Cell

Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes

Nicholas T. Ingolia,[1,3,*] Liana F. Lareau,[2] and Jonathan S. Weissman[1]
[1]Howard Hughes Medical Institute, Department of Cellular and Molecular Pharmacology, University of California, San Francisco and California Institute for Quantitative Biosciences, San Francisco, CA 94158, USA
[2]Department of Biochemistry, Stanford University, Stanford, CA 94305, USA
[3]Present address: Department of Embryology, Carnegie Institution for Science, Baltimore, MD 21218, USA
*Correspondence: ingolia@ciwemb.edu

DOI 10.1016/j.cell.2011.10.002

SUMMARY

The ability to sequence genomes has far outstripped approaches for deciphering the information they encode. Here we present a suite of techniques, based on ribosome profiling (the deep sequencing of ribosome-protected mRNA fragments), to provide genome-wide maps of protein synthesis as well as a pulse-chase strategy for determining rates of translation elongation. We exploit the propensity of harringtonine to cause ribosomes to accumulate at sites of translation initiation together with a machine learning algorithm to define protein products systematically. Analysis of translation in mouse embryonic stem cells reveals thousands of strong pause sites and unannotated translation products. These include amino-terminal extensions and truncations and upstream open reading frames with regulatory potential, initiated at both AUG and non-AUG codons, whose translation changes after differentiation. We also define a class of short, polycistronic ribosome-associated coding RNAs (sprcRNAs) that encode small proteins. Our studies reveal an unanticipated complexity to mammalian proteomes.

INTRODUCTION

In the 10 years since the publication of draft human genomes (Lander et al., 2001; Venter et al., 2001), extraordinary advances in DNA sequencing technology (Bentley et al., 2008) have made it possible to obtain comprehensive genomic information rapidly and at low cost. Decoding the information contained in these genomes represents a central challenge for the biological community. Protein-coding regions have been defined according to simple rules about the nature of translation—for example, that open reading frames (ORFs) have a minimum length, have biased codon usage, and start at the first AUG in a transcript (Brent, 2005). Yet there are many exceptions to these rules, including internal ribosome entry sites, initiation at non-AUG codons, leaky scanning, translational reinitiation, and translational frameshifts (Atkins and Gesteland, 2010). Additionally, an abundant class of large intergenic noncoding RNAs (lincRNAs) that do not contain canonical ORFs has been recently been described (Guttman et al., 2009, 2010). Many of these lincRNA transcripts are likely to be functional RNAs, but there are well-documented cases of biologically important short coding regions. For example, the *Drosophila tarsal-less/polished rice* gene was originally thought to be a lincRNA (Tupy et al., 2005) but actually encodes a series of short peptides that modulate the activity of the shavenbaby transcription factor (Kondo et al., 2010). The question of which of the potential lincRNAs are actually translated remains largely unaddressed.

We also know that the rate of translation is not constant across a message, and translation pauses can regulate synthesis (Darnell et al., 2011; Morris and Geballe, 2000), folding (Kimchi-Sarfaty et al., 2007; Zhang et al., 2009), and localization of a protein (Mariappan et al., 2010) or mRNA (Yanagitani et al., 2011). These pauses can results from codon usage (Irwin et al., 1995), mRNA structure (Namy et al., 2006), or peptide sequence (Nakatogawa and Ito, 2002; Tenson and Ehrenberg, 2002), but little information exists on how generally they occur, let alone their functional impact.

Recently, we described a strategy, termed ribosome profiling, based on deep sequencing of ribosome-protected mRNA fragments, that makes it possible to monitor translation with a depth, speed, and accuracy that rival existing approaches for following mRNA levels (Guo et al., 2010; Ingolia et al., 2009). By revealing the precise locations of ribosomes on each mRNA, ribosome profiling also has the potential to identify protein-coding regions. However, initiation from multiple sites within a single transcript makes it challenging to define all ORFs, especially in complex transcriptomes. Additionally, ribosome profiling provides a snapshot of ribosome positions but does not report directly on the kinetics of translational elongation or distinguish stalled ribosomes from those engaged in active elongation.

Here we describe a simplified, robust protocol for ribosome profiling in mammalian systems. We have used this technique to determine the kinetics of translation by following run-off elongation after stalling new initiation using the drug harringtonine (Fresno et al., 1977; Huang, 1975; Robert et al., 2009; Tscherne and Pestka, 1975). We further employ harringtonine, which causes ribosomes to accumulate precisely at initiation codons, together with a machine learning algorithm, to define the sites of translation initiation genome-wide. Application of our approach to mouse embryonic stem cells (mESCs) reveals a wide range of unannotated or modified ORFs, including highly translated short ORFs in the majority of annotated lincRNAs. We now classify these atypical protein-coding transcripts as short, polycistronic ribosome-associated RNAs (sprcRNAs). Additionally, we identify over a thousand strong translational pauses that could act as key regulatory sites. Our approach is readily applicable to other cells and organisms and as such provides a general scheme for decoding complex genomes, monitoring rates of protein production, and exploring the molecular mechanisms used to regulate translation.

RESULTS

A Simplified Mammalian Ribosome-Profiling Assay

We first describe a simplified ribosome-profiling strategy suitable for the analysis of mammalian cells. In general terms, the assay involves three distinct steps, each of which has been refined: (1) generation of cell extracts in which ribosomes have been faithfully halted along the mRNA they are translating in vivo; (2) nuclease digestion of RNAs that are not protected by the ribosome followed by recovery of the ribosome-protected mRNA fragments; (3) quantitative conversion of the protected RNA fragments into a DNA library that can be analyzed by deep sequencing (Ingolia, 2010; Ingolia et al., 2009; Lau et al., 2001; Pfeffer et al., 2005). After nuclease treatment, we purified ribosomes and the associated mRNA footprints by ultracentrifugation through a sucrose cushion rather than by sucrose density gradient fractionation, which is a more specialized technique. Protected mRNA fragments from single ribosomes were purified by PAGE, as fragments that derive from other ribosomal complexes are longer—tightly packed ribosome pairs protect 58–62 nt of mRNA (Wolin and Walter, 1988), and 48S preinitiation complexes are reported to protect 50 nt or 70 nt under different conditions (Lazarowitz and Robertson, 1977; Pisarev et al., 2008; Ule et al., 2003; Chi et al., 2009; Lunde et al., 2007). We generated libraries from these purified fragments using our previous published protocol (Ingolia, 2010; Ingolia et al., 2009), modified to use RNA ligation to attach a linker to the 3′ end of the protected RNA fragment (Lau et al., 2001; Pfeffer et al., 2005). Additionally, we used subtractive hybridization to substantially deplete the majority of contaminating ribosomal RNA fragments.

We explored the effects of stabilizing ribosome-mRNA interactions with elongation inhibitors before cell lysis. We compared cycloheximide (Schneider-Poetsch et al., 2010) and emetine pretreatment to a ‘‘no drug’’ approach in which unperturbed cells were lysed in a buffer that should not support continued elongation. The density of ribosome footprints on each coding sequence (CDS), which measures the translation of the gene, agreed well across the three approaches (cycloheximide versus no drug, standard deviation (SD) of $\log_2$ ratio 0.20, corresponding to a typical 15% inter-replicate difference; cycloheximide versus emetine, SD $\log_2$ ratio 0.40; emetine versus no drug, SD $\log_2$ ratio 0.41) (Figure 1A). We concluded that brief treatment of cells with elongation inhibitors did not significantly change which transcripts were associated with ribosomes and did not distort translation measurements made by ribosome profiling. Thus, pretreatment can be chosen based on experimental constraints. For example, elongation inhibitors would preserve the cellular state of translation during manipulations such as fluorescence-activated cell sorting (FACS), whereas flash-freezing and cryogenic lysis would enable the analysis of tissues where infusion of translation inhibitors is challenging.

Nonetheless, elongation inhibitors do alter the pattern of ribosome footprints along transcripts. Footprints derived from emetine-treated cells are slightly longer than those from untreated or cycloheximide-treated cells (Figure 1B and Figures S1A and S1B available online), suggesting that emetine stabilizes a different ribosome conformation that protects more mRNA. Furthermore, a metagene analysis, in which many gene profiles are aligned and then averaged, revealed global differences in ribosome density at the beginning and ends of ORFs. The excess of ribosomes at the initiation site and extending over the first 5 to 10 codons is essentially absent from untreated cells (Figure 1C). Such an excess would result from the inhibition of translation elongation in the presence of continuing initiation. Beyond the initial 5 to 10 codon window, we saw no global variation in ribosome density along CDSes in any sample. An earlier analysis had suggested that the excess ribosomes extending over ~100 codons at the beginning of *Saccharomyces cerevisiae* ORFs reflected a broadly conserved ‘‘ramping’’ strategy that minimized ribosome stacking and collisions later in the messages (Tuller et al., 2010). Although it is possible that such a ramping effect occurs in *S. cerevisiae*, it does not appear to occur mammalian cells.

Drug pretreatment also eliminates the excess of ribosomes seen at the stop codon in untreated cells (Figure 1D). The accumulation is still seen when cells are lysed in the presence of a nonhydrolyzable GTP analog, suggesting that it does not result from continued elongation in the lysate (N. Stern-Ginossar and J.S.W., unpublished data). Interestingly, we saw longer footprints at stop codons (Figures 1B and S1B), suggesting that the accumulating ribosomes are in a different conformation, as has been seen during termination in vitro (Alkalaeva et al., 2006). In summary, although drug pretreatment does not distort measurements of the overall level of translation of a given message (Figure 1A), caution should be used in interpreting position-specific information.

We characterized translation in an mESC line (E14 mESCs), with matched ribosome-profiling and mRNA-seq data. We used ribosome footprint density within a CDS as a measure of protein synthesis and determined levels of gene expression genome-wide (Figure S1C and Tables S1A and S1B). We also compared protein synthesis with mRNA abundance and showed that there was a broad distribution encompassing over a 10-fold range in the amount of protein produced per transcript (Figure S1D and Tables S1C and S1D). This distribution is asymmetric, suggesting a maximal rate of protein production from an mRNA and substantial dynamic range for decreased translational efficiency. Our data are consistent

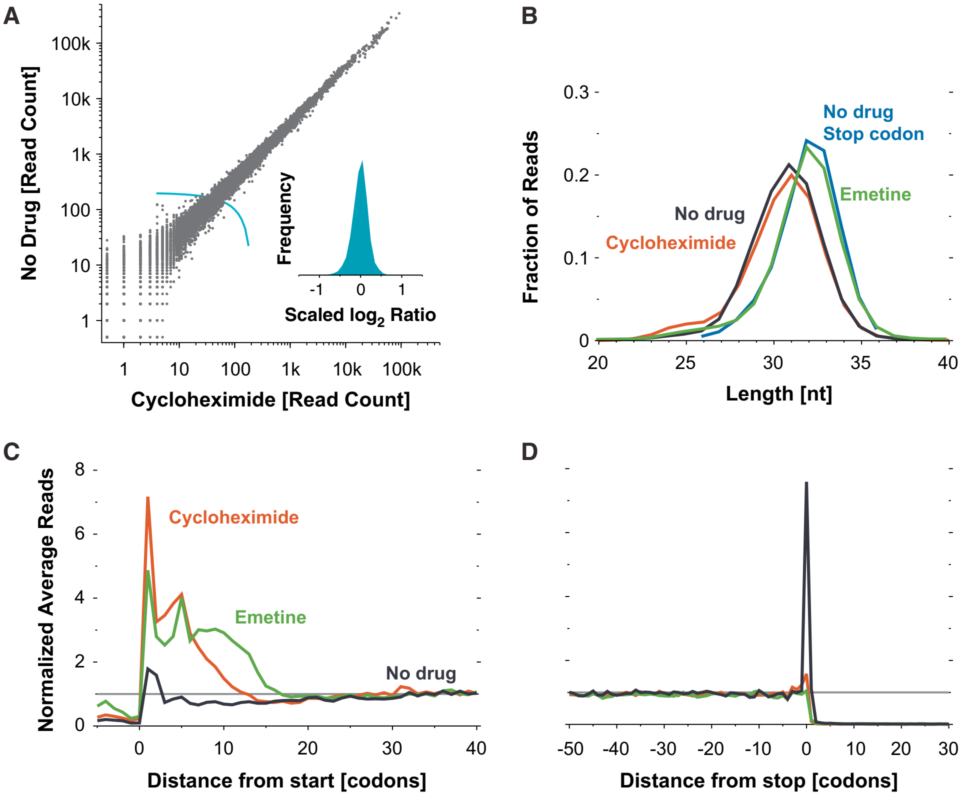

Figure 1. Ribosome Profiling in mESCs

(A) Effect of elongation inhibitors on ribosome density. The number of ribosome footprint reads that align to the body of each CDS (Experimental Procedures) is plotted for cells that were either untreated or pretreated with cycloheximide (Spearman r = 0.99). The inset shows a histogram of $\log_2$ ratios for genes with at least 200 total reads (the threshold shown by the light blue line) normalized by the median ratio (N = 10045, SD = 0.20, corresponding to 15% difference in measurements).

(B) Ribosome-protected fragment lengths. Plotted is the length distribution of ribosome footprints over the body of messages prepared from cells treated as indicated, as well as for footprints centered on the stop codon for the untreated cells.

(C) Metagene analysis of translation initiation. Average ribosome read density profiles over 4,994 well-expressed genes (Table S1), aligned at their start codon, are shown for untreated and drug-treated samples.

(D) Metagene analysis of translation termination. As in (C), but alignment was from stop codons.

See also Figure S1 and Table S1.

with recent work that indirectly infers translation levels from absolute mRNA and protein abundance measurements (Schwanhäusser et al., 2011). Notably, they found that translation was the single largest contributor to protein abundance, highlighting the value of direct measurements of protein synthesis.

Widespread Presence of Strong Ribosomal Pauses

The density of ribosome footprint reads varies substantially at different codons within an individual message (Figure 2A). The footprint count on a codon should be proportional to the average ribosome dwell time there, so this density variation represents differences in the speed of the ribosome. Position-specific variability is pervasive in both yeast and mESC ribosome-profiling data (Ingolia et al., 2009), but in mammalian translation, we find more pronounced pauses where ribosome density is 25-fold or greater than the median density observed across the body of the gene. Based on a typical elongation rate of ~6 codons per second (see below), the pauses we see last for several seconds (Figure 2A), which is enough time for the paused ribosome-nascent chain (RNC) complex to bind cotranslational chaperones.

