Published as: *Cell Rep*. 2014 September 11; 8(5): 1365–1379.

Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes

Nicholas T. Ingolia[1,5,7], **Gloria A. Brar**[2,5], **Noam Stern-Ginossar**[2,6], **Michael S. Harris**[1,3,5], **Gaëlle J. S. Talhouarne**[1,3], **Sarah E. Jackson**[4], **Mark R. Wills**[4], and **Jonathan S. Weissman**[2]

[1]Department of Embryology, Carnegie Institution for Science, Baltimore, MD 21218, USA

[2]Department of Cellular and Molecular Pharmacology, Howard Hughes Medical Institute, Center for RNA Systems Biology, California Institute for Quantitative Biosciences, University of California, San Francisco, CA 94158, USA

[3]Department of Biology, The Johns Hopkins University, Baltimore, MD 21218, USA

[4]Department of Medicine, University of Cambridge, Cambridge CB2 0QQ, UK

SUMMARY

Ribosome profiling suggests that ribosomes occupy many regions of the transcriptome thought to be non-coding, including 5′ UTRs and lncRNAs. Apparent ribosome footprints outside of protein-coding regions raise the possibility of artifacts unrelated to translation, particularly when they occupy multiple, overlapping open reading frames (ORFs). Here we show hallmarks of translation in these footprints: co-purification with the large ribosomal subunit, response to drugs targeting elongation, trinucleotide periodicity, and initiation at early AUGs. We develop a metric for distinguishing between 80S footprints and nonribosomal sources using footprint size distributions, which validates the vast majority of footprints outside of coding regions. We present evidence for polypeptide production beyond annotated genes, including induction of immune responses following human cytomegalovirus (HCMV) infection. Translation is pervasive on cytosolic transcripts outside of conserved reading frames, and direct detection of this expanded universe of translated products enables efforts to understand how cells manage and exploit its consequences.

INTRODUCTION

Identifying the genomic regions that are transcribed and translated is a fundamental step in annotating a genome and understanding its expression. A variety of microarray- and sequencing-based approaches can reveal the mRNA content of the cell (Bertone et al., 2004; Carninci et al., 2005; Wang et al., 2009), but it has proven more challenging to experimentally define translated sequences within the genome or the transcriptome. Historically, protein-coding sequences were discovered by search for long (> 100 codon) open reading frames, which are unlikely to occur in the absence of selection against stop codons. Widespread use of this approach has also been based on the assumption that short

[7]To whom correspondence should be addressed: Tel: (510) 664-7071; ingolia@berkeley.edu.
[5]Present address: Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA.
[6]Present address: Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel.

peptides are unlikely to fold into stable structures and thus perform robust biological functions. Recently, more sophisticated conservation-based metrics, such as PhyloCSF, were developed for the computational identification of sequences that appear to encode proteins over a broad size range (Lin et al., 2008; Lin et al., 2011). However, these approaches focus on identifying regions of the genome experiencing selective pressure to maintain a reading frame encoding a functional protein. The question of which parts of the genome are translated, whether or not the protein product has an adaptive function in the cell, is related but distinct; it can be answered by experimentally finding the locations of ribosomes on mRNAs.

Global profiling of transcription and mRNA abundance has revealed a class of transcripts with no clear protein-coding potential (Bertone et al., 2004; Carninci et al., 2005; Guttman et al., 2009). Many of these RNAs were long RNA polymerase II products, transcribed from genomic regions far from known protein-coding genes and thus were named long non-coding RNAs (lncRNAs).

The discovery of these surprising RNAs in the transcriptome as well as the existence of short upstream open reading frames (uORFS) in 5′ leader regions (often referred to as 5′ untranslated regions (UTRs) (Calvo et al., 2009; Wethmar et al., 2013), highlight the need for comparable direct, experimental maps of translation. While, based on both lack of conservation and the distribution of ribosome protected fragments, there is strong evidence that most lncRNAs do not encode proteins with conserved adaptive cellular roles (Cabili et al., 2011; Chew et al., 2013; Guttman et al., 2013), these computational approaches could miss functional coding sequences, particularly those that are short and/or species-specific (Reinhardt et al., 2013). Furthermore, translation and protein synthesis have impacts beyond the production of stable proteins with discrete molecular functions – polypeptide products from all cellular translation must be degraded, and non-canonical translation products yield unanticipated antigens that may play roles in viral detection or in autoimmunity (Starck et al., 2012). Finally, the process of translation can affect the stability of the template message, by triggering co-translational decay pathways including nonsense-mediated decay (NMD) (Rebbapragada and Lykke-Andersen, 2009). Knowing what transcripts are translated has important implications for the fate of the RNA, the ribosome, and the cell. The ribosome profiling technique provides a unique opportunity to experimentally address this question.

Ribosome profiling is an approach for mapping the exact position of translating ribosomes across the transcriptome by deep sequencing of the mRNA footprints that are occupied by the ribosomes and thereby physically protected from nuclease digestion (Ingolia et al., 2009; Steitz, 1969; Wolin and Walter, 1988). Analysis of these ribosome-protected mRNA fragments yields a quantitative and detailed map of ribosome occupancy that reveal translation in the cell with single nucleotide resolution. Most ribosome footprints fall within known coding sequences, where they showed three-nucleotide periodicity reflecting the triplet nature of the genetic code. However, ribosome profiling data suggested that some predicted non-coding regions of the transcriptome were translated (Ingolia et al., 2011). In some cases, these footprints were organized on single reading frames that closely resembled known coding sequences except for their shorter length (Brar et al., 2012; Stern-Ginossar et al., 2012). In other cases, footprints were not restricted to a single predominant reading

frame based on metrics such as the ribosome release score (RRS) or the disengagement score (DS) (Chew et al., 2013; Guttman et al., 2013). This second group of predicted translated sequences, present on some lncRNAs as well as the 5′ leaders of many mRNAs, can be distinguished both from conserved protein-coding genes, where one single reading frame does predominate, and from the 3′ UTRs of most mRNAs, which are devoid of ribosome footprints (Chew et al., 2013). The high ribosome occupancy on some of these regions, comparable to that on protein-coding genes, suggests a similar stoichiometry of polypeptide production.

The broad implications of pervasive translation and the discrepancy between ribosome profiling and conservation analysis pose an immediate question: do the footprint sequences detected in these profiling experiments indicate the presence of assembled (80S) ribosomes? Here we address this question and present several ways to distinguish true 80S footprints in ribosome profiling data. We first classify protected RNA fragments based on their size distribution, a purely computational analysis that can be applied to existing data and to new profiling data collected without experimental modification. Our analysis discriminates cleanly between true footprints and known sources of contamination. We validate the results from our fragment length classifier with two new lines of experimental evidence, drugs that target the elongating 80S ribosome specifically and affinity purification of the large ribosomal subunit, both of which support the translation of lncRNAs and 5′ UTRs. We also show that footprints on these non-coding sequences demonstrate hallmarks of eukaryotic translation. Finally, we verify the accumulation of protein products from non-canonical translation and demonstrate the potential functional impact of novel HCMV proteins as a source of viral antigens. Our results show that the universe of translated regions extends beyond long conserved regions encoding large, well-conserved proteins.

RESULTS

The characteristic length of ribosome footprints distinguishes them from background RNA fragments

The ribosome physically encloses its mRNA template and protects a characteristic length of this RNA from nuclease digestion (Steitz, 1969; Wolin and Walter, 1988). In ribosome profiling data, the overall size distribution of fragments derived from protein-coding sequences, which should predominantly reflect true ribosome footprints, differs from the lengths of the abundant rRNA contamination found in profiling samples (Ingolia et al., 2009; Ingolia et al., 2011). We reasoned that fragment size could likewise distinguish true ribosome footprints from other, non-ribosomal contaminants, such as RNA regions that are protected by protein complexes or stable RNA secondary structure. The exact length distribution of protected fragments can vary slightly between samples, likely due to differences in digestion conditions (Ingolia et al., 2012). Furthermore, distinct ribosome conformations can lead to significantly different mRNA footprints lengths (Lareau et al., 2014), and the predominant conformation may vary between samples. In order to avoid these confounding effects, we compared the size distributions of fragments derived from non-coding sequences to those on protein-coding genes within a single sample, treated with

translation elongation inhibitors that should capture most ribosomes in a specific state (Lareau et al., 2014; Wolin and Walter, 1988).

