *Gene* 299:1–34.
5. Kozak M (1991) Structural features in eukaryotic mRNAs that modulate the initiation of translation. *J Biol Chem* 266:19867–19870.
6. Kozak M (1990) Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes. *Proc Natl Acad Sci USA* 87:8301–8305.
7. Maag D, Fekete CA, Gryczynski Z, Lorsch JR (2005) A conformational change in the eukaryotic translation preinitiation complex and release of eIF1 signal recognition of the start codon. *Mol Cell* 17:265–275.
8. Martin-Marcos P, Cheung YN, Hinnebusch AG (2011) Functional elements in initiation factors 1, 1A, and 2β discriminate against poor AUG context and non-AUG start codons. *Mol Cell Biol* 31:4814–4831.
9. Iacono M, Mignone F, Pesole G (2005) uAUG and uORFs in human and rodent 5′untranslated mRNAs. *Gene* 349:97–105.
10. Morris DR, Geballe AP (2000) Upstream open reading frames as regulators of mRNA translation. *Mol Cell Biol* 20:8635–8642.
11. Calvo SE, Pagliarini DJ, Mootha VK (2009) Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc Natl Acad Sci USA* 106:7507–7512.

12. Resch AM, Ogurtsov AY, Rogozin IB, Shabalina SA, Koonin EV (2009) Evolution of alternative and constitutive regions of mammalian 5′UTRs. *BMC Genomics* 10:162.
13. Touriol C, et al. (2003) Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons. *Biol Cell* 95:169–178.
14. 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.
15. Guo H, Ingolia NT, Weissman JS, Bartel DP (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* 466:835–840.
16. 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.
17. Schneider-Poetsch T, et al. (2010) Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin. *Nat Chem Biol* 6:209–217.
18. Klinge S, Voigts-Hoffmann F, Leibundgut M, Arpagaus S, Ban N (2011) Crystal structure of the eukaryotic 60S ribosomal subunit in complex with initiation factor 6. *Science* 334:941–948.
19. 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.
20. Sugawara K, et al. (1992) Lactimidomycin, a new glutarimide group antibiotic. Production, isolation, structure and biological activity. *J Antibiot (Tokyo)* 45: 1433–1441.
21. Steitz TA (2008) A structural understanding of the dynamic ribosome machine. *Nat Rev Mol Cell Biol* 9:242–253.
22. Fresno M, Jiménez A, Vázquez D (1977) Inhibition of translation in eukaryotic systems by harringtonine. *Eur J Biochem* 72:323–330.
23. Shalak V, Kaminska M, Mirande M (2009) Translation initiation from two in-frame AUGs generates mitochondrial and cytoplasmic forms of the p43 component of the multisynthetase complex. *Biochemistry* 48:9959–9968.
24. Kochetov AV (2008) Alternative translation start sites and hidden coding potential of eukaryotic mRNAs. *Bioessays* 30:683–691.
25. Spriggs KA, Bushell M, Willis AE (2010) Translational regulation of gene expression during conditions of cell stress. *Mol Cell* 40:228–237.
26. Harding HP, Calfon M, Urano F, Novoa I, Ron D (2002) Transcriptional and translational control in the Mammalian unfolded protein response. *Annu Rev Cell Dev Biol* 18:575–599.