We find thousands of pauses in the body of genes (1500 pauses in 1100 genes found in a set of 4994 well-expressed genes; Tables S2A and S2B) and at termination codons (420 pauses; Table S2C). Interestingly, we see no evidence that pausing causes secondary ribosome accumulation ~10 codons upstream, where a following ribosome would collide with the stalled one (Wolin and Walter, 1988), nor a depletion of ribosomes within the 10 codon "shadow" resulting from paused ribosomes (Figure 2B). The lack of packed ribosomes at pause sites suggests that ribosome density is typically too low to cause frequent encounters between upstream elongating ribosomes and a transiently stalled downstream ribosome (Arava et al., 2003). Alternately, such a collision might relieve ribosomal stalling, allowing for the continual presence of a ribosome at a pause site while minimizing ribosome sequestration.

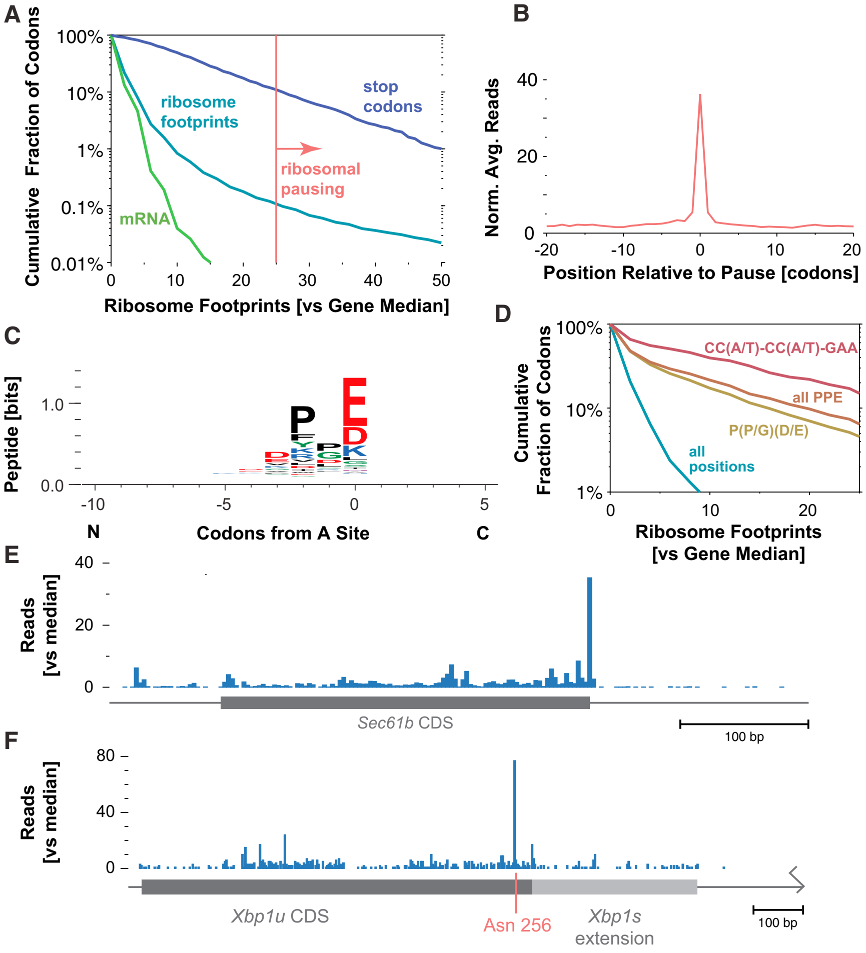

Figure 2. Analysis of Translational Stall Sites

(A) Distribution of per-codon ribosome footprint counts. The cumulative distribution of footprint counts at each codon, relative to the median density across the gene, is plotted, and the 25 × median threshold used to identify ribosomal stall sites is indicated. The distribution of density at stop codons, which are excluded from the overall distribution, is shown as well, along with the read densities in randomly fragmented mRNA, which controls for library generation.

(B) Metagene analysis of translational stalling. Ribosome footprint densities were averaged after aligning gene density profiles at internal translational stall positions (Table S2B).

(C) Peptide motif associated with internal translational stalling.

(D) Ribosome footprints over peptide motif enriched in stall sites. The cumulative distribution of relative ribosome footprint counts for the all Pro-Pro-Glu sites and for those encoded by CC(A/T)-CC(A/T)-GAA are shown along with the more lenient Pro-(Pro/Gly)-(Asp/Glu) sites and the overall data from (A).

(E) Ribosome footprint profile on the *Sec61b* transcript (median 22.5 footprints per codon).

(F) Ribosome footprint profile on the *Xbp1* transcript (median 1.0 footprint per codon). *Xbp1* undergoes a nonconventional splicing event (Calfon et al., 2002). The unspliced (*Xbp1u)* CDS is indicated, along with the site of translational stalling at Asn256 and the extended CDS in the spliced (*Xbp1s*) message.

See also Table S2.

The absence of downstream depletion also argues that the majority of ribosomes continue elongation following these pause sites.

Analysis of the sequence around the pause sites reveals a consensus peptide motif (Figure 2C). There is strong enrichment of glutamate or asparate in the A site at strong pauses, preceded by a proline or glycine and then another proline, with an additional bias toward the GAA glutamate codon and CC(A/T) proline codons. Importantly, we see no enrichment for residues or codons downstream of the A site, which are not yet being decoded. We also see no evidence that the pause sites are enriched for rare codons. Sites that match the full three-residue consensus have dramatically reduced elongation rates overall (Figure 2D). Translation in *E. coli* is stalled by similar peptide motifs with a terminal Pro-Pro peptide, in some cases with an Asp codon in the A site (Tanner et al., 2009). Our findings suggest that transfer RNA (tRNA) identity and nascent peptide sequence can influence the kinetics of elongation, whereas even for rare codons, tRNA recruitment is not rate limiting.

Our analysis also provides insights into the limited number of previously documented translational pauses. A recent study observed slow termination of two tail-anchored (TA) proteins (*Sec61b* and *Vamp*) during in vitro translation (Mariappan et al., 2010). Pausing at the termination codon of TA proteins has been proposed to provide time for the recruitment of the insertion machinery before the release of the C-terminal transmembrane domain from the ribosome exit channel. Our data confirmed termination pausing during the *Sec61b* and *Vamp* translation in vivo (Figure 2E), but we found no evidence for this phenomenon in the majority of other TA proteins (3/32 have pauses), nor was it restricted to TA proteins (stop codon ribosome density does not differ significantly, Kruskal-Wallis p ~0.25). Instead, pausing at termination codons is a common feature of translation.

A second prominent example of a translation pause follows a hydrophobic sequence in the *Xbp1* transcription factor (Yanagitani et al., 2011). This hydrophobic domain interacts with the endoplasmic reticulum (ER) membrane and recruits the *Xbp1* message RNC complex (Yanagitani et al., 2009). Ribosome pausing facilitates this cotranslational localization by delaying the dissociation of the RNC. We confirmed the presence of this pause and identified its precise position as residue Asn256, which is the last codon required for translational arrest (Yanagitani et al., 2011) (Figure 2F). The biological roles of the pauses we identify remain to be established, but many mRNAs are localized to specific subcellular regions (Martin and Ephrussi, 2009), including a number of mRNAs found on the ER surface that, like *Xbp1*, do not enter the secretory pathway (Kraut-Cohen

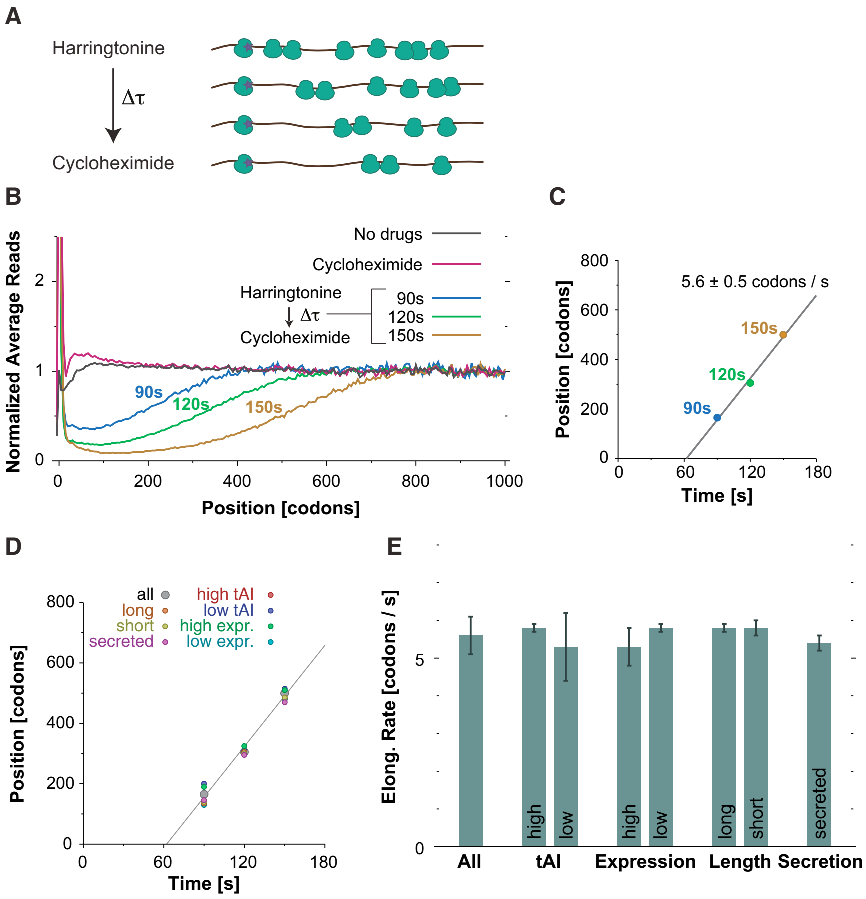

Figure 3. A Pulse-Chase Strategy for Measuring Translation Elongation Rates

(A) Schematic of the in vivo run-off elongation experiment.

(B) Metagene analysis of run-off elongation. Ribosome read density was averaged across 5 codon windows for samples prepared with the indicated drug treatments.

(C) Rate of ribosome depletion. The codon position of 50% ribosome depletion is plotted as a function of harringtonine treatment time. Linear fit is x(t) = ax + b, a = 5.6 ± 0.5 codons/s, b = −347 ± 65 codons, root-mean-square deviation (rmsd) 22.5.

(D) Ribosome depletion on subsets of genes. Data from (C) are plotted, along with comparable measurements made from the indicated gene subsets.

(E) Elongation rates on subsets of genes. Elongation rates, inferred from linear fit as described in (C), are plotted along with the standard error of the regressed coefficient.

and Gerst, 2010), and so the mechanism described for *Xbp1* localization may be more general.

Monitoring Kinetics of Translation

Our knowledge of the kinetics of protein synthesis in vivo has been based on a limited number of specific messages (Boström et al., 1986). We reasoned that we could monitor the kinetics of in vivo translation directly by tracing run-off elongation using ribosome profiling. We first stopped new translation using harringtonine, which effectively blocks initiation by inhibiting elongation during the first rounds of peptide bond formation following subunit joining (Fresno et al., 1977; Robert et al., 2009). We then allowed a short time for run-off elongation before adding cycloheximide to halt translation by all active ribosomes. We varied the time allowed for run-off elongation to generate a series of snapshots that could be assembled into a moving picture of translation in vivo (Figure 3A). Metagene analyses revealed a progressive depletion of ribosomes from the 5′ to the 3′ ends of the messages after harringtonine treatment. Following a delay of ∼60 s, which presumably reflects the time required for engagement of harringtonine, ribosomes progress from the 5′ ends of transcripts at a rate of 5.6 amino acids per second (Figures 3B and 3C), which is consistent with values from previous single-gene measurements (Boström et al., 1986).

The rate of translation is remarkably consistent between different classes of messages (Figures 3D and 3E). The kinetics of elongation are independent of length and protein abundance and are the same in secreted proteins, whose translation occurs on the ER surface. Translation speed is also independent of codon usage, which is consistent with the absence of pauses at rare codons. This is surprising as it is often assumed that codons corresponding to low-abundance tRNAs are decoded more slowly than those read by abundant tRNAs. Although this may be the case for specific examples, we find no evidence for a large effect on the overall rate of elongation. An important practical implication for the universality of the average rate of elongation is that ribosome footprint density provides a reliable measure of protein synthesis independent of the particular gene being translated.