We gathered new ribosome profiling data from mouse embryonic stem (ES) cells (mESCs) treated with the translation elongation inhibitor emetine in order to obtain footprints with stronger reading frame bias (Ingolia et al., 2012; Ingolia et al., 2011). Fragment size distributions in this sample clearly distinguished true ribosome footprints, which predominate on coding sequences, from background RNA contained in non-ribosomal ribonucleoprotein (RNP) complexes such as telomerase (Figure 1A). They also separated footprints of the 80S ribosome from fragments of mitochondrial coding sequences that likely reflect footprints of the distinct mitochondrial ribosome (Figure 1B), and non-coding sRNAs that associate with the cytosolic ribosome or its precursors, such as small nucleolar RNAs (snoRNAs) (Figure 1C). By contrast, RNA fragments derived from lncRNAs and from 5′ UTRs showed a size distribution much like that seen on coding sequences (Figure 1D and 1E). This similarity provides evidence that the protected fragments on these two classes of non-coding sequences consist principally of 80S ribosome footprints, and thus that translation occurs outside of annotated protein-coding regions.

Classifying the translation status of individual transcripts and sub-regions

We next adapted our fragment length distribution analysis to distinguish between individual transcripts that show substantial background fragments from those having true 80S footprints. When hundreds or thousands of ribosome footprint sequencing reads are available for a single transcript, their length distribution should converge to match the characteristic ribosome footprint size. We define a fragment length organization similarity score (FLOSS) that measures the magnitude of disagreement between these two distributions, with lower scores reflecting higher similarity (Figure 1F). Thousands of well-expressed protein coding transcripts almost uniformly scored well, and the similarity improved with increasing read counts, as expected (Figure 1G). As with many sequencing-based analyses, this metric is less informative on transcripts with few reads -- an inevitable consequence of sampling error in estimating the fragment length distribution– but we are most interested in the transcripts with many reads, and thus clear FLOSS results.

In order to contrast non-ribosomal background with true ribosome footprints, we needed canonical set of non-translated RNAs to compare with annotated protein-coding sequences. We selected transcripts with well-established molecular functions as RNAs and features likely to suppress their translation, such as an absence of 5′ methylguanosine caps or assembly into stable ribonucleoprotein structures inaccessible to the translational machinery. Many of these transcripts, defined in previous studies as “classical” non-coding RNAs (Guttman et al., 2013), in fact yielded very few protected fragments. We did find several (including telomerase RNA, vault RNA, and RNase P) that we could test, however, and found that each could be distinguished clearly from annotated coding sequences. Likewise, every individual mitochondrially encoded message stood out clearly from nuclear genes. We concluded that this metric discriminates reliably between true 80S ribosome footprints and background RNA fragments on specific transcripts as well as on broad classes of RNAs.

FLOSS analysis revealed that ribosome profiling-derived reads from lncRNAs and 5′ UTRs overwhelmingly reflect true ribosome footprints. Protected fragments on nearly every individual lncRNA showed a FLOSS value very similar to that seen on coding sequences, in contrast to background from classical non-coding RNAs (Figure 1G). Individual 5′ UTRs also grouped very well with coding sequences (Figure 1H).

We formalized this classification by defining a threshold FLOSS value excluding transcripts that differed greatly from annotated protein-coding genes. We set this threshold based on the read counts and FLOSS values for known protein-coding genes using Tukey's method, a widely-accepted non-parametric criterion for extreme outliers (Tukey, 1977). This cutoff eliminated all classical non-coding RNAs with substantial (> 100 reads) expression while retaining almost all annotated mRNAs (99.6%). The perfect specificity and extraordinary sensitivity likely overestimates the true performance of this metric, especially on transcripts that contribute a mixture of true translation and background. Nonetheless, the vast majority of 5′ UTRs (96%) and lncRNAs (90%) were classified with protein-coding genes (Figures S1A and S1B). Not all 5′ UTRs or lncRNAs produced protected RNA fragments in profiling experiments, but when fragments did appear, they generally resembled the ribosome footprints of coding sequences, suggesting true translation in these regions.

We previously reported apparent ribosome occupancy on the abundant and prototypical lncRNA *Malat1*, which is predominantly nuclear, and thus is largely separated from the translational apparatus (Wilusz et al., 2008). This surprising result led us to investigate protected Malat1 RNA fragments more closely (see Figure 1I). We saw a pattern that was highly suggestive of ribosome occupancy near the 5′ end of the transcript, covering the first AUG-initiated reading frame with substantially lower ribosome density after the corresponding in-frame stop codon. We also saw several other sites in *Malat1* that produced abundant protected RNA fragments. While the overall distribution of *Malat1* fragment lengths did not resemble the profile of true ribosome footprints, the first short reading frame did appear to contain 80S ribosomes (Figures 1I and 1J). Similarly, while the full *Malat1* transcript stood out from protein-coding genes by fragment length analysis, the upstream reading frame resembled those of ordinary protein-coding genes. Thus, *Malat1* RNA fragments appear both to contribute non-ribosomal background, like telomerase or RNase P, and also to represent footprints from ribosomes translating its first ORF. As *Malat1* is predominantly nuclear, while the translation occurs in the cytosol, it would be interesting to find the ribosome density and the relative background contribution in the cytoplasmic fraction. MALAT1 is also unusual in that the mature form is not polyadenylated, but the triple helix structure that protects its non-adenylated 3′ end also supports efficient translation (Wilusz et al., 2012); the role of these ribosomes, if any, in the function of *Malat1* remains to be determined.

The non-coding RNA *Gas5* also yielded a complex mixture of translation and background RNA that could be separated by fragment length analysis. *Gas5* is a snoRNA host gene whose introns contain several snoRNAs; there are no long or conserved reading frames in the mature message. Nonetheless, the spliced RNA associates with ribosomes in order to trigger its degradation by NMD (Smith and Steitz, 1998). Fragment length analysis of the primary *Gas5* transcript indicates that it is a source of background RNA in profiling

experiments, corresponding principally to the intronic snoRNAs (Figure 1K). Fragments that mapped to the fully processed *Gas5* transcript, with no remaining snoRNA sequences, resembled 80S footprints on coding sequences (Figures 1K and 1L). They were also concentrated in reading frames near the 5′ end of the transcript, where translation is expected to occur.

Taken together, these analyses show that fragment length analysis can discriminate between true 80S footprints and background RNA reads in ribosome profiling data. Furthermore, this simple metric can be applied to existing profiling data sets as well as incorporated into computational workflows with no change to experimental protocols. It provides strong evidence for the presence of ribosomes based on comparisons with RNAs whose biology is well understood. As this analysis is correlative, however, we performed direct experimental tests to confirm that footprints on non-coding sequences reflected true translation.

Drugs that inhibit translation specifically affect elongating ribosome footprints on coding and non-coding sequences

Diverse translation inhibitors target distinct sites on the ribosome with high affinity and selectivity (McCoy et al., 2011; Schneider-Poetsch et al., 2010). We previously observed that mammalian cells treated with one such drug, cycloheximide, yielded ~1 nt shorter ribosome footprints over the body of open reading frames than those treated with another, emetine (Figure 2A) (Ingolia et al., 2011). Both emetine and cycloheximide target the ribosome specifically, and so the differences observed in mammalian cells between these two drugs should appear only in true footprints of elongating ribosomes.

We set out to use the selectivity of these drugs for the ribosome as an additional test to distinguish true footprints. In aggregate, fragments on lncRNAs and on 5′ UTRs showed a similar, but more modest, length shift to that seen on protein-coding genes -- the cumulative length distribution on both non-coding regions is larger in emetine than in cycloheximide (Figures 2B and 2C). Drug treatment may affect footprints on non-coding RNAs less than those on coding sequences because the translated reading frames on these RNAs are short and thus terminating ribosomes, whose footprints appear to differ slightly from elongating ribosomes (Ingolia et al., 2011), comprise a larger fraction of the total ribosomes. Alternately, a fraction of these footprints may reflect post-termination ribosome footprints, which can accumulate in yeast defective for ribosome recycling factors, and which should not respond to drugs targeting elongation (Guydosh and Green, 2014). Non-ribosomal background fragments do not shift in length between these two elongation inhibitors (Figure 2D).