27. Vattem KM, Wek RC (2004) Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. *Proc Natl Acad Sci USA* 101:11269–11274.
28. Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) Coding-sequence determinants of gene expression in Escherichia coli. *Science* 324:255–258.
29. Kochetov AV, et al. (2007) AUG_hairpin: Prediction of a downstream secondary structure influencing the recognition of a translation start site. *BMC Bioinformatics* 8:318.
30. Kertesz M, et al. (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature* 467:103–107.
31. Mattick JS (2005) The functional genomics of noncoding RNA. *Science* 309:1527–1528.
32. Pauli A, Rinn JL, Schier AF (2011) Non-coding RNAs as regulators of embryogenesis. *Nat Rev Genet* 12:136–149.
33. Siepel A, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15:1034–1050.
34. Wolin SL, Walter P (1989) Signal recognition particle mediates a transient elongation arrest of preprolactin in reticulocyte lysate. *J Cell Biol* 109:2617–2622.
35. Sachs MS, et al. (2002) Toeprint analysis of the positioning of translation apparatus components at initiation and termination codons of fungal mRNAs. *Methods* 26: 105–114.
36. Vagner S, et al. (1996) Translation of CUG- but not AUG-initiated forms of human fibroblast growth factor 2 is activated in transformed and stressed cells. *J Cell Biol* 135:1391–1402.
37. Meiron M, Anunu R, Scheinman EJ, Hashmueli S, Levi BZ (2001) New isoforms of VEGF are translated from alternative initiation CUG codons located in its 5′UTR. *Biochem Biophys Res Commun* 282:1053–1060.
38. Hann SR, King MW, Bentley DL, Anderson CW, Eisenman RN (1988) A non-AUG translational initiation in c-myc exon 1 generates an N-terminally distinct protein whose synthesis is disrupted in Burkitt's lymphomas. *Cell* 52:185–195.
39. Pöyry TA, Kaminski A, Jackson RJ (2004) What determines whether mammalian ribosomes resume scanning after translation of a short upstream open reading frame? *Genes Dev* 18:62–75.
40. Chang KJ, Wang CC (2004) Translation initiation from a naturally occurring non-AUG codon in Saccharomyces cerevisiae. *J Biol Chem* 279:13778–13785.
41. Porras P, Padilla CA, Krayl M, Voos W, Bárcena JA (2006) One single in-frame AUG codon is responsible for a diversity of subcellular localizations of glutaredoxin 2 in Saccharomyces cerevisiae. *J Biol Chem* 281:16551–16562.
42. Descombes P, Schibler U (1991) A liver-enriched transcriptional activator protein, LAP, and a transcriptional inhibitory protein, LIP, are translated from the same mRNA. *Cell* 67:569–579.
43. Hill JR, Morris DR (1993) Cell-specific translational regulation of S-adenosylmethionine decarboxylase mRNA. Dependence on translation and coding capacity of the cis-acting upstream open reading frame. *J Biol Chem* 268:726–731.
44. Kondo T, et al. (2010) Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis. *Science* 329:336–339.