Defining Translation Start Sites

We found that harringtonine treatment also leads to a profound accumulation of ribosomes at the sites of translation initiation (Figures 4A and 4B). This effect likely occurs because harringtonine binds to free 60S subunits but not those that are joined into an 80S ribosome. Thus, elongating ribosomes are immune to harringtonine, whereas a 60S subunit bound by harringtonine will form an 80S at a start site that does not move forward (Fresno et al., 1977; Robert et al., 2009). We reasoned that this accumulation of ribosomes could serve as a basis for objectively detecting translation initiation. Accordingly, we used a support vector machine (SVM)-based machine learning strategy (Joachims, 1999; Noble, 2006) to comprehensively identify initiation sites from harringtonine-treated ribosome footprint profiles, using a “vector” of footprint counts around a candidate translation start site. The SVM model was trained

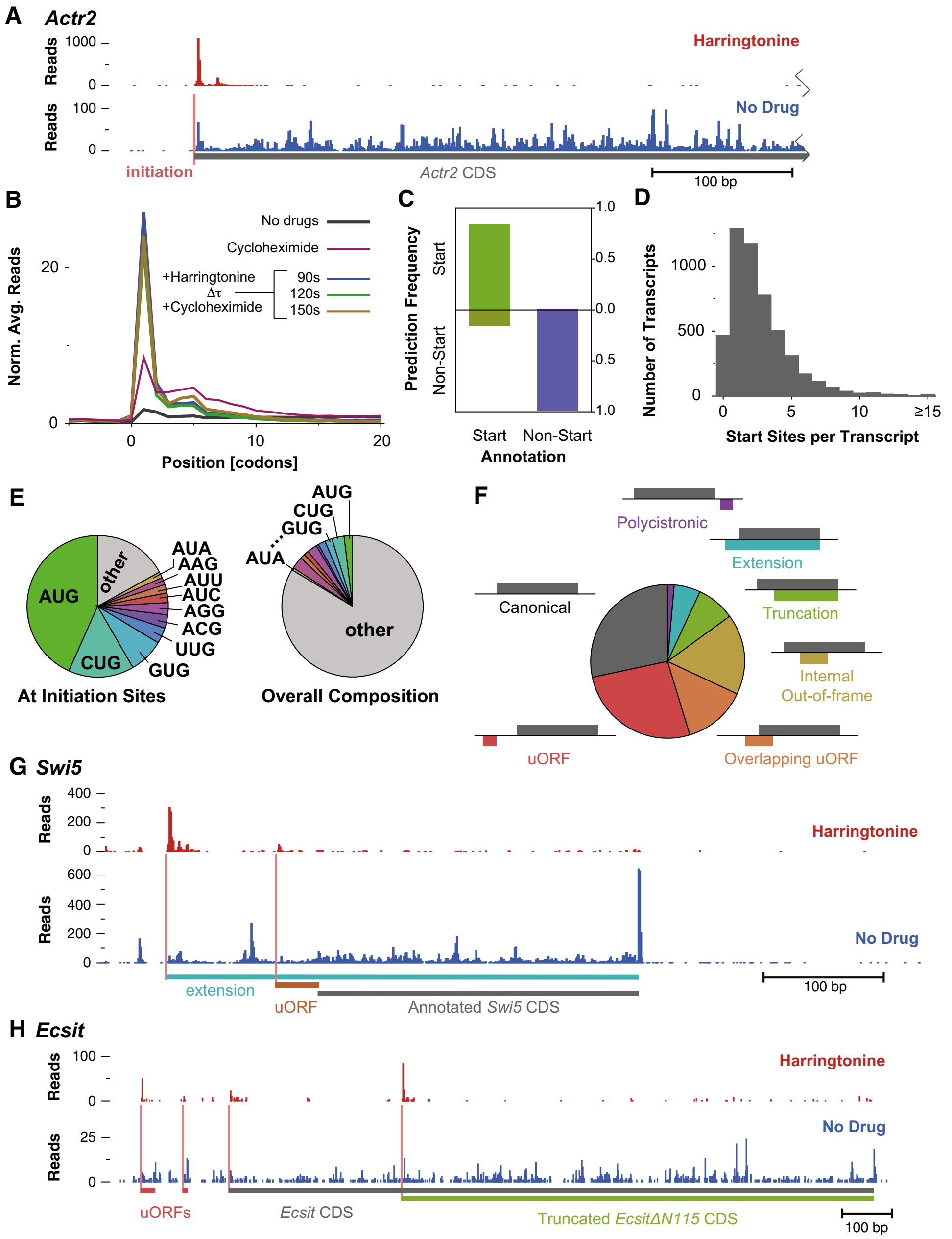

Figure 4. Harringtonine Enables Automated Identification of Translation Initiation Sites

(A) Effect of harringtonine on ribosome density for a typical gene. Ribosome footprint read count is shown prior to and following harringtonine treatment (150 s) along the 5′ UTR and the beginning of the CDS of *Actr2*.

(B) Metagene analysis of ribosome footprints surrounding start codons after harringtonine treatment. As in Figure 3B, focusing specifically on the site of translation initiation and the surrounding codons.

(C) Evaluation of start site prediction analysis. Plotted is the fraction of positive and negative initiation site predictions for start and selected nonstart codons that were excluded from the training set.

(D) Histogram of initiation sites predicted per transcript.

(E) Distribution of AUG codons and near-AUG codons at predicted sites of translation initiation (left), compared with the overall codon distribution (right).

(F) Classification of reading frames at predicted initiation sites relative to the annotated CDS.

(G) Pattern of initiation and translation on the *Swi5* transcript. As in Figure 4A, with the two detected initiation sites shown along with the respective reading frames, one of which produces a conserved amino-terminal extension on the *Swi5* protein.

(H) Pattern of initiation and translation on the *Ecsit* transcript. Four AUG initiation sites are present, two associated with uORFs and two with alternate protein isoforms of *Ecsit*.

See also Figure S2 and Table S3.

on a set of annotated genes to identify features of footprint profiles that distinguish the start codon from other positions. These profiles capture not just the accumulation of ribosomes at the start codon but also the distinctive asymmetric pattern of reads across flanking codons. Analysis of a distinct testing set of transcripts not used for training established that this model recognized 86% of annotated start codons as sites of translation initiation in comparison to only ~1% of other positions (Figures 4C and S2A). Actual false-negative and false-positive rates may be considerably lower, as not all annotated start sites are correct and there is a substantial rate of translation initiation from noncanonical start sites.

We applied the SVM approach to identify 13,454 candidate translation start sites within ~5000 transcripts that were well-expressed in our mESCs (Table S2A). The majority (65%) of these transcripts contain more than one detectable site of translation initiation, with 16% containing four or more sites (Figure 4D and Table S3). Although the analysis examined all potential translation start sites, we observed a dramatic enrichment for AUG (23-fold; Figure 4E), which provides an independent line of evidence for the accuracy of the SVM approach. We also found a strong enrichment for a specific subset of the near-cognate codons (i.e., codons that differ from AUG by a single nucleotide) at initiation sites (Figure 4E). Initiation at near-cognate sites is sometimes resistant to harringtonine (Starck et al., 2008; N. Stern-Ginossar and J.S.W., unpublished data), so our analysis may underestimate the true prevalence of near-cognate initiation.

Characterization of Alternate Open Reading Frames

We classified the reading frames downstream of the initiation sites we identified based on their relationship to the annotated ORF (Figure 4F). Nearly half (44%) of the AUG initiation sites that we found are unannotated, and the majority of these were downstream of the annotated start and were predicted to produce N-terminally truncated proteins or ORFs encoded in alternate reading frames (Figure S2B). In many cases, the annotated AUG was also used, and the alternate protein may not be the primary translation product. However, 280 of the genes with N-terminal truncations lacked detectable initiation on the annotated AUG, either because the annotated start codon is skipped in favor of the internal start site that we identified, or because the transcript is truncated and the annotated start codon is absent.

A substantial fraction (14%) of the initiation sites we observed are predicted to produce alternate protein isoforms of known genes (Figure 4F). We identified 570 genes with potential N-terminal extensions and 870 with N-terminal truncations in the 4,994 genes we analyzed. Extensions most often resulted from near-cognate initiation (Figure S2B), probably because computational gene annotation selects the first in-frame AUG, though conservation has been used to identify N-terminal extensions from near-cognate initiation (Ivanov et al., 2011). We found an N-terminal extension on the DNA repair protein *Swi5* (Figure 4G); its protein sequence is conserved, and there is experimental evidence that endogenous mouse *Swi5* is larger than the annotated 89 amino acid protein (Akamatsu and Jasin, 2010). Our data also revealed information about the protein products resulting from alternative splicing, which are often difficult to annotate. For instance, the growth factor *Igf2* has two 5′ untranslated region (UTR) variants with the same reading frame annotated in both transcripts, but we observed an isoform-specific N-terminal extension (Figure S2C).

The N-terminal truncations are of particular note as they can produce functionally distinct protein isoforms that lack an entire amino-terminal domain. For example, alternate start codons in the *Cebpa* gene can result in either a full-length transcription factor or a truncated dominant-negative isoform that contains the DNA-binding domain but not the full transactivation domain (Lin et al., 1993). We observe clear evidence of additional N-terminal truncations that could produce similar antagonistic products. Internal initiation in the Ets family transcription factor *Etv5* produces a product that lacks the predicted activation domain (Monté et al., 1996) but contains the domain that mediates DNA binding (Monté et al., 1994) (Figure S2D). This mechanism is not limited to transcription factors—internal initiation in the signaling scaffold *Ecsit* produces a protein nearly identical to a dominant-negative form created by designed N-terminal deletion (Figure 4H) (Kopp et al., 1999).

Exploring Translation of sprcRNAs

The above analysis focuses on known coding transcripts, but recently an abundant class of RNAs, referred to as lincRNAs, have been identified that lack the characteristics of conventional protein-coding genes. A limited number of lincRNAs such as *Xist* and *HotAir* have been shown to act at the RNA level in the nucleus (Brockdorff et al., 1992; Khalil et al., 2009), but the extent to which putative lincRNAs are translated is not known. Accordingly, we searched for translated regions within candidate lincRNAs (Guttman et al., 2009, 2010) by finding the most highly ribosome-occupied 90 nt window within the lincRNA and determining its translational efficiency as the ratio of ribosome footprint and mRNA-seq reads (Guttman et al., 2010, 2009). This analysis was very effective at distinguishing between traditional translated CDSes and their 3′ UTRs, which are poorly translated (Figure 5A).

Remarkably, the majority of putative lincRNAs contain regions of high translation comparable to protein-coding genes (Figure 5A and Table S4). We saw specific start sites marked by harringtonine followed by ribosome footprints extending to the first in-frame stop codon (Figures 5B–5D) (Clemson et al., 2009). These data establish that the majority of lincRNAs are exported to the cytoplasm and effectively engaged by the protein translation machinery. These included roughly half of the lincRNA candidates that were recently shown to be required for maintenance of pluripotency (Guttman et al., 2011). We classify these RNAs as sprcRNAs based on our observation that they contain small CDSes that are bound by elongating ribosomes and frequently contain multiple ORFs. We also identify a significant subset of true lincRNAs that are not translated, including the well-documented RNA element NEAT1, which regulates mRNA export (Clemson et al., 2009). The extent to which various RNAs act through their translation products and/or directly through their transcript remains a central open question that our dataset should provide a critical resource for addressing.

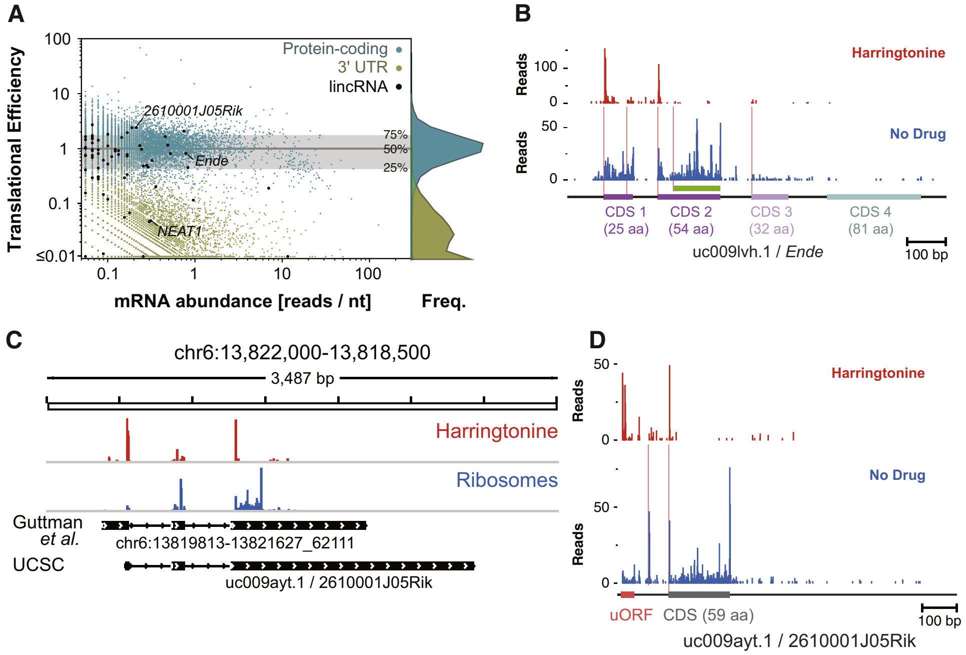

Figure 5. Translation of sprcRNAs

(A) Translational efficiency of putative lincRNAs. The translational efficiency, a normalized ratio of ribosome footprint density to mRNA-seq read density, is plotted for the most highly occupied 90 nt window of each lincRNA, protein-coding gene, and coding transcript 3′ UTR, along with a histogram of translational efficiency values for CDSes and 3′ UTRs and the median and quartile values for protein-coding genes.

(B) Ribosome footprint profile of the uc009lvh.1 transcript. This RNA is annotated as a noncoding RNA, but we identify two short (25 and 54 amino acids) well-translated ORFs and see little translation from a longer (81 amino acid) downstream CDS hypothesized to encode a protein (Hassan et al., 2010).

(C) Ribosome footprint profile of the 2610001J05Rik genomic locus. The profile includes transcript-aligned reads mapped to corresponding genomic positions and genomic-aligned reads with no transcript alignment. The annotated noncoding uc009ayt.1 transcript is shown along with the reconstructed transcript (Guttman et al., 2010).

(D) Ribosome footprint profile of the uc009ayt.1 transcript.

See also Table S4.

Widespread Translation of uORFs

The majority of unannotated near-cognate initiation sites we detected drive the translation of upstream open reading frames (uORFs) (Figures 6A and S4B). This is consistent with the high level of translation that we observe on many 5′ UTRs as opposed to 3′ UTRs, which are almost devoid of ribosomes. These uORF initiation sites are accompanied by elongating ribosome footprints in the untreated sample that are depleted during harringtonine treatment, indicating that they are involved in active translation (Figure 4B). In a few well-studied examples, uORFs have been shown to affect translation of downstream genes. The first uORF in the *Atf4* transcript is constitutively translated, and ribosomes then reinitiate at either the second uORF or the CDS (Calvo et al., 2009; Lu et al., 2004; Morris and Geballe, 2000) (Figure 6B). This exemplifies two roles of uORFs—some permit downstream reinitiation, whereas others capture some fraction of scanning preinitiation complexes and decrease CDS translation. There are a small number of well-documented uORFs with near-cognate start codons (Ivanov et al., 2008), but there are no effective computational approaches for identifying them. Our observations suggest that near-cognate uORFs are quite common. The ribosome footprint profiles of *Myc* and *Nanog*, two genes that play a critical role in pluripotency, illustrate the complexity of translation; both have multiple uORFs and alternate translation products initiating at both AUG and near-cognate sites (Figures 6C and 6D).