We gathered new ribosome profiling data from cycloheximide- as well as emetine-treated mESCs and included a small amount of cycloheximide-stabilized yeast polysomes in each sample in order to monitor any differences in the extent of nuclease digestion between samples (Figure 2E). The true ribosome footprints on annotated coding sequences were again shorter from cycloheximide-treated than from emetine-treated cells, though the difference was less pronounced (Figure 2F). The length of footprints on lncRNAs also shifted in response to treatment with elongation inhibitors (Figure 2G), and these length shifts were significant on protein-coding genes ($p < 1e\text{-}4$), 5′ UTRs ($p < 1e\text{-}4$), and on

lncRNAs ($p < 0.01$) (Figure S2). In contrast, the footprints from the yeast ribosomes included as an internal control showed, if anything, a very modest shift in the opposite direction (Figure 2H) that did not rise to the level of significance ($p > 0.05$) (Figure S2), arguing that the reproducible difference between cycloheximide and emetine treated polysomes did not result from differences in nuclease digestion or library generation that affect all RNA fragments in a sample.

Ribosome footprints on classical coding sequences, 5′ UTRs, and lncRNAs co-purify with the large ribosomal subunit

We next sought to verify that footprints seen outside of annotated coding regions co-purified specifically with the ribosome. Ribosome affinity purification would provide strong evidence that footprints on lncRNAs and on 5′ UTRs were bound to the ribosome (Figure 3A). We typically recover ribosomes by sedimentation in an ultracentrifuge, but this purification provides little specificity for ribosomes over other large RNPs. The most prominent classical non-coding RNAs that contribute to background in ribosome profiling experiments are components of non-ribosomal RNPs, such as RNase P, telomerase, and the vault RNP (Figure 1G). We infer that these RNP assemblies both protect RNA fragments from digestion and then sediment with ribosomes, and it seemed possible that some apparent ribosome footprints on lncRNAs actually reflected the incorporation of the lncRNA into a similar RNP complex.

Specific affinity purification of the ribosome would deplete background from these RNPs. The large (60S) subunit joins at the last step in translation initiation and does not associate with mRNA prior to this time, and so any footprint associated with the 60S subunit derives from a ribosome that has completed initiation and begun translation (Aitken and Lorsch, 2012). Ribosome profiling data are unlikely to include footprints of small (40S) subunits scanning 5′ UTRs prior to initiation, because these complexes are unstable in the absence of chemical cross linking and are expected to protect a different mRNA footprint size from assembled 80S ribosomes (Valasek et al., 2007). Nonetheless, we wished to verify that footprints on 5′ UTRs reflected post-initiation assembled (80S) ribosomes.

In order to purify 80S (and 60S) ribosomes specifically, we developed an affinity-tagged version of large subunit ribosomal protein L1 (formerly RPL10A). Several ribosome epitope tags have been developed for lineage-specific polysome isolation, including the translating ribosome affinity purification (TRAP) tag, in which L1 is fused to EGFP (Heiman et al., 2008). We believed that in vivo biotinylation of L1 would offer advantages over epitope tags, allowing us to exploit the high affinity and rapid association of biotin and streptavidin to purify tagged ribosomes. We placed a biotin acceptor peptide at the end of a long, flexible linker at the C-terminus of L1 and co-expressed this tagged protein along with birA, the cognate *E. coli* biotin ligase, in human HEK293 cells. Tagged L1 was biotinylated, dependent on the presence of birA, and L1-biotin was incorporated into ribosomes.

In order to test our enrichment of tagged ribosomes, we mixed lysate from human cells expressing L1-biotin (in addition to their endogenous L1) with a control yeast lysate lacking biotinylated ribosomes and compared the fate of the human ribosome footprints to footprints from yeast genes. We performed nuclease footprinting this mixture, collected all ribosomes

by filtration through Sephacryl S400 columns, and purified the tagged human ribosomes by streptavidin affinity. Footprints from human protein-coding genes were strongly enriched in the streptavidin-bound sample relative to footprints from yeast transcripts (Figure 3B). The only exception was the yeast gene *ACC1*, which encodes the endogenous yeast biotin carrier protein. We assume that it is biotinylated co-translationally in vivo and so footprints recovered by affinity purification through the nascent chain. Consistent with this model, only footprints from the 3′ end of *ACC1*, corresponding to ribosomes that have synthesized the biotin acceptor site of Acc1p, are enriched. Importantly, the observed specificity for human mRNAs also excluded post lysis association of human ribosomes to yeast mRNAs, arguing strongly that footprints seen in ribosome profiling experiments reflects translation that initiated in vivo prior to cell lysis. Fragment length distribution analysis provided further evidence against human ribosomes subject to affinity enrichment on yeast mRNAs, as protected fragments on human and yeast ribosomes are distinct in the mixed lysate and there was no evidence for a shift towards human fragment lengths on yeast messages following affinity purification. Human snoRNA reads also co-purified with biotinylated L1, though somewhat less efficiently than ribosome footprints, as we expect due to their binding to pre-ribosomal complexes in order to guide pre-rRNA modification (Figures S3A–C).

We then investigated the fate of other human-derived background reads following affinity purification of ribosomes. As noted above, profiling data after conventional ribosome sedimentation in HEK cells contained fragments mapping to several classical non-coding RNAs that also appeared in the mESC profiling, such as RNase P. Fragment length analysis using the FLOSS reliably discriminated this background from footprints on coding sequences (Figure 3D). These same transcript fragments were also depleted in affinity-purified profiling samples, at least as strongly as were yeast coding sequences (Figures 3E and 3F). Fragments from mitochondrial coding sequences were also strongly depleted, as the mitochondrial ribosome, which is entirely distinct from the cytosolic ribosome, lacked a biotin tag.

Having established affinity purification as a physical separation of background RNA fragments from true ribosome footprints, we next turned to investigate the status of apparent ribosome footprints in non-coding regions. We first verified that, as in mESCs, the protected fragments size distribution on HEK cell 5′ UTRs closely resembled ribosome footprints from the coding sequences (Figure S3D). These 5′ UTR protected fragments also co-purified with the large ribosomal subunit in nearly all cases (Figure 3C). We thus conclude that these fragments are true 80S ribosome footprints and do not reflect scanning 40S subunits. Likewise, we find that protected fragments on most HEK lncRNAs are physically bound to the ribosome and likely reflect true translation of these non-coding RNAs (Figures 3G–I). Furthermore, the small number of lncRNAs yielding substantial non-ribosome-associated fragments were independently identified as sources of background by the FLOSS analysis.

Translation on lncRNAs occurs in AUG-initiated reading frames near the 5′ end of the transcript

LncRNAs lack a conserved, protein-coding reading frame by definition, and accordingly, ribosome footprints on these transcripts are not organized into a single, discrete reading

frame without downstream translation in the manner seen on mRNAs (Chew et al., 2013; Guttman et al., 2013; Ingolia et al., 2011). Translation on lncRNAs and on mRNAs could differ fundamentally, however, and we wished to determine whether ribosome occupancy on lncRNAs show key features of eukaryotic translation. While translation outside of annotated protein coding regions often initiates at a variety of near cognate codons in overlapping reading frames, obscuring some features of translation that manifest clearly on transcripts encoding a conserved protein, initiation should nonetheless be strongly biased towards AUG codons near the 5′ end of RNAs, and elongating ribosomes should show enrichment in the reading frame that follows until it ends at a stop codon. In order to evaluate the pattern of translation on lncRNA, we analyzed the initiation site profiling data we gathered from mESCs (Ingolia et al., 2011). We previously reported that brief treatment with the drug harringtonine causes ribosomes to accumulate at start codons while allowing run-off depletion of ribosomes over the rest of the coding sequence. This can be used to robustly identify translation initiation sites (Ingolia et al., 2011; Stern-Ginossar et al., 2012). Here, we use a simplified criterion to detect peaks of ribosome occupancy over AUG codons following harringtonine treatment (Figure 4A). This approach is robust against the possibility of concurrent translation of other, overlapping reading frames. While we considered only AUG codons as candidate start sites, we found that on the majority of lncRNAs, the start site we selected was the highest occupancy ribosome position of the entire RNA (Figure 4B), suggesting that this assumption was reasonable.