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.

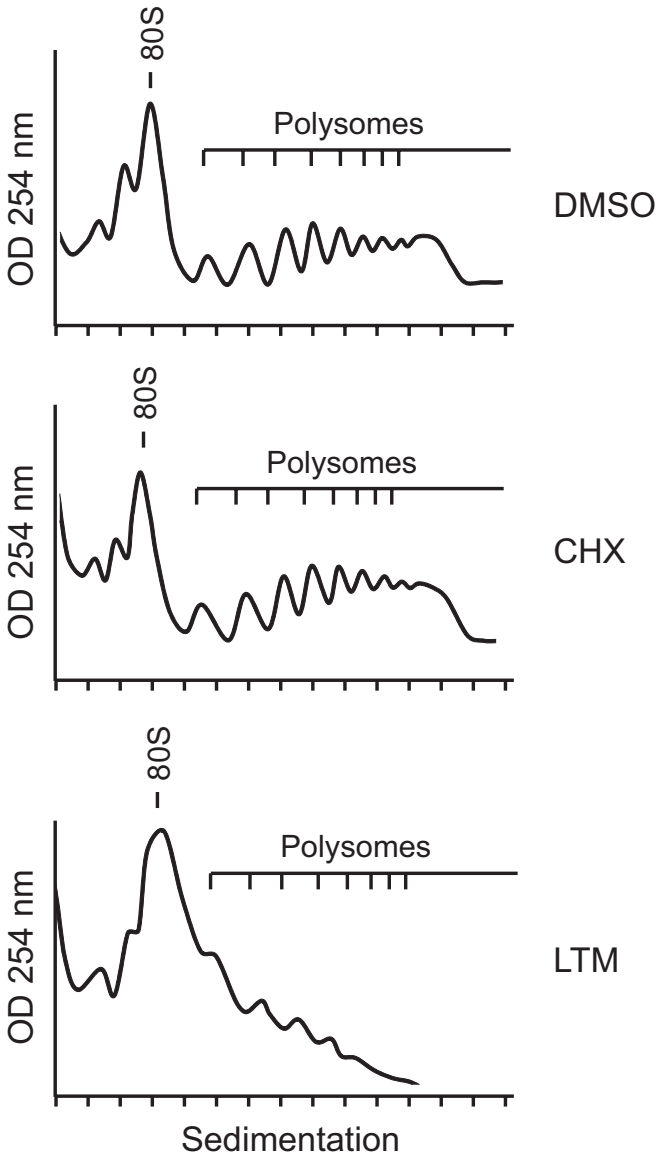


Fig. S1. Polysome profile analysis in cells treated with ribosome exit-site translation inhibitors. HEK293 cells were pretreated with equal volume of DMSO, 100 μM CHX, or 50 μM LTM for 30 min followed by sucrose gradient sedimentation. Both 80S monosome and polysome peaks are indicated.

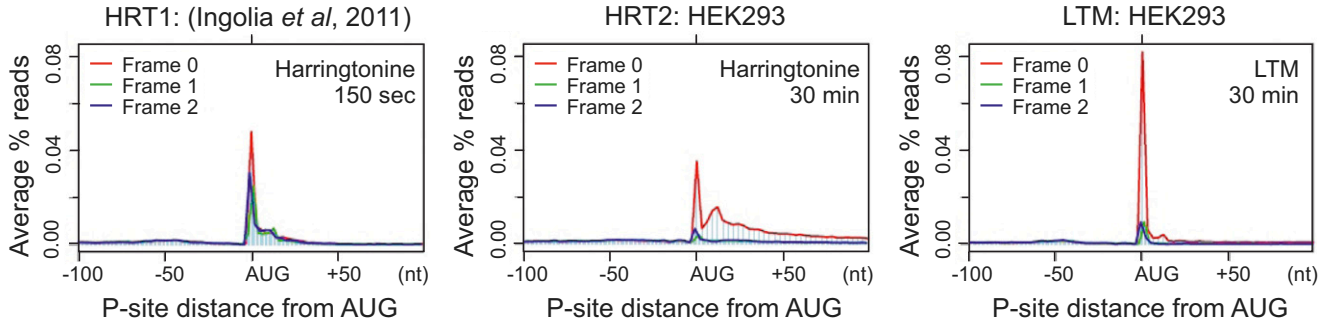


Fig. S2. Metagene analysis of RPFs obtained using different approaches. RPF reads reported by Ingolia et al. (1) using harringtonine in mouse embryonic stem cells were replotted after peptidyl (P)-site adjustment based on the original report (HRT1, *Left*). RPF reads obtained from HEK293 cells treated 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.

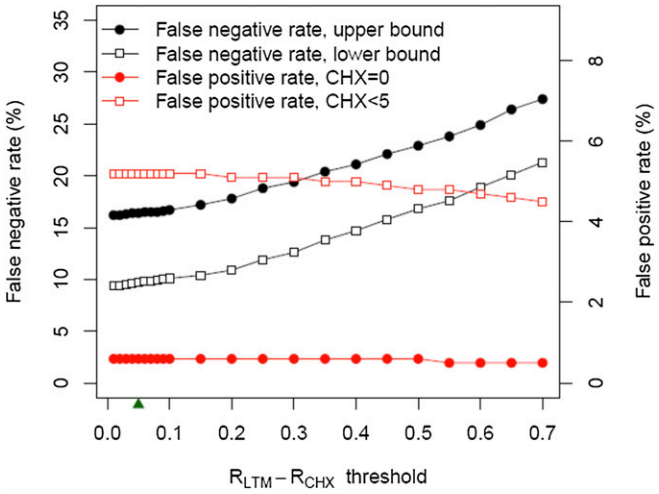


Fig. S3. False-positive and false-negative rates at various $R_{LTM} - R_{CHX}$ thresholds. The false-negative rate is computed as the percentage of undetected aTIS among the top 10% translated aTIS codons based on CHX reads within five codons downstream of the aTIS. The lower and upper bounds of false-negative 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.