Due to the prevalence of alternative transcription start sites and alternative splicing, many genes have multiple 5′ UTR isoforms, potentially including distinct regulatory information (Hughes, 2006). Many initiation sites occurred in alternative UTRs; we found 1,800 genes showing differential initiation of uORFs in distinct 5′ UTR isoforms. We additionally observed that at least 30% of these genes showed a significant difference in the ratio of ribosome footprint to mRNA-seq reads between the distinct 5′ UTRs of different isoforms. Thus, alternative splicing generates transcripts with different upstream initiation sites and results in different uORF translation. For example, the transcription factor *Atf5* is regulated by well-characterized uORFs in one mRNA isoform that are missing from a less-abundant isoform expressed in early development (Hansen et al., 2002). We observe robust translation initiation at a distinct uORF in this second isoform (Figure 6E). Alternative inclusion of uORFs was also seen in ribosomal proteins, including *Rps27a*, where a small fraction of transcripts had a retained 5′ UTR intron that introduced a uORF (Figure 6F). In the particular

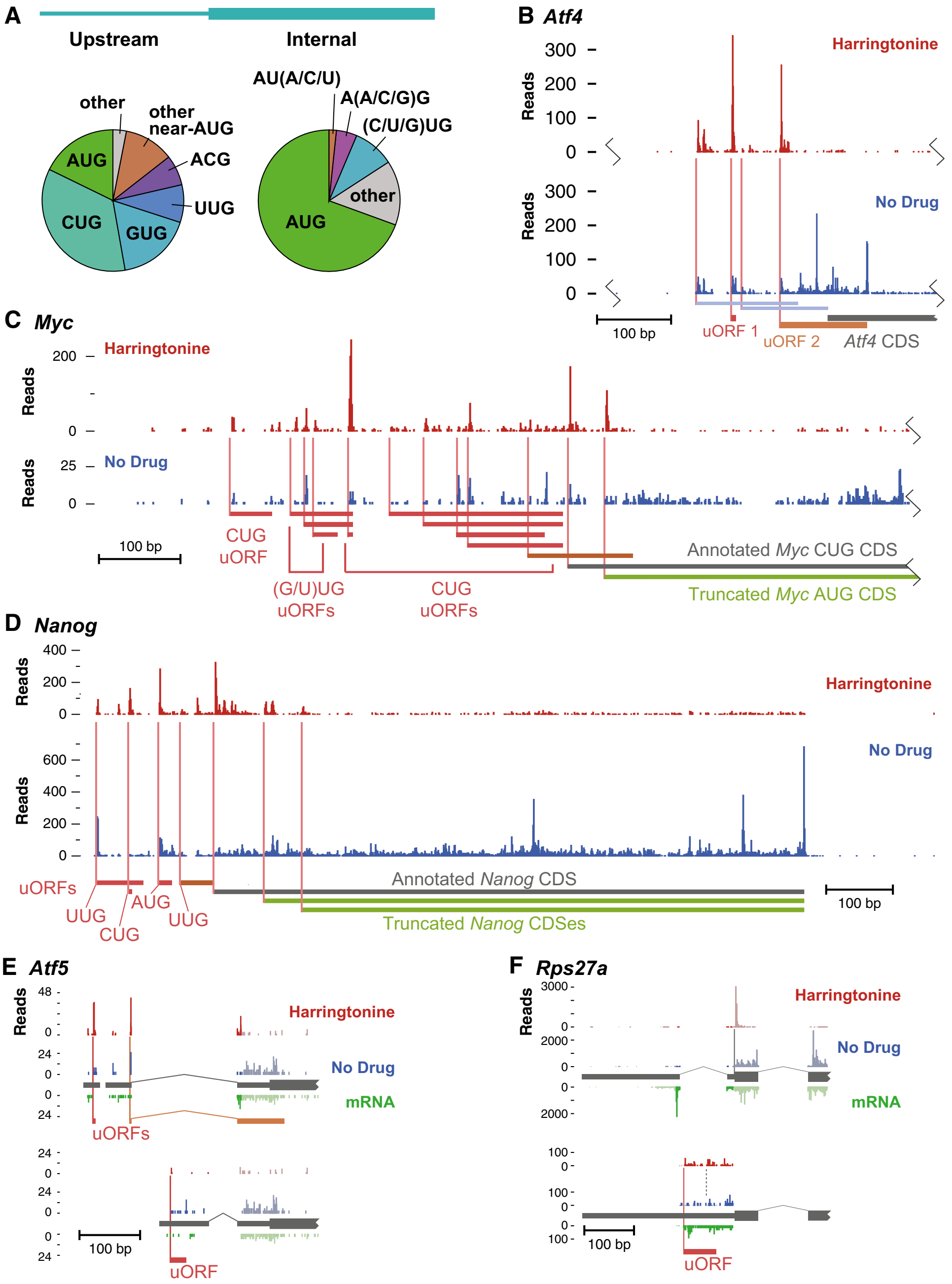

Figure 6. Translation of Regulatory uORFs and Alternatively Processed Transcripts

(A) Codon distribution at upstream (left) and internal (right) translation initiation sites. Internal sites are only taken from codons 15 through 300, as internal sites further downstream are affected by incomplete ribosome run-off during short harringtonine treatment.

(B) Patterns of initiation and translation on the *Atf4* transcript. The two characterized regulatory uORFs, initiated by AUG codons, are highlighted. Two weak non-AUG reading frames are shown in blue.

(C) As in (B) on the *Myc* transcript. Several near-cognate sites of upstream initiation are shown, along with the annotated CUG initiation codon and the alternate AUG initiation codon.

(D) As in (C) on the *Nanog* transcript. Upstream ORFs are shown, along with the CDS and two in-frame AUG initiation sites within the CDS.

(E) Patterns of initiation and translation on the 5′ ends of two transcripts of the *Atf5* gene. The exon structure is shown with thin gray rectangles for the 5′ UTR and thick gray rectangles for the annotated CDS. An mRNA-seq read profile is shown on an inverted y axis. Isoform-specific transcript positions are shown in dark colors and non-isoform-specific positions are shown with faint colors. The major isoform (top) has two uORFs that confer translational regulation on the CDS; a distinct uORF is observed in the minor embryonic isoform (bottom).

(F) As (E), for the 5′ ends of the *Rpl27a* transcripts. Only the isoform-specific positions are shown for the minor isoform (bottom), scaled 10×.

See also Figure S3.

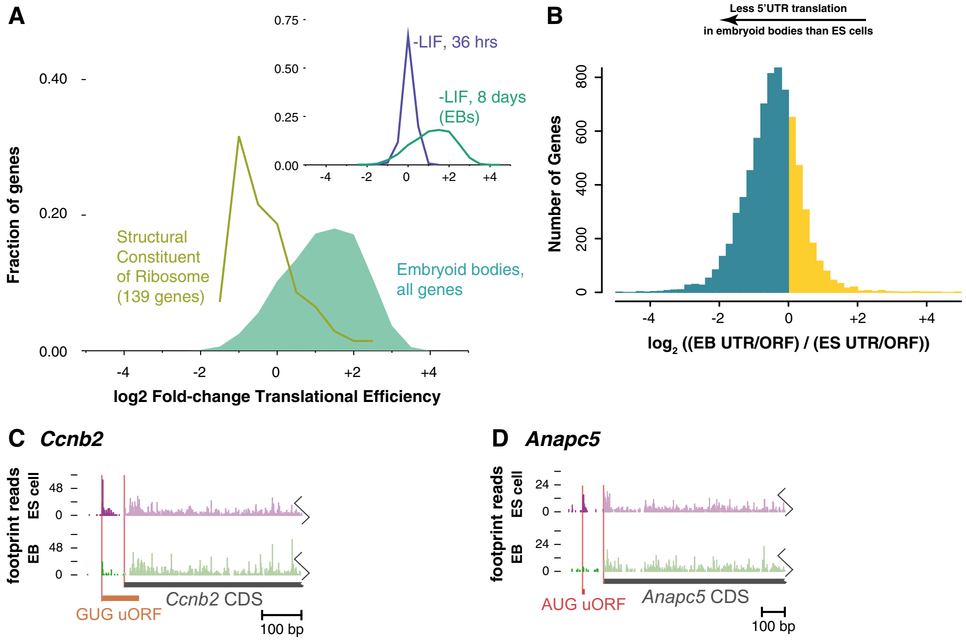

Figure 7. Changes in Upstream Translation during Differentiation

(A) Translational regulation following LIF withdrawal. The distribution of $\log_2$ fold-changes of translational efficiency (ratio of sample-normalized ribosome footprint density to mRNA-seq density) is shown for all genes and for those with the GO annotation "structural constituent of ribosome" (see Table S5D). Inset: distributions for all genes, 36 hr and 8 days after LIF withdrawal (see Tables S5A and S5D).

(B) Changes in relative upstream translation in EBs versus ESCs. The ratio of footprints between the 5′ UTR and the ORF was computed for each gene, and the distribution of $\log_2$ change in the 5′ UTR/ORF ratio is plotted, with decreases in EB shown in blue and increases in EB shown in yellow.

(C and D) Patterns of translation on the *Ccnb5* (C) and *Anapc5* (D) transcripts. Ribosome footprints that map to the 5′ UTR are in dark colors, and the CDS in faint colors. The average, sample-normalized ribosome footprint density on the CDS is slightly higher in the EB sample than in the ESC sample for both.

See also Figure S4 and Table S5.

case of isoforms where an alternative UTR splice junction is quite close to the shared start codon, ribosome footprints from initiation at the start codon can include enough distinct upstream sequence to distinguish the effect of different UTRs. The gene *Pih1d1* has two 5′ UTR variants with distinct uORFs. Strong initiation of the uORF in one isoform led to 50% less initiation of its protein-coding reading frame as compared to initiation of the same protein-coding reading frame in the second isoform (Figure S3). This effect demonstrates the potential impact of the widespread upstream initiation we observe in both alternative and constitutive 5′ UTRs.

Changes in Translation during Embryoid Body Formation

We next asked how the landscape of translation changes when proliferative, pluripotent ESCs undergo differentiation into embryoid bodies (EBs). Withdrawal of leukemia inhibitory factor (LIF) induced differentiation (Figure S4A), which we assessed visually and by the downregulation of the direct LIF target Klf4 (Niwa et al., 2009), followed by loss of Oct4 expression and the induction of developmental and lineage-specific genes (Figures S4B and S4C and Tables S5A–S5F). We then looked for translational control of gene expression during differentiation and observed strong repression of ribosomal proteins (RPs) in EBs relative to ESCs (Figures 7A and S4D and Table S5F). Although these genes were still highly expressed in EBs, they were translated 3- to 4-fold less efficiently than the typical transcript (Tables S5D–S5F). The translation of RPs is regulated in response to proliferation and nutrient status (Hamilton et al., 2006), and here we show that this response is a notable feature of EB formation. Polysome-profiling experiments have suggested a global increase in cellular translation during early ESC differentiation, and we see a modest upregulation of RPs in our early time point (Sampath et al., 2008). This might lead to a surfeit of ribosomes at the later stage of EB formation. Intriguingly, Akt/mTOR signaling controls RP expression and may regulate translation during differentiation more generally (Di Cristofano et al., 1998; Sampath et al., 2008). We also observed a modest but quite significant increase in the translational efficiency of integral membrane proteins in EBs (Figure S4E and Table S5F), which could result directly from a redirection of ribosomes to the rough ER, or indirectly through regulatory programs whose targets are enriched for membrane proteins.

Translation of uORFs also declined substantially during differentiation. We measured the level of upstream translation using the ratio of ribosome footprint reads in the 5′ UTR to the CDS of each gene and found that the typical transcript showed an ~25% decrease in 5′ UTR translation during differentiation (Figure 7B). This shift can be observed on the 5′ UTR of individual genes with defined uORFs (Figures 7C and 7D). It reflects a broad change in the translational apparatus with the potential to impact gene expression genome-wide. Reduced upstream translation might reflect a relative decrease in cap-dependent versus cap-independent initiation, as cap-dependent initiation would be expected to favor upstream sites near the cap. Such a shift has been associated with proliferation in tumorigenesis and has been linked to the translational control of RPs (Mamane et al., 2006; Ruggero and Sonenberg, 2005). This tumor cell translational program may also be active in ESCs.

DISCUSSION

Here we present a range of ribosome-profiling techniques, based on deep sequencing of ribosome-protected fragments, that dramatically expand our ability to define and quantitatively monitor mammalian proteomes. Our approaches provide experimentally based maps of the protein-coding potential of complex genomes and reveal in-depth information about the kinetics and mechanism of translation elongation and coupled cotranslational events. Finally, ribosome profiling allows high-precision, genome-wide measures of the rate of protein synthesis from the density of ribosome footprints, much as RNA-seq experiments measure mRNA abundance from read density; such gene expression measurements may represent the most frequent application of ribosome profiling even after the proteome is fully defined. Although there have been remarkable advances in quantitative mass spectrometry (Nilsson et al., 2010), it is difficult to match the large dynamic range and comprehensive nature of deep sequencing. More generally mass spectrometry and ribosome profiling represent highly complementary approaches; for example, comparison between changes in rate of synthesis measured by ribosome profiling and abundance measured by mass spectrometry should reveal examples of regulated degradation of proteins.

A number of features of mammalian proteomes emerge from our studies, including the ubiquitous use of alternate initiation sites that drive the production of extended or truncated isoforms of known proteins as well as the translation of sprcRNAs, whose protein-coding potential was not initially apparent. We also observe widespread translation upstream of mammalian protein-coding genes, similar to but more extensive than upstream translation that we observed in yeast (Ingolia et al., 2009). Translation of uORFs can modulate the expression of the downstream protein-coding gene in response to global (Sonenberg and Hinnebusch, 2009) or gene-specific regulatory signals (Medenbach et al., 2011). We have shown that upstream translation decreases as ESCs undergo differentiation, indicating that it is subject to regulation and may be part of a major program of translational control.

Our studies also establish that many sites of translation initiation, especially upstream initiation, occur at non-AUG codons. Although most productive protein synthesis starts at a classical AUG codon, initiation at CUG and GUG codons is widespread and is likely to have broad biological significance. An important open question is how this non-AUG initiation differs mechanistically from AUG initiation and what factors regulate initiation site selection. The bias toward upstream non-AUG initiation seems to conflict with a pure scanning model for start codon recognition, as a preinitiation complex that bypasses the annotated AUG is no less likely to recognize a subsequent CUG, though the difference could reflect heterogeneous stringencies in scanning complexes.

Non-AUG initiation clearly impacts many aspects of translation. The extensive upstream non-AUG initiation we observe is likely to regulate protein synthesis from specific transcripts in response to global changes in initiation. It is also regulated during EB formation, suggesting a global link with growth and proliferation, and is involved in the synthesis of functional proteins, including the well-studied oncogene and pluripotency factor *Myc* (Hann et al., 1988). More broadly, it has been implicated in the production of peptides for immune surveillance (Malarkannan et al., 1999), and additional roles will likely emerge as we understand more about which non-AUG codons are used and how this selection is regulated.