Initiation sites on lncRNAs detected in harringtonine profiling data showed hallmarks of eukaryotic translation. In the canonical initiation pathway, factors bound to the 5′ cap recruit a pre-initiation complex that scans the RNA directionally to identify a start codon. Consistent with this mechanism of translation, the start sites detected in harringtonine profiling generally fell near the beginning of the lncRNA, within a few hundred nucleotides of the 5′ end (Figure 4C) and at one of the first AUG codons on the transcript (Figure 4D). This bias towards early AUG codons is well explained by the classical model of eukaryotic initiation. By contrast, it is not likely that background RNA fragments not indicative of translation would show a strong preference for AUG codons near the 5′ end of transcripts.

Based on these observations, we next looked for evidence of elongating ribosome footprints in the reading frames associated with these initiation sites. Earlier studies argued against the predominance of a single open reading frame on lncRNAs. Both studies employed similar metrics -- the ribosome release score (RRS) or the disengagement score (DS) -- to demonstrate that the abrupt drop in ribosome occupancy at the end of coding sequences was not seen for short reading frames in 5′ UTRs and on lncRNAs (Chew et al., 2013; Guttman et al., 2013). The absence of clear termination in any single reading frame argues that multiple, overlapping reading frames are translated on these RNAs. Nonetheless, we expected that the start sites we detected should result in elevated ribosome occupancy in the downstream open reading frame relative to the overall transcript. Indeed, we found the observed number of ribosome footprints within predicted reading frames on lncRNAs exceeded the number expected based on the overall ribosome density the length of the reading frame, often 10-fold or more, and never strongly depleted relative to the transcript overall (Figure 4E). This comparison is related to the inside/outside score (IO), the ratio of footprints inside versus outside a candidate reading frame, used by Chew et al. (Chew et al.,

2013). Furthermore, we found that footprints within the open reading frame immediately following the predicted strongest initiation site on a lncRNA showed codon periodicity relative to that start site, similar to the periodicity seen in annotated protein-coding genes, whereas footprints outside of these reading frames do not (Figure 4F). This pattern of footprint occupancy is consistent with substantial in-frame translation from the predicted start site occurring alongside translation of many other reading frames on the transcript including those initiating at near cognate, non AUG sites. This translation, particularly the downstream component that lacks a reading frame signal relative to the strongest AUG start site and thus reflects overlapping translation in alternate reading frames, would reduce RRS and DS metrics on these lncRNAs relative to annotated mRNAs.

Fragment length analysis supports translation on novel reading frames in meiotic yeast

In previous studies, we defined novel translated reading frames in meiotic budding yeast using ribosome profiling data (Brar et al., 2012). We wished to determine whether FLOSS analysis could be applied in this distantly related organism to support our novel annotations. Cycloheximide-stabilized ribosome footprints lying within yeast coding sequences show a tight size distribution, as we observed previously, which could be readily distinguished from background RNA fragments derived from non-translated yeast RNAs, including tRNAs and isolated snoRNAs, and from the validated yeast meiotic non-coding RNAs *IRT1*, *RME2*, and *RME3* (Figure 5A). As in mammals, we also found fragments of mitochondrial mRNAs, likely representing footprints of the mitoribosome, which were larger than cytosolic ribosome footprints. By contrast, the protected fragments on the large majority of new, independent ORFs and on upstream ORFs in the 5′ UTRs of annotated protein-coding genes matched the size of true ribosome footprints closely (Figures 5B to 5D). FLOSS analysis discriminated well between individual annotated coding sequences and non-coding transcripts (Figure 5E), and classified nearly all novel ORFs with known protein-coding genes (Figure 5F). Thus, considered singly or as a group, our reading frame annotations, defined solely by ribosome profiling data, represent the presence of 80S ribosomes and not background RNA fragments.

We also sought to test whether productive translation could be detected from the ribosomes occupying these novel short reading frames. We integrated a GFP reading frame at the 3′ end of meiotically regulated short reading frames in yeast (Figures 5G and 5H). Fusion protein from one short (72 codon) reading frame accumulated in mid-meiotic cells, as predicted from translation data, and localized to mitochondria (Figures 5I and 5J). GFP fused to another, 78 codon reading frame showed robust expression in vegetative cells that decreased in meiosis, consistent with expression profiling data (Figure 5K). The fusion protein colocalized with the nucleus in vegetative cells (Figure 5L). The translational fusion of these short peptides with the large and well-folded GFP may artificially stabilize the protein products and enhance their accumulation. Nonetheless, these data confirm that the novel ORFs defined by ribosome profiling result in the synthesis of proteins, and further, that these short proteins can confer specific localization on a GFP fusion, suggesting that they can display some molecular activity in the cell.

Fragment length analysis supports translation on novel reading frames in human cytomegalovirus

We recently published a new annotation of human cytomegalovirus (HCMV) open reading frames based on ribosome profiling of infected human foreskin fibroblasts (Stern-Ginossar et al., 2012). This annotation included many entirely novel reading frames as well as alternate versions of known proteins. The translation of many of our novel HCMV reading frames was confirmed previously by epitope tagging and by direct detection of native protein products through mass spectrometry (Stern-Ginossar et al., 2012). Our fragment length analysis revealed little difference between human protein-coding genes, well-known viral coding sequences and newly identified ORFs (Figures 6A–6D). We next tested the FLOSS on individual HCMV ORFs and found that nearly all fell among the annotated human protein-coding genes (Figures 6E and 6F).

We may fail to detect proteins from other novel reading frames, despite the fact that they are actually synthesized in the cell, if they are highly unstable and thus low abundance. However, all translated polypeptides can serve as antigens, even if they are rapidly degraded and never accumulate within the cell. In fact, breakdown products from co-translational degradation may be preferentially targeted for display as antigens. The adaptive immune system thus records signatures of past protein expression, and we wanted to mine this record by testing the antigenicity of the novel reading frames we identified in HCMV. We reasoned that if humans with a history of CMV infection displayed T cell responses against novel peptides, as they do against canonical CMV proteins (Sylwester et al., 2005), it would indicate that these peptides were produced in the course of the normal viral life cycle in a human host. Furthermore, the T cell response would directly demonstrate the functional impact of short reading frame translation in viral infection.

We focused on the beta 2.7 transcript in HCMV. Despite its designation as a long noncoding RNA, ribosome profiling data identified eight new, moderately sized ORFs, two of which (ORFL7C and ORFL6C) were identified in lysates from infected cells by mass spectrometry (Stern-Ginossar et al., 2012) (Figure 6G). Human T cells from anonymous HCMV positive donors revealed robust cellular immune responses to ORFL7C and ORFL6C, as well as to other short reading frames on beta 2.7 and other ORFs that we had identified by ribosome profiling (Figures 6H and 6I). These responses were absent from HCMV negative individuals (Figure S2), supporting the natural exposure of HCMV infected individuals specifically to these newly annotated translation products. Neither ORFL6C nor ORFL7C resembled annotated reading frames by the RRS metric, consistent with the polycistronic and overlapping translation on the beta 2.7 transcript (Figure 6G), but the encoded proteins are synthesized in culture models and in infected humans.

DISCUSSION

In this study, we establish the validity of ribosome profiling as the first global and experimental strategy for identifying translated regions of a genome. Profiling data are an excellent complement to computational analyses, which detect conserved protein-coding regions of the genome, and to proteomic approaches for identifying stable proteins. These three techniques answer different but related questions. Conserved functional proteins are a

subset of the total polypeptide content of the cell, which in turn is a subset of all products that are produced, however transiently, by translation. Ribosome profiling thus provides the most expansive view of the proteome, and has thereby helped us appreciate a wider universe of translated sequences.

We present multiple lines of evidence that true ribosome footprints are pervasive on cytosolic RNAs, independent of the presence of conserved reading frames. These footprints change in response to translation inhibitors, co-purify with the large ribosomal subunit, and fall preferentially in reading frames near the 5′ ends of transcripts. The size distribution of ribosome-protected mRNA fragments also distinguishes them from the background present in profiling data. This observation allowed us to develop a fragment length analysis, the FLOSS, that very accurately predicts the results of ribosome affinity purification, which separate true footprints from background RNA by physical rather than computational means. In fact, because some non-coding RNAs do associate with the ribosome for reasons that are unrelated to their actual translation, the FLOSS appears to exclude background more effectively than ribosome pull-down. The large majority of regions identified in profiling experiments reflect true translation; background originates from a handful of known, abundant non-coding RNAs. The FLOSS can be easily incorporated into ribosome profiling workflows and we here provide tools for applying it based on the widely used Bioconductor project (Gentleman et al., 2004). The specific length distribution and FLOSS cutoff for each individual data set can be determined empirically based on annotated protein-coding genes serving as examples of true translation. Adoption of the FLOSS should further increase confidence that profiling measurements on individual transcripts reflect their translation and aid in removing the small number of RNAs that yield non-ribosomal background.