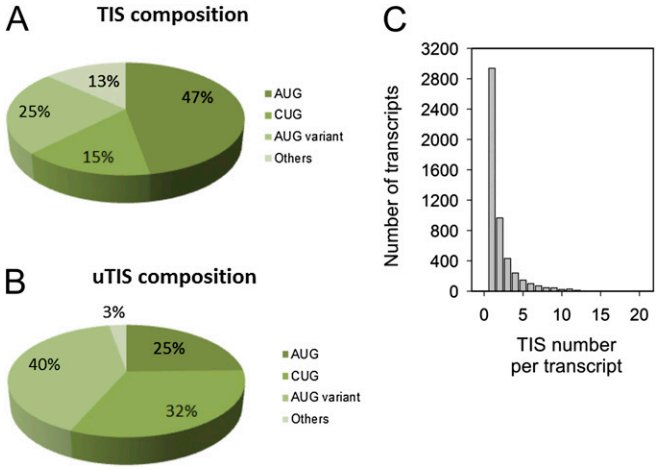
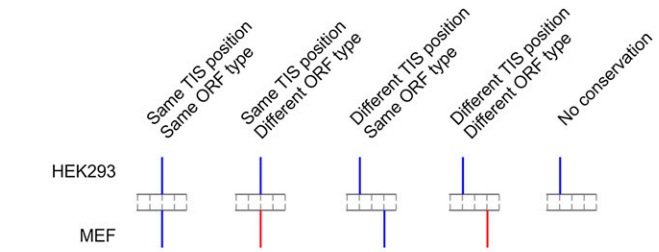


Fig. S4. Global TIS identification in MEF cells. (*A*) Codon composition of all TIS codons identified by global translation initiation sequencing (GTI-seq) in MEF cells. (*B*) Codon composition of uTIS codons identified by GTI-seq in MEF cells. (*C*) Histogram showing the overall distribution of the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

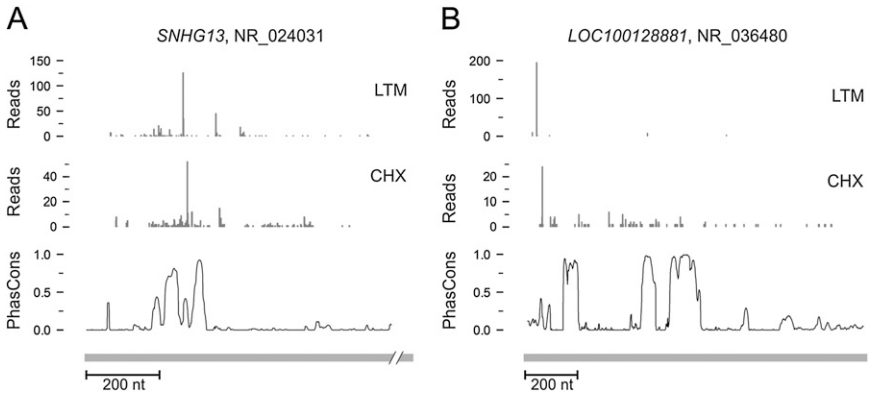


Fig. S6. ORF conservation in ncRNAs. (*A*) Translation in ncRNA *SNHG13* is illustrated by LTM- and CHX-associated RPF reads. PhastCons scores retrieved from the primate genome sequence alignment are plotted also (*B*) Translation in ncRNA *LOC100128881* is illustrated by LTM- and CHX-associated RPF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