EXPERIMENTAL PROCEDURES

Ribosome Footprinting

E14 mESCs were propagated in standard culture feeder-free conditions (Tremml et al., 2008), and differentiation was induced by transferring cells to media lacking LIF in low-adhesion dishes. Cells were pretreated with harringtonine (2 μg/ml), cycloheximide (100 μg/ml), and/or emetine (20 μg/ml) as indicated, and detergent lysis was performed in the dish. The lysate was DNase-treated and clarified, and a sample was taken for mRNA-Seq analysis. Lysates were subject to ribosome footprinting by nuclease treatment. Footprint fragments were purified, and deep sequencing libraries were generated from these fragments, as well as from poly(A) mRNA purified from untreated lysate. These libraries were analyzed by sequencing on the Illumina GAII and HiSeq.

Footprint Sequence Alignment

Sequences were aligned to a library of transcripts derived from the UCSC Known Genes data set (Hsu et al., 2006) and the reconstructed mESC transcriptome of Guttman et al. (Guttman et al., 2010), and those with no acceptable transcript alignment were then aligned against the genome. Because sequencing reads comprise a variable-length RNA fragment followed by a linker sequence, the first 26 nucleotides were aligned against the reference database using Bowtie, and this alignment was extended until it reached the known linker sequence. Alignments were accepted with up to two mismatches, and multiple alignments were allowed for a single sequence, but alignments with fewer mismatches were preferred.

For most analyses, footprint alignments were assigned to specific A site nucleotides by using the position and total length of each alignment, calibrated from footprints at the beginning and the end of CDSes (Figures S1A and S1B) as previously described (Ingolia et al., 2009).

Footprint Profile Analysis

Profiles of ribosome footprints across a transcript were constructed by quantifying the number of footprints assigned to each nucleotide position. A set of well-expressed genes was selected based on median footprint density across the CDS, excluding the first 15 and last 5 codons due to the accumulation of ribosomes (Figures 1C and 1D). To construct metagene density profiles, individual gene profiles were scaled by their footprint density in the untreated control, and all were averaged with equal weight.

Harringtonine Depletion Profile Analysis

Metagene profiles from harringtonine run-off were further normalized by the median value over codons 800–1000, which appeared undepleted at harringtonine treatment times used in this study, and smoothed by averaging disjoint 5 codon windows. The extent of depletion was defined as the earliest codon position, beyond the first 40, that retained at least 50% of the full ribosome density. Subsets of genes for elongation rate analysis were as follows: (1) lowest and highest quintile of tAI, computed according to dos Reis et al. (2004); (2) lowest and highest quintile of ribosome footprint density; (3) short genes, 750–1,000 codons, and long genes, over 1,000 codons; (4) secreted proteins that were identified using SignalP data from Ensembl.

Initiation Site Prediction

Initiation site predictions for each nucleotide position were based on a vector of footprint read counts over 15 codons around the position for each harringtonine sample, concatenated to produce an overall vector. The SVMlight pattern-recognition tool (Joachims, 1999) was trained on an arbitrary set of 3,200 transcripts, using the annotated start codon as a positive example and ten other positions as negative examples.

Initiation sites were defined as one or more consecutive nucleotide positions that passed an SVM score threshold as well as a minimum of 50 harringtonine footprints total among all samples. These consecutive blocks were typically (91%) three or fewer nucleotides long and in no case longer than six nucleotides (Table S3). Initiation sites that contained an AUG codon were assigned to that codon or, if none was present, to any near-cognate codons, and the reading frame was predicted from that codon. Sites with no recognizable initiation codon or with multiple potential near-cognate codons could not be assigned to a specific reading frame and were eliminated from further analyses. The preferential assignment of initiation sites to AUG codons may lead to a modest bias against detecting near-cognate initiation.

lincRNA Analysis

lincRNAs were collected from reconstructed transcripts (Guttman et al., 2010) that lay entirely within the lincRNA chromatin signatures identified by Guttman et al. (2009), which excluded known protein-coding genes. Footprint density profiles from the untreated sample were analyzed to identify the 90 nt window with the most positions occupied by at least one ribosome footprint among all transcripts in the chromatin region. For annotated protein-coding transcripts, the CDS and the 3′ UTR were analyzed separately. The mRNA abundance was calculated as the density of mRNA-seq reads in the window, and the translational efficiency was calculated as the ratio between the ribosome footprint and the mRNA-seq read density in the window.

ACCESSION NUMBERS

Sequencing data were deposited in the GEO database with accession number GSE30839.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, four figures, and five tables and can be found with this article online at doi:10.1016/j.cell.2011.10.002.

ACKNOWLEDGMENTS

We thank M. Barna, S. Luo, D. Ruggero G. Schroth, P. Brown, and members of the Weissman lab including A. Frost and S. Churchman for advice and helpful discussions; C. Chu for help with sequencing; S. Rouskin for help developing the ligation protocol; G. Brar for help developing rRNA depletion; and J. Hunkapiller and J. Reiter for E14 mESCs. This investigation was supported by the NIH P01 (AG10770) and a Ruth L. Kirschstein NRSA (GM080853) (N.T.I.), the Damon Runyon Cancer Research Foundation (DRG-2033-09) (L.F.L.), and the Howard Hughes Medical Institute (J.S.W.)

Received: February 26, 2011
Revised: June 19, 2011
Accepted: September 13, 2011
Published online: November 3, 2011

REFERENCES

Akamatsu, Y., and Jasin, M. (2010). Role for the mammalian Swi5-Sfr1 complex in DNA strand break repair through homologous recombination. PLoS Genet. *6*, e1001160.

Alkalaeva, E.Z., Pisarev, A.V., Frolova, L.Y., Kisselev, L.L., and Pestova, T.V. (2006). In vitro reconstitution of eukaryotic translation reveals cooperativity between release factors eRF1 and eRF3. Cell *125*, 1125–1136.

Arava, Y., Wang, Y., Storey, J.D., Liu, C.L., Brown, P.O., and Herschlag, D. (2003). Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. USA *100*, 3889–3894.

Atkins, J.F., and Gesteland, R.F. (2010). Recoding: Expansion of Decoding Rules Enriches Gene Expression (New York: Springer).

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature *456*, 53–59.

Boström, K., Wettesten, M., Borén, J., Bondjers, G., Wiklund, O., and Olofsson, S.O. (1986). Pulse-chase studies of the synthesis and intracellular transport of apolipoprotein B-100 in Hep G2 cells. J. Biol. Chem. *261*, 13800–13806.

Brent, M.R. (2005). Genome annotation past, present, and future: how to define an ORF at each locus. Genome Res. *15*, 1777–1786.

Brockdorff, N., Ashworth, A., Kay, G.F., McCabe, V.M., Norris, D.P., Cooper, P.J., Swift, S., and Rastan, S. (1992). The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. Cell *71*, 515–526.

Calfon, M., Zeng, H., Urano, F., Till, J.H., Hubbard, S.R., Harding, H.P., Clark, S.G., and Ron, D. (2002). IRE1 couples endoplasmic reticulum load to secretory capacity by processing the XBP-1 mRNA. Nature *415*, 92–96.

Calvo, S.E., Pagliarini, D.J., and Mootha, V.K. (2009). Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc. Natl. Acad. Sci. USA *106*, 7507–7512.

Chi, S.W., Zang, J.B., Mele, A., and Darnell, R.B. (2009). Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature *460*, 479–486.

Clemson, C.M., Hutchinson, J.N., Sara, S.A., Ensminger, A.W., Fox, A.H., Chess, A., and Lawrence, J.B. (2009). An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. Mol. Cell *33*, 717–726.

Darnell, J.C., Van Driesche, S.J., Zhang, C., Hung, K.Y., Mele, A., Fraser, C.E., Stone, E.F., Chen, C., Fak, J.J., Chi, S.W., et al. (2011). FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. Cell *146*, 247–261.

Di Cristofano, A., Pesce, B., Cordon-Cardo, C., and Pandolfi, P.P. (1998). Pten is essential for embryonic development and tumour suppression. Nat. Genet. *19*, 348–355.

dos Reis, M., Savva, R., and Wernisch, L. (2004). Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. *32*, 5036–5044.

Fresno, M., Jiménez, A., and Vázquez, D. (1977). Inhibition of translation in eukaryotic systems by harringtonine. Eur. J. Biochem. *72*, 323–330.

Guo, H., Ingolia, N.T., Weissman, J.S., and Bartel, D.P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature *466*, 835–840.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223–227.

Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C., et al. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat. Biotechnol. *28*, 503–510.

Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., et al. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature. Published online August 28 2011. 10.1038/nature10398.

Hamilton, T.L., Stoneley, M., Spriggs, K.A., and Bushell, M. (2006). TOPs and their regulation. Biochem. Soc. Trans. *34*, 12–16.

Hann, S.R., King, M.W., Bentley, D.L., Anderson, C.W., and Eisenman, R.N. (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.

Hansen, M.B., Mitchelmore, C., Kjaerulff, K.M., Rasmussen, T.E., Pedersen, K.M., and Jensen, N.A. (2002). Mouse Atf5: molecular cloning of two novel mRNAs, genomic organization, and odorant sensory neuron localization. Genomics *80*, 344–350.

Hassan, A.S., Hou, J., Wei, W., and Hoodless, P.A. (2010). Expression of two novel transcripts in the mouse definitive endoderm. Gene Expr. Patterns *10*, 127–134.

Hsu, F., Kent, W.J., Clawson, H., Kuhn, R.M., Diekhans, M., and Haussler, D. (2006). The UCSC Known Genes. Bioinformatics *22*, 1036–1046.

Huang, M.T. (1975). Harringtonine, an inhibitor of initiation of protein biosynthesis. Mol. Pharmacol. *11*, 511–519.

Hughes, T.A. (2006). Regulation of gene expression by alternative untranslated regions. Trends Genet. *22*, 119–122.

Ingolia, N.T. (2010). Genome-wide translational profiling by ribosome footprinting. Methods Enzymol. *470*, 119–142.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218–223.

Irwin, B., Heck, J.D., and Hatfield, G.W. (1995). Codon pair utilization biases influence translational elongation step times. J. Biol. Chem. *270*, 22801–22806.

Ivanov, I.P., Loughran, G., and Atkins, J.F. (2008). uORFs with unusual translational start codons autoregulate expression of eukaryotic ornithine decarboxylase homologs. Proc. Natl. Acad. Sci. USA *105*, 10079–10084.

Ivanov, I.P., Firth, A.E., Michel, A.M., Atkins, J.F., and Baranov, P.V. (2011). Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences. Nucleic Acids Res. *39*, 4220–4234.

Joachims, T. (1999). Making large-scale SVM learning practical. In Advances in Kernel Methods - Support Vector Learning, B. Schölkopf, C. Burges, and A. Smola, eds. (Cambridge, MA: MIT Press).

Khalil, A.M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B.E., van Oudenaarden, A., et al. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc. Natl. Acad. Sci. USA *106*, 11667–11672.

Kimchi-Sarfaty, C., Oh, J.M., Kim, I.W., Sauna, Z.E., Calcagno, A.M., Ambudkar, S.V., and Gottesman, M.M. (2007). A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science *315*, 525–528.

Kondo, T., Plaza, S., Zanet, J., Benrabah, E., Valenti, P., Hashimoto, Y., Kobayashi, S., Payre, F., and Kageyama, Y. (2010). Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis. Science *329*, 336–339.

Kopp, E., Medzhitov, R., Carothers, J., Xiao, C., Douglas, I., Janeway, C.A., and Ghosh, S. (1999). ECSIT is an evolutionarily conserved intermediate in the Toll/IL-1 signal transduction pathway. Genes Dev. *13*, 2059–2071.

Kraut-Cohen, J., and Gerst, J.E. (2010). Addressing mRNAs to the ER: cis sequences act up!. Trends Biochem. Sci. *35*, 459–469.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al; International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Lau, N.C., Lim, L.P., Weinstein, E.G., and Bartel, D.P. (2001). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science *294*, 858–862.

Lazarowitz, S.G., and Robertson, H.D. (1977). Initiator regions from the small size class of reovirus messenger RNA protected by rabbit reticulocyte ribosomes. J. Biol. Chem. *252*, 7842–7849.

Lin, F.T., MacDougald, O.A., Diehl, A.M., and Lane, M.D. (1993). A 30-kDa alternative translation product of the CCAAT/enhancer binding protein alpha message: transcriptional activator lacking antimitotic activity. Proc. Natl. Acad. Sci. USA *90*, 9606–9610.

Lu, P.D., Harding, H.P., and Ron, D. (2004). Translation reinitiation at alternative open reading frames regulates gene expression in an integrated stress response. J. Cell Biol. *167*, 27–33.

Lunde, B.M., Moore, C., and Varani, G. (2007). RNA-binding proteins: modular design for efficient function. Nat. Rev. Mol. Cell Biol. *8*, 479–490.

Malarkannan, S., Horng, T., Shih, P.P., Schwab, S., and Shastri, N. (1999). Presentation of out-of-frame peptide/MHC class I complexes by a novel translation initiation mechanism. Immunity *10*, 681–690.

Mamane, Y., Petroulakis, E., LeBacquer, O., and Sonenberg, N. (2006). mTOR, translation initiation and cancer. Oncogene *25*, 6416–6422.

Mariappan, M., Li, X., Stefanovic, S., Sharma, A., Mateja, A., Keenan, R.J., and Hegde, R.S. (2010). A ribosome-associating factor chaperones tail-anchored membrane proteins. Nature *466*, 1120–1124.

Martin, K.C., and Ephrussi, A. (2009). mRNA localization: gene expression in the spatial dimension. Cell *136*, 719–730.

Medenbach, J., Seiler, M., and Hentze, M.W. (2011). Translational control via protein-regulated upstream open reading frames. Cell *145*, 902–913.

Monté, D., Baert, J.L., Defossez, P.A., de Launoit, Y., and Stéhelin, D. (1994). Molecular cloning and characterization of human ERM, a new member of the Ets family closely related to mouse PEA3 and ER81 transcription factors. Oncogene *9*, 1397–1406.

Monté, D., Coutte, L., Dewitte, F., Defossez, P.A., Le Coniat, M., Stéhelin, D., Berger, R., and de Launoit, Y. (1996). Genomic organization of the human ERM (ETV5) gene, a PEA3 group member of ETS transcription factors. Genomics *35*, 236–240.

Morris, D.R., and Geballe, A.P. (2000). Upstream open reading frames as regulators of mRNA translation. Mol. Cell. Biol. *20*, 8635–8642.

Nakatogawa, H., and Ito, K. (2002). The ribosomal exit tunnel functions as a discriminating gate. Cell *108*, 629–636.