Pervasive ribosome occupancy outside of annotated coding regions has been seen in diverse organisms, and we here present further evidence for the existence of protein products resulting from translation by these ribosomes. The biological implications of this translation remain to be explored, however. In part, it may reflect an imprecision that leads to translation with no functional relevance. We do not know of molecular features that would enable the translational apparatus to distinguish an mRNAs from a capped, polyadenylated, cytosolic lncRNA, and so it may not be surprising to find ribosomes on many lncRNAs. Imperfect rejection of near-AUG codons during translation initiation, combined with the presence of actual AUGs, could explain ribosome occupancy in many 5′ UTRs. However, translation of these non-coding sequences has many potential consequences and non-coding sequences likely experience selection against translation with harmful effects. For example, AUG codons are depleted in many 5′ UTRs, as they interfere with translation of the downstream protein coding sequence, though this interference is exploited as a regulatory mechanism controlling the expression of genes such as Atf4 (Sonenberg and Hinnebusch, 2009). Other side effects of non-coding translation may likewise be avoided in some RNAs and coopted in others.

The translation of an RNA can impact the transcript itself, and lncRNAs with specific molecular functions are likely subject to selective pressure to manage this translation and avoid interference with their other activities. The translating ribosome acts as a potent helicase that can remodel RNA structure and remove RNA-binding proteins, potentially

disrupting functional ribonucleoprotein complexes. We have shown that initiation and translation are biased towards the 5′ ends of lncRNAs, as expected in eukaryotes, and so non-coding cytosolic transcripts may experience selection for benign 5′ reading frames that capture ribosomes and protect functional elements occurring in the 3′ end of the RNA (Ulitsky and Bartel, 2013). Short reading frames with atypical amino acid composition may resemble those found in aberrant mRNAs and trigger RNA decay through NMD or no-go decay, which were originally characterized as mRNA quality control pathways (Perez-Ortin et al., 2013). Translated sequences may also exert cis-acting effects through the peptides they encode, for example by co-translational recruitment of the nascent chain, attached to the ribosome and the transcript, to specific structures in the cell (Yanagitani et al., 2009).

Translation results in the synthesis of a polypeptide regardless of whether an RNA sequence encodes a functional protein constrained by selection, and we have now detected proteins synthesized from novel translated sequences predicted by ribosome profiling in yeast and given evidence for their presence in humans during CMV infection. These unconstrained peptide sequences may not adopt a specific fold and may occupy co-translational folding or degradation machinery, and those peptides escaping surveillance may aggregate and contribute to the burden of unfolded proteins. Some subset of this large pool of newly identified short peptides may play cellular roles that we have yet to discover, akin to the important roles recently shown for the 11 and 32 amino acid peptides synthesized from the *polished rice* and *sarcolambin* loci in Drosophila and the 58 amino acid peptide encoded by the zebrafish *toddler* gene (Kondo et al., 2010; Magny et al., 2013; Pauli et al., 2014).

All RNA sequences subject to translation will experience selection against encoding proteins with detrimental impact on the cell or on the organism. These benign proteins may occasionally provide an adaptive molecular function; for example, a surprisingly large fraction (~20%) of random nucleotide sequences encode functional secretion signals (Kaiser et al., 1987). Further evolution may refine their expression, folding, and activity, ultimately giving rise to the birth of a new gene (Carvunis et al., 2012; Reinhardt et al., 2013).

Regardless of their original cellular role, degraded proteins are the substrates for antigens presented to the cellular immune system, and proteins synthesized by non-canonical translation may be shunted preferentially for degradation and presentation as antigens, expanding the range of epitopes displayed by virus-infected or transformed (Yewdell, 2011). The apparent elevation of non-canonical translation in stress could aid the body in detecting these pathological cells, and differences in translation between normal and transformed cells could yield cancer-specific antigens for immunomodulatory therapy (Mellman et al., 2011). The same processes producing cryptic viral and tumor antigens could also expose cryptic self-antigens that could initiate or sustain an autoimmune response.

In summary, translation of non-coding RNA has the potential to impact the cell directly and to constrain the evolution of genomic sequences. A better understanding of these molecular and evolutionary implications relies, first, on a reliable means for unbiased detection of translation. Ribosome profiling provides a starting point for exploring the role of the translational apparatus in truly non-coding RNAs as well as revealing novel short, functional proteins and offering a window into the murky gradations in between.

EXPERIMENTAL PROCEDURES

Ribosome Footprinting

E14 mESCs and were pretreated with cycloheximide (100 μg/ml) or emetine (50 μg/ml) for 1 minute as indicated, followed by detergent lysis and ribosome footprinting by RNase I digestion (Ingolia et al., 2012). Deep sequencing libraries were generated from 26 – 34 nt footprint fragments (Ingolia et al., 2012) and sequenced on an Illumina HiSeq.

Ribosome Affinity Purification

The ribosome affinity tag construct comprised human ribosomal protein L1 fused to the biotin acceptor peptide (Beckett et al., 1999; de Boer et al., 2003), co-expressed with a biotin ligase using a 2A peptide (de Felipe et al., 2006), as a stable transgene in HEK293 cells using the Flp-In system (Invitrogen). Yeast lysates were prepared as described (Ingolia, 2010). Following nuclease digestion, lysates were loaded onto a Sepharcryl S-400 gel filtration spin column (Boca Scientific) and the flow-through was collected. One aliquot of flow-through was bound to streptavidin-coated magnetic beads (Invitrogen) and RNA was recovered by Trizol extraction directly from beads; another aliquot was used directly for library generation following Trizol extraction. Extracted RNA was converted into deep sequencing libraries.

Footprint Sequence Alignment

Footprint sequences were trimmed to remove 3′ adapter sequence and aligned using TopHat v2.0.7 (Kim et al., 2013) with Bowtie v0.12.9.0 and samtools v0.1.18.0. The composite reference genomes comprised either the mm10 mouse genome with Ensembl GRCm38.72 transcripts or the human hg19 genome with Gencode v17 transcripts (Harrow et al., 2012), supplemented with the yeast genome with *de novo* transcript annotations (Brar et al., 2012). Alignments were filtered to remove those containing more than one mismatch.

Footprint Sequence Data Analysis

Footprints were assigned to specific A site nucleotide positions ~15 bases from their 5′ ends, depending on the exact fragment length, as described previously. Reads assigned between 15 nucleotides before the start codon and 45 nucleotides after the start codon were excluded, as were all reads falling after the position 15 nucleotides upstream of the stop codon. All footprint data analysis was implemented in R/Bioconductor and is provided in a format allowing the direct reproduction of the analyses presented here.

We used our previously published (Stern-Ginossar et al., 2012), simplified approach to detect sites of AUG-mediated initiation in harringtonine-treated mESCs. We identified all AUG codons and selected harringtonine peaks by finding codons where A site occupancy on the +1 codon (i.e., AUG in the P site) as greater than occupancy on the +2 codon, *and* greater than the sum of occupancy on the −1 and the 0 codon, in *both* replicates. Among these AUG harringtonine peaks, we then selected the highest footprint occupancy on the +1 codon.

We computed the footprint A site occupancy at all codons on the transcript (not restricted to AUG codons with a harringtonine peak) and found the rank of the candidate initiation site relative to all other codons.

We also indexed all AUG codons on the transcript, starting from the 5′ end, and found the the candidate initiation site among all AUG codons on the transcript.

Fragment Length Organization Similarity Score

The fragment length organization similarity score (FLOSS) was computed from a histogram of read lengths for footprints on a transcript or reading frame. A reference histogram was produced using raw counts on all annotated nuclear protein-coding transcript, excluding those whose gene overlapped a gene annotated as non-coding. The FLOSS was defined as

$$0.5 \times \sum_{l=26}^{34} \|f(l) - f_{\mathrm{ref}}(l)\|$$

where *f(l)* is the fraction of reads at length *l* in the transcript histogram and $f_{ref}(l)$ is the corresponding fraction in the reference histogram. The FLOSS cutoff score, as a function of the total number of reads, was counted from a rolling window of individual annotated genes and the computing the upper extreme outlier cutoff for each window.