Namy, O., Moran, S.J., Stuart, D.I., Gilbert, R.J., and Brierley, I. (2006). A mechanical explanation of RNA pseudoknot function in programmed ribosomal frameshifting. Nature *441*, 244–247.

Nilsson, T., Mann, M., Aebersold, R., Yates, J.R., 3rd, Bairoch, A., and Bergeron, J.J. (2010). Mass spectrometry in high-throughput proteomics: ready for the big time. Nat. Methods *7*, 681–685.

Niwa, H., Ogawa, K., Shimosato, D., and Adachi, K. (2009). A parallel circuit of LIF signalling pathways maintains pluripotency of mouse ES cells. Nature *460*, 118–122.

Noble, W.S. (2006). What is a support vector machine? Nat. Biotechnol. *24*, 1565–1567.

Pfeffer, S., Sewer, A., Lagos-Quintana, M., Sheridan, R., Sander, C., Grässer, F.A., van Dyk, L.F., Ho, C.K., Shuman, S., Chien, M., et al. (2005). Identification of microRNAs of the herpesvirus family. Nat. Methods *2*, 269–276.

Pisarev, A.V., Kolupaeva, V.G., Yusupov, M.M., Hellen, C.U., and Pestova, T.V. (2008). Ribosomal position and contacts of mRNA in eukaryotic translation initiation complexes. EMBO J. *27*, 1609–1621.

Robert, F., Carrier, M., Rawe, S., Chen, S., Lowe, S., and Pelletier, J. (2009). Altering chemosensitivity by modulating translation elongation. PLoS ONE *4*, e5428.

Ruggero, D., and Sonenberg, N. (2005). The Akt of translational control. Oncogene *24*, 7426–7434.

Sampath, P., Pritchard, D.K., Pabon, L., Reinecke, H., Schwartz, S.M., Morris, D.R., and Murry, C.E. (2008). A hierarchical network controls protein translation during murine embryonic stem cell self-renewal and differentiation. Cell Stem Cell *2*, 448–460.

Schneider-Poetsch, T., Ju, J., Eyler, D.E., Dang, Y., Bhat, S., Merrick, W.C., Green, R., Shen, B., and Liu, J.O. (2010). Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin. Nat. Chem. Biol. *6*, 209–217.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–342.

Sonenberg, N., and Hinnebusch, A.G. (2009). Regulation of translation initiation in eukaryotes: mechanisms and biological targets. Cell *136*, 731–745.

Starck, S.R., Ow, Y., Jiang, V., Tokuyama, M., Rivera, M., Qi, X., Roberts, R.W., and Shastri, N. (2008). A distinct translation initiation mechanism generates cryptic peptides for immune surveillance. PLoS ONE *3*, e3460.

Tanner, D.R., Cariello, D.A., Woolstenhulme, C.J., Broadbent, M.A., and Buskirk, A.R. (2009). Genetic identification of nascent peptides that induce ribosome stalling. J. Biol. Chem. *284*, 34809–34818.

Tenson, T., and Ehrenberg, M. (2002). Regulatory nascent peptides in the ribosomal tunnel. Cell *108*, 591–594.

Tremml, G., Singer, M., and Malavarca, R. (2008). Culture of mouse embryonic stem cells. In Current Protocols in Stem Cell Biology Chapter 1, Unit 1C 4.

Tscherne, J.S., and Pestka, S. (1975). Inhibition of protein synthesis in intact HeLa cells. Antimicrob. Agents Chemother. *8*, 479–487.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., and Pilpel, Y. (2010). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell *141*, 344–354.

Tupy, J.L., Bailey, A.M., Dailey, G., Evans-Holm, M., Siebel, C.W., Misra, S., Celniker, S.E., and Rubin, G.M. (2005). Identification of putative noncoding polyadenylated transcripts in Drosophila melanogaster. Proc. Natl. Acad. Sci. USA *102*, 5495–5500.

Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R.B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. Science *302*, 1212–1215.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science *291*, 1304–1351.

Wolin, S.L., and Walter, P. (1988). Ribosome pausing and stacking during translation of a eukaryotic mRNA. EMBO J. *7*, 3559–3569.

Yanagitani, K., Imagawa, Y., Iwawaki, T., Hosoda, A., Saito, M., Kimata, Y., and Kohno, K. (2009). Cotranslational targeting of XBP1 protein to the membrane promotes cytoplasmic splicing of its own mRNA. Mol. Cell *34*, 191–200.

Yanagitani, K., Kimata, Y., Kadokura, H., and Kohno, K. (2011). Translational pausing ensures membrane targeting and cytoplasmic splicing of XBP1u mRNA. Science *331*, 586–589.

Zhang, G., Hubalewska, M., and Ignatova, Z. (2009). Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. Nat. Struct. Mol. Biol. *16*, 274–280.

EXTENDED EXPERIMENTAL PROCEDURES

Cell Culture and Drug Treatment

E14 mESCs were plated on 15 cm dishes coated with gelatin (0.1% in Dulbecco's phosphate-buffered saline without calcium or magnesium salts) and grown in GMEM supplemented with 10% FBS, nonessential amino acids, glutamine, pyruvate, β-mercaptoethanol, and leukemia inhibitory factor (Tremml et al., 2008). Harringtonine (LKT Laboratories) treatment was performed by adding the drug to a final concentration of 2 μg/ml from a 2 mg/ml stock in DMSO. Cells were returned to 37°C following drug addition. Cycloheximide (Sigma) treatment was performed by adding the drug to a final concentration of 100 μg/ml from a 50 mg/ml stock in 100 percent EtOH. Cells were returned to 37°C for 1 min following drug addition. Emetine (Sigma) treatment was performed in a similar manner, adding the drug to a final concentration of 20 μg/ml from a 100 mg/ml stock in DMSO.

Lysis

Medium was aspirated from dishes, which were immediately placed on ice and rinsed with 10 ml ice-cold PBS supplemented with drugs used in pretreatment of the cells. PBS was aspirated and 800 μl ice-cold lysis buffer consisting of polysome buffer (20 mM Tris, pH 7.4, 250 mM NaCl, 15 mM $MgCl_2$, 1 mM dithiothreitol) supplemented with 0.5% Triton X-100 and 24 U / ml Turbo DNase (Ambion, AM2239), along with any drugs used for sample treatment) was dripped onto dishes. Cells were scraped extensively and triturated through a pipette. The lysate was removed and incubated 10 min on ice with periodic agitation. The lysate was then clarified by centrifugation for 10 min at 20,000 × g, 4°C and ~1.1 ml supernatant was recovered.

Ribosome Footprinting

A 600 μl aliquot of lysate was treated with 15 μl RNase I 100 U/μl (Ambion, AM2295) for 45 min at room temperature with gentle agitation. The digestion was stopped by the addition of 30 μl SuperaseIn 20 U/μl (Ambion, AM2696). Digestions were then immediately loaded onto a 1 M sucrose cushion, prepared in polysome buffer containing 0.1 U/μl SuperaseIn. Ribosomes were pelleted by centrifugation for 4 hr at 70,000 rpm, 4°C in a TLA-110 rotor. The liquid was removed and the pellet was resuspended in 570 μl 10 mM Tris (pH 7), followed by the immediate addition of 30 μl 20% SDS. The sample was heated to 65°C and RNA was extracted using two rounds of acid phenol/chloroform followed by chloroform alone. RNA was precipitated from the aqueous phase by adding sodium acetate to a final concentration of 300 mM followed by at least one volume of isopropanol. Precipitation was carried out at −30°C for 30 min and RNA was then pelleted by centrifugation for 30 min at 20,000 × g, 4°C. The supernatant was discarded, the pellet was air-dried, and the RNA was resuspended in 150 μl Tris (pH 7). The typical RNA yield was 100 to 200 μg.

Size Selection

RNA was first filtered through a Microcon YM-100 (Millipore) to remove of RNAs greater than 100 nt. Subsequent experiments have shown that this filtration is unnecessary and that samples can be prepared by performing electrophoretic separation directly on resuspended footprinting pellets (our unpublished observations). To perform filtration, a 50 μg RNA aliquot was diluted to a total volume of 500 μl and SuperaseIn was added to a final concentration of 0.1 U/μl. The sample was loaded onto a Microcon YM-100 (Millipore) and 50 U of SuperaseIn was placed in the collection tube. The sample was centrifuged for roughly 30 min at 510 × g, to recover 425 μl filtrate containing only small RNAs. RNA was then precipitated from the filtrate as described above, except that 30 μg GlycoBlue (Ambion AM9515) was added as a coprecipitant.

Filtered RNA was resuspended in 10 μl 10 mM Tris (pH 7). Samples were separated by denaturing polyacrylamide gel electrophoresis (PAGE) in a 15% polyacrylamide gel with urea. Marker oligos oNTI199 and oNTI265 were used to demarcate the 28– 34 nt region, inclusive, that was excised. The gel slices were physically disrupted by centrifugation through a needle hole from an inner 0.5 ml microfuge tube nested in an outer 1.5 ml collection microfuge tube. The gel was extracted in 200 μl RNase-free water for 10 min at 70°C. The eluate was recovered by loading the gel slurry onto a Spin-X column (Corning 8160) and centrifuging to recover the eluate in the collection tube. RNA was then precipitated from the filtered eluate as described above, including the use of a coprecipitant.

mRNA-Seq Fragment Preparation

A 300 μl aliquot of lysate was diluted with 300 μl 10 mM Tris (pH 7) and RNA was extracted as described above. Poly(A)$^+$ mRNA was purified from the total RNA sample using the Oligotex mRNA Mini kit (QIAGEN) according to the manufacturer's instructions. The resulting mRNA was fragmented by partial hydrolysis in a bicarbonate buffer as previously described (Ingolia et al., 2009). The fragmented mRNA was separated by denaturing PAGE as described above and fragments of 50–80 nt length were selected. The fragmented mRNA was recovered from the gel as described above.

Library Generation

RNA samples were dephosphorylated using T4 polynucleotide kinase, which also possesses a 3′-phosphatase activity that is capable of removing 2′,3′-cyclic phosphodiesters. Ribosome footprints and mRNA fragments were resuspended in 25 μl 10 mM Tris (pH 8) and denatured for 2 min at 75°C. Samples were then equilibrated to 37°C and brought to a volume of 50 μl in 1× T4 polynucleotide kinase reaction buffer with 25 U T4 polynucleotide kinase (NEB M0201S) and 12.5 U SuperaseIn. This dephosphorylation reaction was incubated 1 hr at 37°C and heat-inactivated 10 min at 70°C. Dephosphorylated RNA was then purified by precipitation as described above.

Linker attachment was then performed either by linker ligation or by enzymatic polyadenylation. Linker ligation was carried out in a 20 μl reaction consisting of dephosphorylated RNA, 12.5% w/v PEG 8000 (prepared from a 50% w/v stock no more than 1 month old), 10% DMSO, 1x T4 Rnl2(tr) reaction buffer, 20 U SuperaseIn, 500 ng preadenylylated miRNA cloning linker (IDT, Linker #1), and 200 U T4 Rnl2(tr) (NEB, M0242L). The ligation was incubated for 2.5 hr at 37°C. Ligation products were separated by denaturing PAGE as described above, the product bands were excised, and RNA was extracted and precipitated.

Ribosome footprint samples were then treated by subtractive hybridization using biotinylated oligos that were reverse complements of abundant rRNA contaminants observed in preliminary sequencing experiments. RNA was suspended in 30 μl 2× SSC and 250 pmol total biotinylated subtraction oligos were added. The sample was denatured 2 min at 70°C, transferred to 37°C, and 20 U SuperaseIn was added. Hybridization was performed for 30 min at 37°C. Biotinylated oligos were then recovered using MyOne streptavidin C1 DynaBeads (Invitrogen) according to the manufacturer's instructions, using 1 mg magnetic beads per sample. RNA was then precipitated as described above.

Reverse transcription was then carried out on all samples. Reactions were prepared in a 18.0 μl volume using SuperScript III (Invitrogen) according to the manufacturer's instructions, using 50 pmol oNTI225-Link1 primer (for linker ligation samples) or oNTI225 (for polyadenylated samples). Reactions were denatured 5 min at 65°C and then equilibrated at 48°C, at which point 10 U SuperaseIn was added along with DTT and SuperScript III enzyme. Reverse transcription was carried out for 30 min at 48°C. RNA template was destroyed by adding 2.0 μl 1N NaOH and incubating 20 min at 98°C. Reverse transcription products were then purified by denaturing PAGE, incorporating an unextended primer control to facilitate the identification of extended products. These extended products were excised, extracted from the gel, and precipitated.

Reverse transcription products were circularized by carrying out a 20 μl CircLigase (Epicenter) reaction according to the manufacturer's instructions, using the entire extracted sample as a substrate.

Some circularized first-strand cDNA libraries were subject to a second round of subtractive hybridization as described above, except that 60 pmol total biotinylated subtraction oligos were used and no SuperaseIn was added. The subtractive oligos used with cDNA libraries were forward strand sequences derived from abundant rRNA contaminants.

One quarter of the circularized cDNA was used as a template for PCR amplification using Phusion (NEB). Reactions were prepared according to the manufacturer's instructions in a 100 μl volume using oligos oNTI230 and oNTI231. Reactions were split into five aliquots of 16.7 μl and amplified with a 30 s denaturation at 98°C followed by cycles of 10 s denaturation at 98°C, 10 s annealing at 65°C, and 5 s extension at 72°C. Reactions were carried out for 6, 8, 10, 12, and 14 cycles. Reactions were then separated by non-denaturing PAGE on an 8% polyacrylamide gel in 1× TBE. Product bands were excised from reactions selected to achieve reasonable yield without saturation, which manifests as reannealed library fragments that migrate slowly due to their imperfect complementarity. DNA was extracted as described above, except elution was carried out by soaking overnight in STE at 4°C. Extracted DNA was resuspended in 10 μl 10 mM Tris (pH 8) and verified using the High Sensitivity DNA assay on the Agilent Bioanalyzer.

Oligonucleotide Sequences

oNTI199 (RNA): 5′-AUGUACACGGAGUCGACCCGCAACGCGA

oNTI225-Link1 (DNA): 5′-(P) GATCGTCGGACTGTAGAACTCTGAACCTGTCGGTGGTCGCCGTATCATT(Sp18)CACTCA(Sp18)CAAGCAGAAGACGGCATACGAATTGATGGTGCCTACAG

oNTI230 (DNA): 5′-AATGATACGGCGACCACCGA

oNTI231 (DNA): 5′-CAAGCAGAAGACGGCATACGA

oNTI265 (RNA): 5′-AUGUACACGGAGUCGAGCUCAACCCGCAACGCGA

oNTI269 (DNA): (biotin)-TGGCGCCAGAAGCGAGAGCC

oNTI270 (DNA): (biotin)-AGACAGGCGTAGCCCCGGGA

oNTI298 (DNA): (biotin)-GGGGGGATGCGTGCATTTATCAGATCA

oNTI299 (DNA): (biotin)-TTGGTGACTCTAGATAACCTCGGGCCGATCGCACG

oNTI300 (DNA): (biotin)-GAGCCGCCTGGATACCGCAGCTAGGAATAATGGAAT

oNTI301 (DNA): (biotin)-TCGTGGGGGGCCCAAGTCCTTCTGATCGAGGCCC

oNTI303m (DNA): (biotin)-GGGGCCGGGCCGCCCCTCCCACGGCGCG

oNTI304m (DNA): (biotin)-CCCAGTGCGCCCCGGGCGTCGTCGCGCCGTCGGGTCCCGGG

oNTI305 (DNA): (biotin)-TCCGCCGAGGGCGCACCACCGGCCCGTCTCGCC

oNTI306 (DNA): (biotin)-AGGGGCTCTCGCTTCTGGCGCCAAGCGT

oNTI307m (DNA): (biotin)-GAGCCTCGGTTGGCCCCGGATAGCCGGGTCCCCGT

oNTI308 (DNA): (biotin)-TCGCTGCGATCTATTGAAAGTCAGCCCTCGACACA

oNTI309 (DNA): (biotin)-TCCTCCCGGGGCTACGCCTGTCTGAGCGTCGCT

(P) designates 5′ phosphorylation, (Sp18) designates a hexa-ethyleneglycol spacer, and (biotin) designates biotin attached to the 5′ terminus by a C6 spacer.

Footprint Sequence Alignment

Ribosome footprints were aligned as described in Ingolia (2010). The first 26 nucleotides of each read were aligned using Bowtie (Langmead et al., 2009) and this alignment was then extended with reference library sequence followed by library generation linker sequence CTGTAGGCACCATCAATTCGTATGCCGTCTTCTGCTTGAA. The length of the reference sequence extension was chosen to minimize the number of mismatches between the read query sequence and the constructed reference-linker target sequence. The use of a 3′ linker, rather than polyadenylation as described previously, avoided most ambiguity in the length of the reference sequence alignment. Alignments with up to two mismatches between the query and target sequence were accepted, and a small fraction of query sequences with over 255 alignments were suppressed.

Alignments were carried out first against a library of transcripts consisting of the mm9 UCSC Known Genes transcript sequences, downloaded on Dec 9, 2009 combined with sequences derived from the genome using the coordinates of the reconstructed transcripts in (Guttman et al., 2010). Reads with no acceptable alignment to these transcript sequences were then aligned against the mm9 genomic sequence.

Footprint alignments were assigned to a specific A site nucleotide based on the length of the fragment. The offset from the 5′ end of the alignment was 29–30 nt long, +15; 31–33 nt long, +16; 34–35 nt long, +17 (Figures S1A and S1B). Transcript density profiles were constructed by determining the number of sequencing reads whose A site was assigned to each nucleotide position.

Metagene Profiles

Footprint profiles within CDSes were produced by assigning reads to a codon when they mapped to nucleotides at positions −1, 0, and +1 relative to the first nucleotide of the codon. Mean and median footprint levels were computed across CDSes, excluding the first fifteen and the last 5 codons. Nonoverlapping 5 codon windows were tiled across the body of the gene and well-translated genes were selected based on a median value of at least 2 reads per window in a cycloheximide-treated ESC sample (Table S2A). Only the representative splice isoform of each gene was considered.

Metagene analyses were performed by normalizing the footprint profile of each well-translated gene by the average footprint density across the body of the gene, again excluding the first fifteen and the last 5 codons. The scaled profiles of all well-translated genes were then averaged at each position, excluding genes from the average at positions that lay outside the bounds of their transcripts.

Translation and Translational Efficiency Calculations

The translation level of a gene was computed as the number of ribosome footprint reads mapping to the gene’s CDS, divided by the length of the CDS in nucleotides, which we call the “ribosome footprint density” in the gene. When comparing translation measurements between different samples (see Figures 7A, S4B, and S4B and Tables S5A and S5D), these measurements were normalized by dividing by the total number of ribosome footprint reads that align to any region of any annotated mouse transcript, which we call the “normalized ribosome footprint density.” This normalization accounts for differences in the total number of sequencing reads and the extent of rRNA contamination in different samples.

The translational efficiency of a gene was computed as the ratio of the normalized ribosome footprint density to the normalized mRNA-seq read density. The latter is computed from mRNA-seq data just as normalized ribosome footprint density is computed and should reflect an estimate of mRNA abundance.

Quantitation of translation was highly reproducible in our experiments—the comparison presented in Figure 1A represents a biological replication, convolved with any cycloheximide-induced differences in translation. When the total number of footprints involved in a comparison is small, statistical fluctuations will dominate the ratio. When there are sufficient footprints for reliable comparison—in the case of Figure 1A, we require 200 in total—then the ratio reflects underlying biological and experimental variability. The typical inter-replicate difference of 15% indicates that measurements of footprint density are highly reproducible.

Gene Ontology Analysis

The Mann-Whitney U test was used to test for significant differences in the translation, translational efficiency, or change in translation or translational efficiency of genes within a certain GO category relative to the full list of genes analyzed. GO categories with at least 16 genes analyzed were tested, and the threshold for significance was an uncorrected $p < 1.7e\text{-}5$, which corresponds to a Bonferroni-corrected $p < 0.01$. Note that the Mann-Whitney is a nonparametric, rank-based test, so the monotonic transformation between raw values and $\log_2$ values does not affect the result. The median value for genes in the GO category and for genes not in the GO category was also computed and the difference was used as a measurement of the magnitude; median values are not directly used in the Mann-Whitney test.

Ribosome Pausing

The restricted set of genes used for metagene analysis was selected for pausing analysis as well (Table S2A) and each footprint profile was then scaled by the median footprint count on the transcript. Pauses were identified as codons whose ribosome footprint count was at least 25-fold the gene median. In order to avoid artifactually high read density caused by reads mapping to degenerate positions in the genome, pause sites were excluded when they aligned with a 28-mer hypothetical footprint, beginning 15 nt upstream of the pause site, that had a perfect match in a distinct transcript. A metagene analysis was performed on these pauses as described above. The nucleotide and peptide sequence flanking these nondegenerate pause sites was used to generate a peptide sequence motif logo.

Ribosome Depletion Analysis

Metagene analysis was performed on well-translated genes following harringtonine treatment using a 5 codon window average as described above. Profiles were then scaled by the average along the metagene profile between codons 800 and 1000 inclusive. Harringtonine depletion did not extend beyond 750 codons at any time point used in this experiment, so the average read density beyond 800 codons serves as an appropriate normalization for overall sequencing coverage in a sample. The half-recovery point for each scaled profile was then defined as the position, beyond codon 40, where the profile first exceeded 0.5.

Low- and high-expression genes were defined as the lowest and highest quintile of well-translated genes sorted by mean ribosomes per codon in the CDS. Low and high TAI genes were defined as the lowest and highest quartile of well-translated genes sorted by TAI. Only genes of at least 750 codons were used for stratifying genes by length; the shorter subset of genes were less than 1000 codons and the longer subset was more than 1000 codons. Secreted proteins were defined based on SignalP predictions in the Ensembl database.

Initiation Site Prediction

Initiation site score vectors were constructed from footprint profiles in four harringtonine-treated samples: the 90S, 120S, and 150S profiles described above, plus a fourth profile from a sample treated with 0.5 μg/ml harringtonine for 180S in a preliminary experiment. An initiation site scoring vector was constructed for a nucleotide position by first collecting a per-codon footprint profile from the per-nucleotide footprint profile by summing reads at nucleotide positions at -1, +0, and +1 relative to the first nucleotide of a codon. Codons at the following positions, relative to the candidate initiation codon at position 0, were then summed as follows: [[-2, -1], [0], [1], [2], [3, 4], [5, 6, 7], [8, 9, 10], [11, 12, 13]]. This provided eight read counts per sample, which were concatenated to produce a 32-element initiation site profile vector. With the initiation codon defined as position 0, the A site position showing the strongest footprint accumulation was position +1. The training set was constructed using the first and second of every three genes in the set of well-translated genes (Table S1). The nucleotide position of the annotated start codon was used as a positive example and the following nucleotide positions, relative to the start codon, were used as negative examples: [-6, -3, 3, 9, 18, 30, 60, 90, 120, 150]. Positions without at least 18 nucleotides on each side of the initiation site scoring window, which covers nucleotides -7 through +40 relative to the candidate initiation site, were excluded. The vectors were used to train the "svm_learn" program within "svm_light" using a radial basis kernel, $\gamma = 2.4$, error/margin trade-off $C = 2.0$, relative positive example weighting of 4.0 (though there were 10-fold more negative than positive examples in the training corpus), using iterative removal and retraining. The model was then tested on the third of every three genes in the set of well-translated genes (Table S2A), using the same positive and negative nucleotide positions.

Initiation sites were predicted based on a minimum score of 0.75 and at least 50 harringtonine footprint reads overall in the scoring vector (covering 48 nucleotides per profile across four profiles). Contiguous blocks of initiation site nucleotide positions were collected into a single initiation site, whose width is given in the "Harr Peak Width" column of Table S3. The site was then assigned to an initiation codon whose first nucleotide was located within the site. If an AUG codon was present, then the initiation site was assigned to it. If no AUG was present but a near-cognate codon was present, then the site was assigned to it. Sites with no candidate initiation codon were excluded from further analysis because it was not typically possible to predict the reading frame being decoded.

Alternative Isoform Analysis

In order to identify initiation sites affected by alternative splicing, UCSC genomic coordinates and cluster IDs were used to determine whether each initiation site was present in all splice isoforms of a gene, or only found in a subset of isoforms. We identified 1827 initiation sites that were absent in some splice isoforms and either annotated as a uORF in all isoforms containing the site, or as an overlapping uORF in all isoforms containing the site.

To identify differences in 5′ UTR translation of distinct isoforms, ribosome footprint and mRNA-seq reads were mapped to the transcripts and classified as alternative or constitutive by the following method. We first created an index of isoform-specific sequence positions in the transcriptome by mapping each 28 nt window of each UCSC transcript with Bowtie against all other transcripts; matches to other transcripts of the same gene indicated that a window was shared by multiple isoforms of the gene. Twenty-eight nucleotide windows that matched multiple genomic positions, e.g., repetitive sequences, were discarded. The result was an annotation of each transcript position as isoform-specific or shared by multiple isoforms of the same gene. We also classified each position as 5′ UTR, CDS, or 3′ UTR in each isoform. Most positions were shared between multiple isoforms, leaving only a fraction of positions that could distinguish between isoforms. (The motivation of this indexing method was to count the total number of isoform-specific positions, not just those observed in sparse ribosome footprint data, to allow calculations of footprint density, i.e., footprints per base.)

We then took the Bowtie alignments of all ribosome footprint and mRNA-seq reads to the transcriptome and tallied the number of reads falling in each category: isoform A 5′ UTR, isoform A CDS + isoform B CDS, etc.

The 906 genes with alternative uORF initiation sites and sufficient isoform-specific sequence (at least 21 isoform-specific 5′ UTR windows for at least two isoforms of the gene) were analyzed for differences in translation. A χ^2 test was used to test each gene for significant differences between ribosome and mRNA-seq read counts across each isoform-specific 5′ UTR region. That is, we constructed a contingency table with the number of footprint (ribosome) reads and mRNA sequence reads specific to each 5′ UTR isoform.

Two hundred and sixty-three genes had a significant difference in ribosome footprint to mRNA ratios between isoforms, with a false discovery rate of 1%. (χ^2 results for genes with expected counts of 5 or less in any cell of the contingency table were discarded as not having enough data for a reliable comparison. This undercounts the biologically interesting examples by discarding genes with two highly expressed, differentially translated 5′ UTR isoforms and a third isoform that is never observed in this cell type.)

lincRNA Analysis

Profiles of ribosome footprints on lincRNAs were computed using only alignments with a unique genomic origin in order to exclude the misassignment of reads that actually derived from a known protein-coding transcript. The genomic origin was defined by determining the genomic coordinate corresponding to the position of each transcript alignment. The 90 nt window with the most nucleotide positions occupied by ribosome footprints was identified for each lincRNA profile, and the ribosome footprint and mRNA-Seq read density was determined for this window. The translational efficiency of the window was defined as the ratio of ribosome footprint read density to mRNA-Seq read density. As a comparison, the same analysis was performed for windows contained entirely within the codon sequence of the set of well-translated protein-coding genes, and similarly for the 3′ UTRs of these genes.

SUPPLEMENTAL REFERENCES

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Leahy, A., Xiong, J.W., Kuhnert, F., and Stuhlmann, H. (1999). Use of developmental marker genes to define temporal and spatial patterns of differentiation during embryoid body formation. J. Exp. Zool. *284*, 67–81.

Niwa, H., Miyazaki, J., and Smith, A.G. (2000). Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. Nat. Genet. *24*, 372–376.

Niwa, H., Ogawa, K., Shimosato, D., and Adachi, K. (2009). A parallel circuit of LIF signalling pathways maintains pluripotency of mouse ES cells. Nature *460*, 118–122.

Tremml, G., Singer, M., and Malavarca, R. (2008). Culture of mouse embryonic stem cells. In Current Protocols in Stem Cell Biology Chapter 1, Unit 1C 4.