Yeast Western blotting and microscopy

Novel ORFs were tagged with C-terminal GFP fusions by the Pringle method (Longtine et al, Yeast 1998). Samples were collected by TCA precipitation and subjected to Western blotting (mouse anti-GFP antibody, Roche; rabbit anti-hexokinase antibody, Rockland antibodies). Samples were also collected for microscopy, which was performed on a Zeiss Axiophot. Samples were costained with either DAPI or Mitotracker Orange (Molecular Probes).

T Cell Response Assays

Tiling peptides (15 amino acids long with 10 amino acid overlap) for novel CMV ORFs were obtained from JPT Peptide Technologies and pooled at 2 μg/ml of each individual peptide in RPMI 1640. PBMCs were isolated by Lymphoprep (Axis-Shield, Norway) and depleted of either CD4+ or CD8+ T cells by magnetic activated cell sorting (MACS; Miltenyi, U.K.), yielding no more than 0.8% residual cells as accessed by flow cytometry. ELISPOT plates (EBioscience) were prepared, coated and blocked, and T cells were plated at 3.0×10^5 cells in 100 μl RPMI-10.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank H. Chang, J. Darnell, R. Darnell, L. Lareau, and members of the Ingolia and Weissman labs for valuable comments; C. Jan and C. Williams for insights into ribosome affinity purification; and A. Pinder for sequencing. This work was supported by the Searle Scholars Program (N.T.I.), the Howard Hughes Medical Institute (J.S.W.) and by a Human Frontiers in Science post-doctoral fellowship (N. S.-G.) and American Cancer Society Postdoctoral Fellowship 117945-PF-09-136-01-RMC (G.A.B.). MW and SJ were funded by British Medical Research Council Grant G0701279 and the Cambridge BRC.

References

Aitken CE, Lorsch JR. A mechanistic overview of translation initiation in eukaryotes. Nat Struct Mol Biol. 2012; 19:568–576. [PubMed: 22664984]

Beckett D, Kovaleva E, Schatz PJ. A minimal peptide substrate in biotin holoenzyme synthetase-catalyzed biotinylation. Protein Sci. 1999; 8:921–929. [PubMed: 10211839]

Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, et al. Global identification of human transcribed sequences with genome tiling arrays. Science. 2004; 306:2242–2246. [PubMed: 15539566]

Brar GA, Yassour M, Friedman N, Regev A, Ingolia NT, Weissman JS. High-resolution view of the yeast meiotic program revealed by ribosome profiling. Science. 2012; 335:552–557. [PubMed: 22194413]

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2011; 25:1915–1927. [PubMed: 21890647]

Calvo SE, Pagliarini DJ, Mootha VK. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc Natl Acad Sci U S A. 2009; 106:7507–7512. [PubMed: 19372376]

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al. The transcriptional landscape of the mammalian genome. Science. 2005; 309:1559–1563. [PubMed: 16141072]

Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, et al. Proto-genes and de novo gene birth. Nature. 2012; 487:370–374. [PubMed: 22722833]

Chew GL, Pauli A, Rinn JL, Regev A, Schier AF, Valen E. Ribosome profiling reveals resemblance between long non-coding RNAs and 5′ leaders of coding RNAs. Development. 2013; 140:2828–2834. [PubMed: 23698349]

de Boer E, Rodriguez P, Bonte E, Krijgsveld J, Katsantoni E, Heck A, Grosveld F, Strouboulis J. Efficient biotinylation and single-step purification of tagged transcription factors in mammalian cells and transgenic mice. Proc Natl Acad Sci U S A. 2003; 100:7480–7485. [PubMed: 12802011]

de Felipe P, Luke GA, Hughes LE, Gani D, Halpin C, Ryan MD. E unum pluribus: multiple proteins from a self-processing polyprotein. Trends Biotechnol. 2006; 24:68–75. [PubMed: 16380176]

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004; 5:R80. [PubMed: 15461798]

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009; 458:223–227. [PubMed: 19182780]

Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell. 2013; 154:240–251. [PubMed: 23810193]

Guydosh NR, Green R. Dom34 rescues ribosomes in 3′ untranslated regions. Cell. 2014; 156:950–962. [PubMed: 24581494]

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012; 22:1760–1774. [PubMed: 22955987]

Heiman M, Schaefer A, Gong S, Peterson JD, Day M, Ramsey KE, Suarez-Farinas M, Schwarz C, Stephan DA, Surmeier DJ, et al. A translational profiling approach for the molecular characterization of CNS cell types. Cell. 2008; 135:738–748. [PubMed: 19013281]

Ingolia NT. Genome-wide translational profiling by ribosome footprinting. Methods in enzymology. 2010; 470:119–142. [PubMed: 20946809]

Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat Protoc. 2012; 7:1534–1550. [PubMed: 22836135]

Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009; 324:218–223. [PubMed: 19213877]

Ingolia NT, Lareau LF, Weissman JS. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell. 2011; 147:789–802. [PubMed: 22056041]

Kaiser CA, Preuss D, Grisafi P, Botstein D. Many random sequences functionally replace the secretion signal sequence of yeast invertase. Science. 1987; 235:312–317. [PubMed: 3541205]

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013; 14:R36. [PubMed: 23618408]

Kondo T, Plaza S, Zanet J, Benrabah E, Valenti P, Hashimoto Y, Kobayashi S, Payre F, Kageyama Y. Small peptides switch the transcriptional activity of Shavenba by during Drosophila embryogenesis. Science. 2010; 329:336–339. [PubMed: 20647469]

Lareau LF, Hite DH, Hogan GJ, Brown PO. Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. eLife. 2014; 3:e01257. [PubMed: 24842990]

Lin MF, Deoras AN, Rasmussen MD, Kellis M. Performance and scalability of discriminative metrics for comparative gene identification in 12 Drosophila genomes. PLoS Comput Biol. 2008; 4:e1000067. [PubMed: 18421375]

Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics. 2011; 27:i275–282. [PubMed: 21685081]

Magny EG, Pueyo JI, Pearl FM, Cespedes MA, Niven JE, Bishop SA, Couso JP. Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. Science. 2013; 341:1116–1120. [PubMed: 23970561]

McCoy LS, Xie Y, Tor Y. Antibiotics that target protein synthesis. Wiley Interdiscip Rev RNA. 2011; 2:209–232. [PubMed: 21957007]

Mellman I, Coukos G, Dranoff G. Cancer immunotherapy comes of age. Nature. 2011; 480:480–489. [PubMed: 22193102]

Pauli A, Norris ML, Valen E, Chew GL, Gagnon JA, Zimmerman S, Mitchell A, Ma J, Dubrulle J, Reyon D, et al. Toddler: an embryonic signal that promotes cell movement via Apelin receptors. Science. 2014; 343:1248636. [PubMed: 24407481]

Perez-Ortin JE, Alepuz P, Chavez S, Choder M. Eukaryotic mRNA decay: methodologies, pathways, and links to other stages of gene expression. J Mol Biol. 2013; 425:3750–3775. [PubMed: 23467123]

Rebbapragada I, Lykke-Andersen J. Execution of nonsense-mediated mRNA decay: what defines a substrate? Curr Opin Cell Biol. 2009; 21:394–402. [PubMed: 19359157]

Reinhardt JA, Wanjiru BM, Brant AT, Saelao P, Begun DJ, Jones CD. De Novo ORFs in Drosophila Are Important to Organismal Fitness and Evolved Rapidly from Previously Non-coding Sequences. PLoS Genet. 2013; 9:e1003860. [PubMed: 24146629]

Schneider-Poetsch T, Ju J, Eyler DE, Dang Y, Bhat S, Merrick WC, Green R, Shen B, Liu JO. Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin. Nat Chem Biol. 2010; 6:209–217. [PubMed: 20118940]

Smith CM, Steitz JA. Classification of gas5 as a multi-small-nucleolar-RNA (snoRNA) host gene and a member of the 5′-terminal oligopyrimidine gene family reveals common features of snoRNA host genes. Mol Cell Biol. 1998; 18:6897–6909. [PubMed: 9819378]

Sonenberg N, Hinnebusch AG. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. Cell. 2009; 136:731–745. [PubMed: 19239892]

Starck SR, Jiang V, Pavon-Eternod M, Prasad S, McCarthy B, Pan T, Shastri N. Leucine-tRNA initiates at CUG start codons for protein synthesis and presentation by MHC class I. Science. 2012; 336:1719–1723. [PubMed: 22745432]

Steitz JA. Polypeptide chain initiation: nucleotide sequences of the three ribosomal binding sites in bacteriophage R17 RNA. Nature. 1969; 224:957–964. [PubMed: 5360547]

Stern-Ginossar N, Weisburd B, Michalski A, Le VT, Hein MY, Huang SX, Ma M, Shen B, Qian SB, Hengel H, et al. Decoding human cytomegalovirus. Science. 2012; 338:1088–1093. [PubMed: 23180859]

Sylwester AW, Mitchell BL, Edgar JB, Taormina C, Pelte C, Ruchti F, Sleath PR, Grabstein KH, Hosken NA, Kern F, et al. Broadly targeted human cytomegalovirus-specific CD4+ and CD8+ T cells dominate the memory compartments of exposed subjects. J Exp Med. 2005; 202:673–685. [PubMed: 16147978]

Tukey, JW. Exploratory Data Analysis. Reading, MA: Addison-Wesley; 1977.

Ulitsky I, Bartel DP. lincRNAs: genomics, evolution, and mechanisms. Cell. 2013; 154:26–46. [PubMed: 23827673]

Valasek L, Szamecz B, Hinnebusch AG, Nielsen KH. In vivo stabilization of preinitiation complexes by formaldehyde cross-linking. Methods Enzymol. 2007; 429:163–183. [PubMed: 17913623]

Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009; 10:57–63. [PubMed: 19015660]

Wethmar K, Barbosa-Silva A, Andrade-Navarro MA, Leutz A. uORFdb--a comprehensive literature database on eukaryotic uORF biology. Nucleic Acids Res. 2013

Wilusz JE, Freier SM, Spector DL. 3′ end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. Cell. 2008; 135:919–932. [PubMed: 19041754]

Wilusz JE, JnBaptiste CK, Lu LY, Kuhn CD, Joshua-Tor L, Sharp PA. A triple helix stabilizes the 3′ ends of long noncoding RNAs that lack poly(A) tails. Genes Dev. 2012; 26:2392–2407. [PubMed: 23073843]

Wolin SL, Walter P. Ribosome pausing and stacking during translation of a eukaryotic mRNA. Embo J. 1988; 7:3559–3569. [PubMed: 2850168]

Yanagitani K, Imagawa Y, Iwawaki T, Hosoda A, Saito M, Kimata Y, Kohno K. Cotranslational targeting of XBP1 protein to the membrane promotes cytoplasmic splicing of its own mRNA. Mol Cell. 2009; 34:191–200. [PubMed: 19394296]

Yewdell JW. DRiPs solidify: progress in understanding endogenous MHC class I antigen processing. Trends Immunol. 2011; 32:548–558. [PubMed: 21962745]

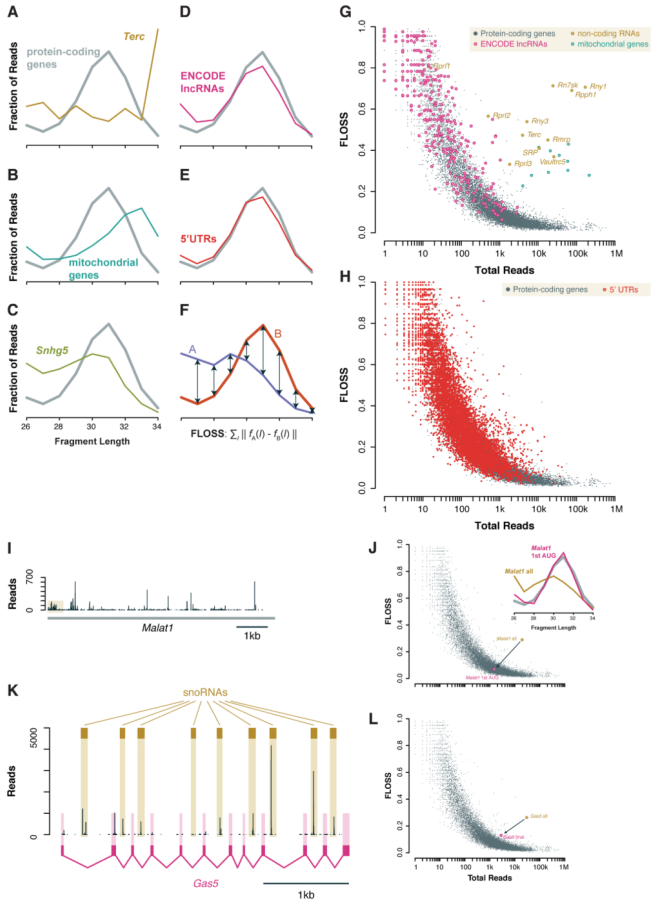

Figure 1. Fragment length analysis distinguishes true ribosome footprints on coding and non-coding sequences

(A – E) Distribution of fragment lengths mapping to nuclear coding sequences (CDSes) compared to (A) the telomerase RNA *Terc*, (B) mitochondrial coding sequences, (C) snoRNA host gene *Snhg5*, (D) ENCODE lncRNAs, and (E) 5′ UTRs of protein-coding genes, in ribosome profiling data from emetine-treated mESCs. (F) Metric comparing the similarity of two length distributions. (G) Fragment length analysis plot of total reads per transcript and FLOSS relative to the nuclear coding sequence average. An FLOSS cutoff is based on an extreme outlier threshold for annotated coding sequences. LncRNAs resemble annotated, nuclear protein-coding genes, whereas functional RNAs and mitochondrial coding sequences are distinct. (H) As (G), comparing 5′ UTRs and coding sequences of nuclear-encoded mRNAs. (I) Read count profile on *Malat1* with an inset showing ribosomes on a non-AUG uORF and the first reading frame at the 5′ end of the transcript. An inset shows the fragment length distribution for the first reading frame, which matches the overall coding sequence average, and the whole transcript, which does not. (J) Fragment length analysis showing the shift from the entire *Malat1* transcript, which contains substantial background, to the first *Malat1* reading frame, which contains true ribosome footprints. (K) Read count profile acros the primary *Gas5* transcript with the snoRNAs and the fully-spliced transcript shown. (L) As (J) for the primary *Gas5* transcript, containing snoRNA precursors, and the fully spliced product.

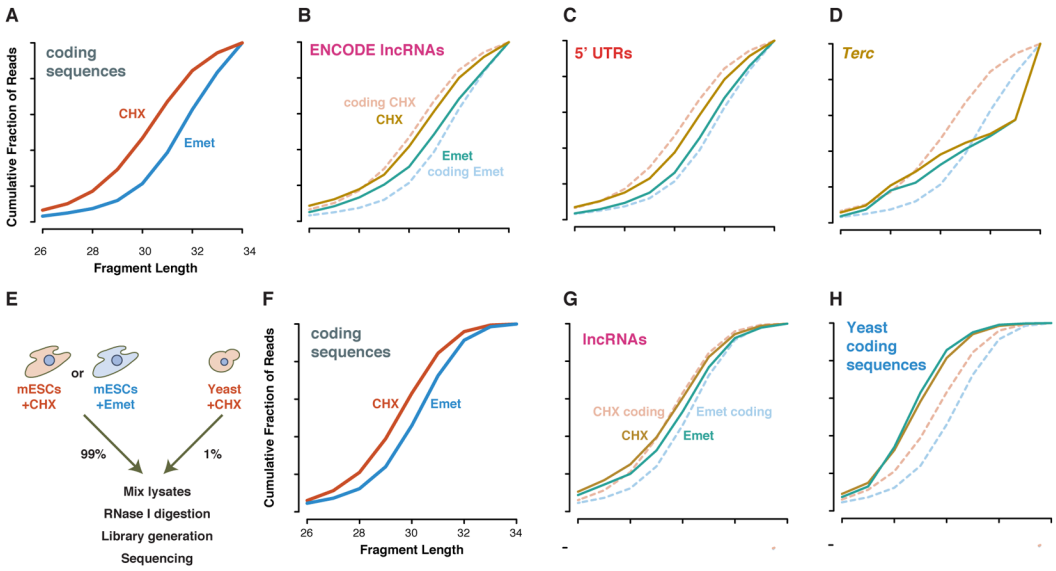

Figure 2. Elongation inhibitors shift ribosome footprint sizes

(A) Cumulative length distribution shows ~1 nt larger footprints on annotated coding sequences from emetine-versus cycloheximide-treated cells (Ingolia et al., 2011). (B) LncRNA and (C) 5′ UTR footprints from transcripts passing the FLOSS cutoff show a similar length shift, whereas background from (D) classical non-coding RNAs do not. (E) Experimental design with e cycloheximide-treated yeast polysomes as an internal standard for nuclease digestion and library generation. (F) Annotated coding seuqences and (G) lncRNAs again show larger footprints in emetine-treated cells. (H) Cycloheximide-stabilizied footprints are not larger in the emetine-treated mESC lysate sample.