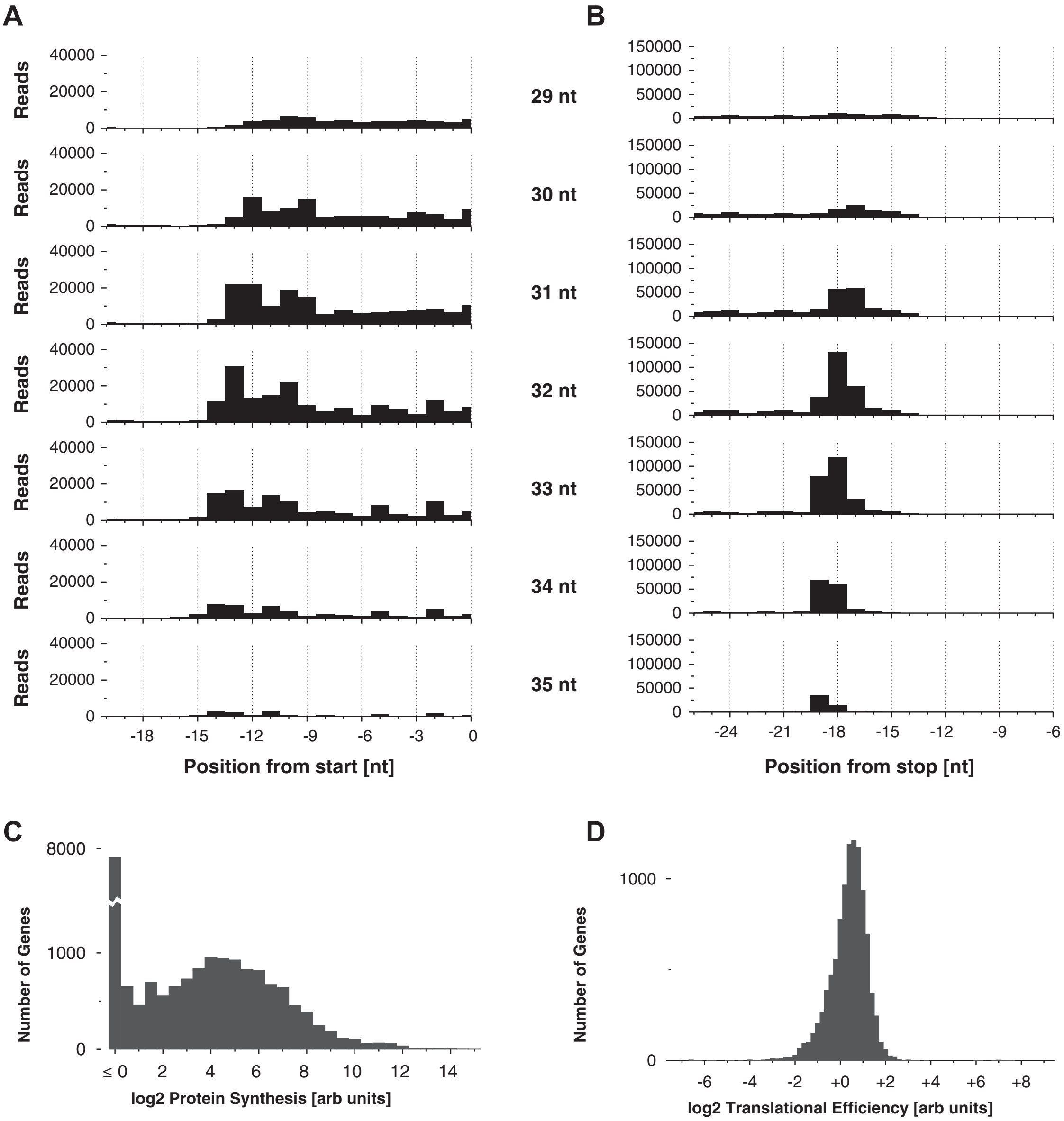

Figure S1. Genome-wide Protein Synthesis in mESCs, Related to Figure 1

(A) Metagene analysis of footprints at the start codon used to calibrate the A site position. Ribosome footprints prepared with no drug treatment were stratified by length, and total footprint count in annotated mouse genes, aligned at their beginning (0 is the first nucleotide of the start codon), shown for the predominant length categories. The A site codon begins at nucleotide +3.

(B) Metagene analysis of footprints at the stop codon. As (A), for genes aligned at the end (0 is the last nucleotide of the stop codon). The A site codon begins at nucleotide −2.

(C) Genome-wide protein synthesis. Total translation for each mouse gene was computed from the ribosome footprint density in the CDS of the canonical isoform (Table S1A). A histogram of log-scaled translation levels is shown.

(D) Genome-wide translational efficiency. Translational efficiency for each mouse gene with detectable mRNA abundance was computed from the ratio of ribosome footprint density to mRNA abundance from mRNA-seq data (Table S1C). A histogram of log-scaled translational efficiencies is shown.

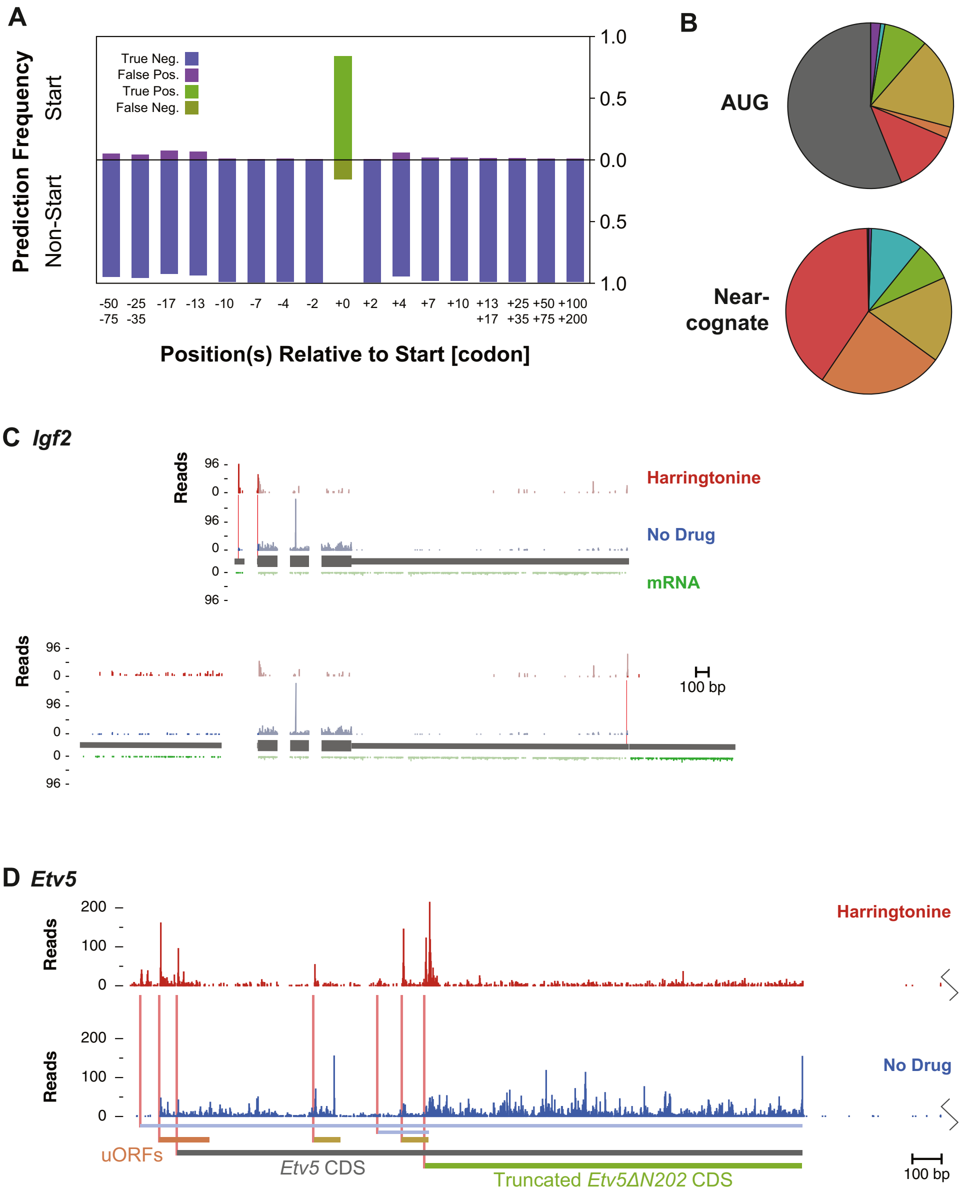

Figure S2. Identification of Translation Initiation Sites and Characterization of Protein Products, Related to Figure 4

(A) Position-specific accuracy of initiation site prediction. As in Figure 4C, separating the non-start-codon test set into specific classes based on their codon position relative to the start codon. Many bars represent data from two positions with similar behavior, grouped together.

(B) Reading frame products by start codon. As in Figure 4F, for initiation sites separated into AUG (top) and near-cognate (bottom).

(C) Patterns of initiation and translation on two transcripts of the *Igf2* gene. The exon structure is shown with thin gray rectangles for the 5′ UTR and thick gray rectangles for the annotated CDS. An mRNA-seq read profile is shown as well on an inverted y axis. Sequencing data for isoform-specific transcript positions are shown in dark colors, and data for non-isoform-specific positions are shown with faint colors.

(D) Pattern of initiation and translation on the *Etv5* transcript. As in Figure 4H. Three alternate AUG reading frames are highlighted, including one that encodes a truncated protein, as well as a prominent short CUG reading frame overlapping the internal AUG initiation site. Two weak non-AUG reading frames are shown in blue.

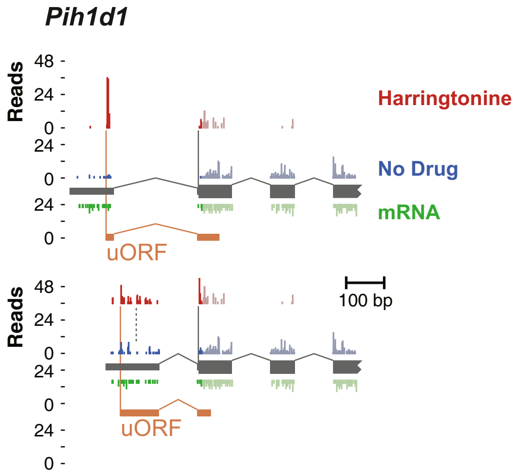

Figure S3. Translation on Alternate *Pih1d1* Transcripts, Related to Figure 6

Patterns of initiation and translation on two *Pih1d1* transcripts. As Figure 6E, for the 5′ ends of the *Pih1d1* transcripts. Ribosome footprints from the annotated start codon overlap the alternative 5′ splice junctions, so it is possible to determine which 5′ UTR is associated with a footprint from an initiating ribosome.

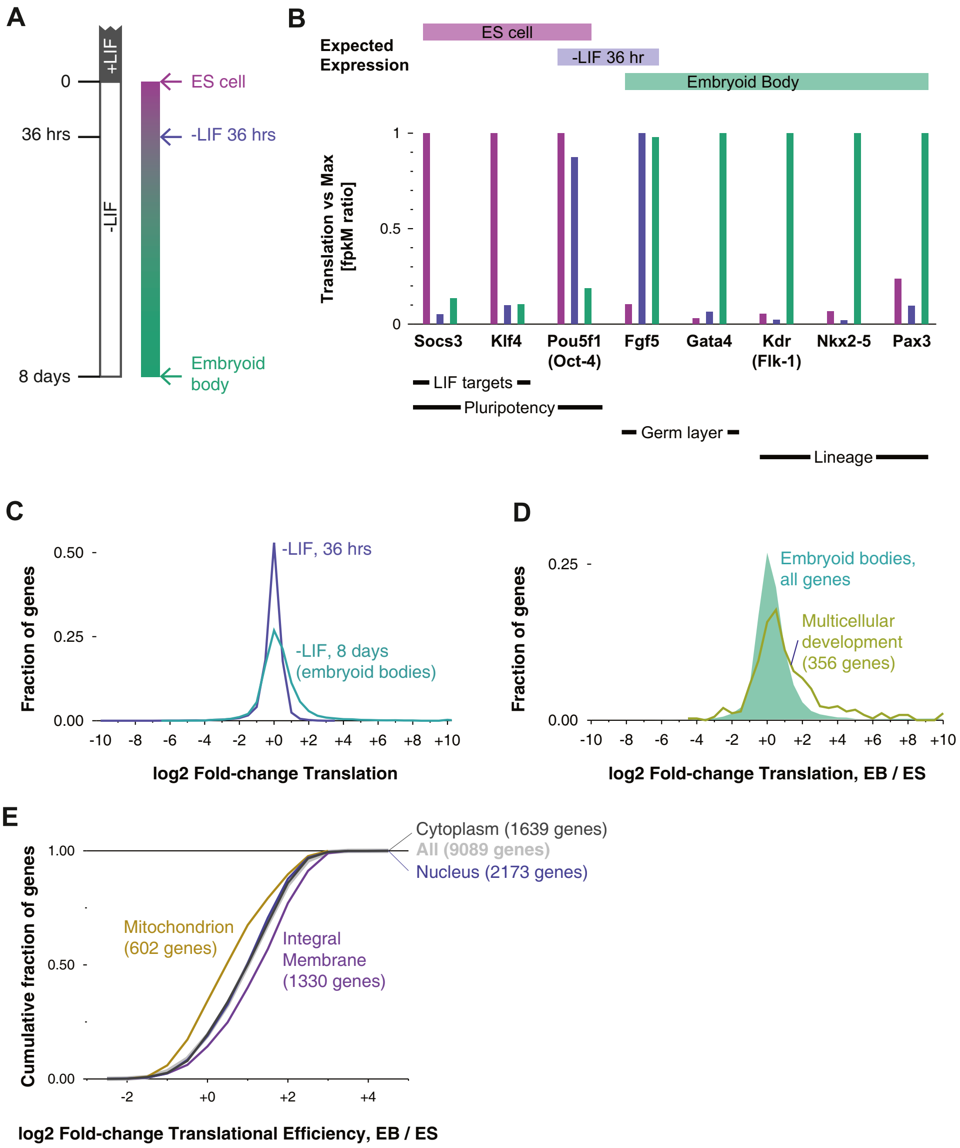

Figure S4. Translation during ESC Differentiation, Related to Figure 7

(A) Schematic of samples during the differentiation timecourse.

(B) Changes in marker gene expression during differentiation. The total translation (sample-normalized ribosome footprint density) for several marker genes was scaled to the maximum translation seen in any sample and the scaled translation levels are plotted. Expected expression periods are shown above (Leahy et al., 1999; Niwa et al., 2000, 2009).

(C) Changes in translation during differentiation. The distribution of $\log_2$ fold-changes of translation (sample-normalized ribosome footprint density) is shown for all genes, 36 hr and 8 days after LIF withdrawal (see Tables S5A and S5D).

(D) Induction of developmental genes in EBs. As (C), showing the distribution for all genes as well as for genes with the "multicellular organismal development" GO annotation.

(E) Translational regulation of localized proteins in embryoid bodies. The cumulative distribution of $\log_2$ fold-changes in translational efficiency is shown for four broad GO localization categories.