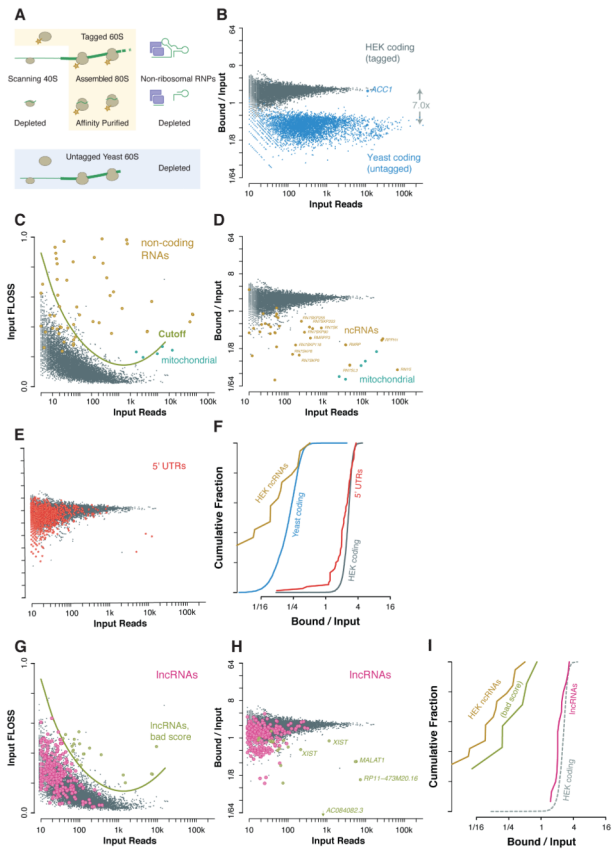

Figure 3. Ribosome affinity purification separates 80S footprints from background RNA

(A) Schematic showing that affinity purification of tagged 60S ribosome subunits recovers 80S footprints but depletes background from non-ribosomal RNPs, potential scanning 40S footprints, and footprints of untagged yeast 80S ribosomes are also depleted. (B) Human ribosome footprints are retained during ribosome affinity purification while yeast ribosome footprints (excepting the yeast biotin carrier *ACC1*) are depleted. (C) Fragment length analysis of nuclear and mitochondrial coding sequences and of functional non-coding RNAs in HEK cells. A fragment length score cutoff based on extreme outliers relative to coding sequences excludes background fragments. (D) Ribosome footprints are retained during ribosome affinity purification while mitochondrial footprints and non-coding RNAs are depleted. (E, F) Ribosome footprints on 5′ UTRs are retained during affinity purification of the 60S ribosomal subunit. (G) Fragment length analysis of ENCODE lncRNAs, identifying a small number of transcripts with likely non-ribosomal contamination. (H, I) Ribosome footprints on lncRNAs are retained during ribosome affinity purification, whereas many sources of non-ribosomal contamination, including the nuclear non-coding RNA *XIST*, are depleted.

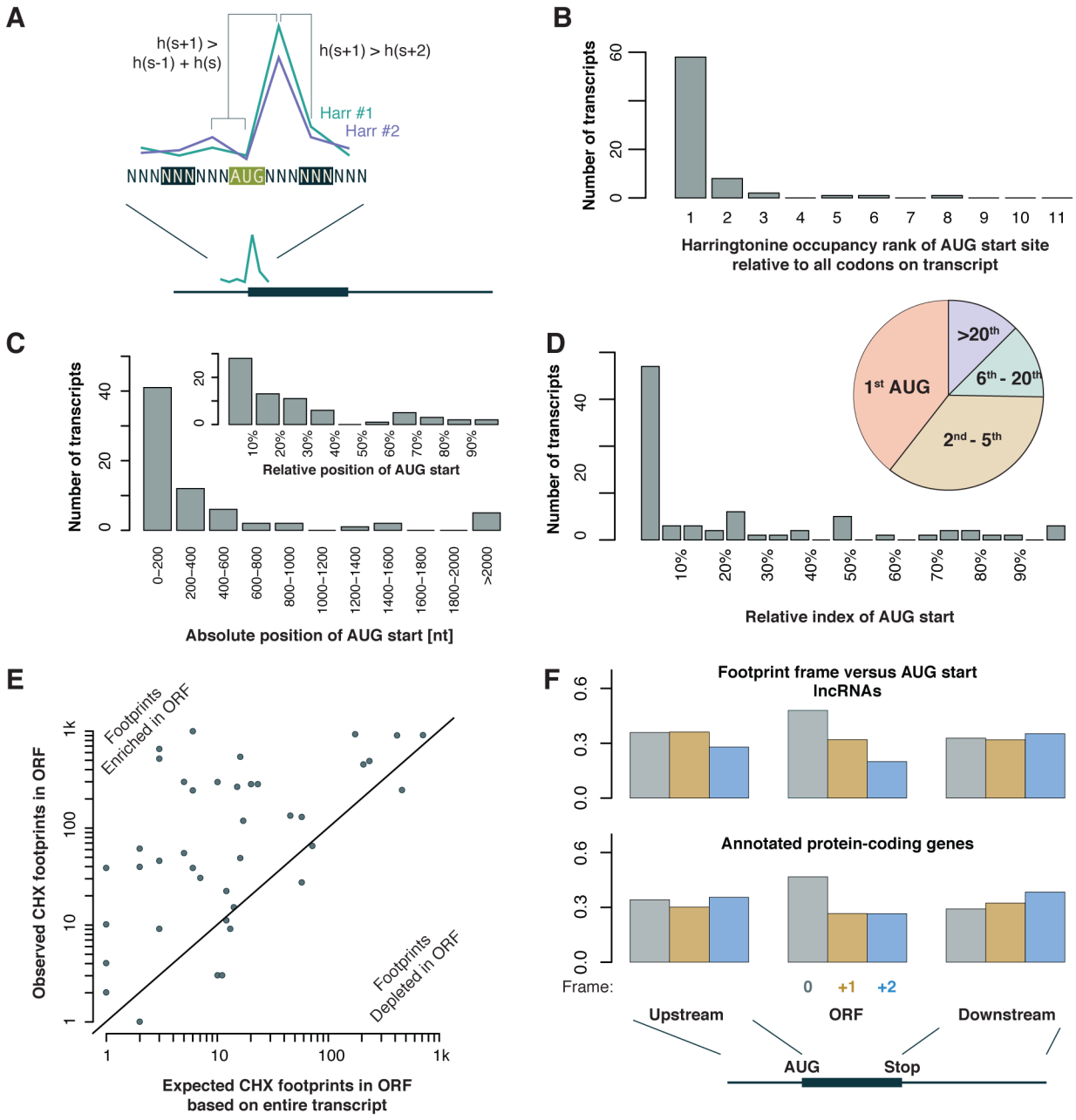

Figure 4. Ribosomes translate detectable reading frames on lncRNAs

(A) Schematic of AUG start site detection using two harringtonine samples (from 120 s and 150 s treatment). The start site is an AUG codon with a peak in footprint density – higher occupancy than flanking codons – selected as the highest occupancy among peaks at AUGs. (B) AUG start sites typically show the highest footprint density among all codons, not just all AUGs with peaks. (C) AUG start sites typically fall in the first few hundred nt of transcripts, and (inset) near the beginning of the transcript. (D) AUG start sites are typically among the first AUG codons on transcripts, with relative positions shown in the histogram and absolute index shown in the pie chart (i.e., nearly half of AUG start sites are the first AUG on the transcript overall). (E) Overall ribosome occupancy is higher in the ORFs

downstream of AUG start sites, relative to the overall density on the transcript. (F) Footprints on As downstream of detected AUG start sites and upstream of the stop codon are biased towards the frame of the ORF. Annotated protein-coding genes show similar reading frame bias within the ORF but not in the 5′ UTR (upstream) or 3′ UTR (downstream).

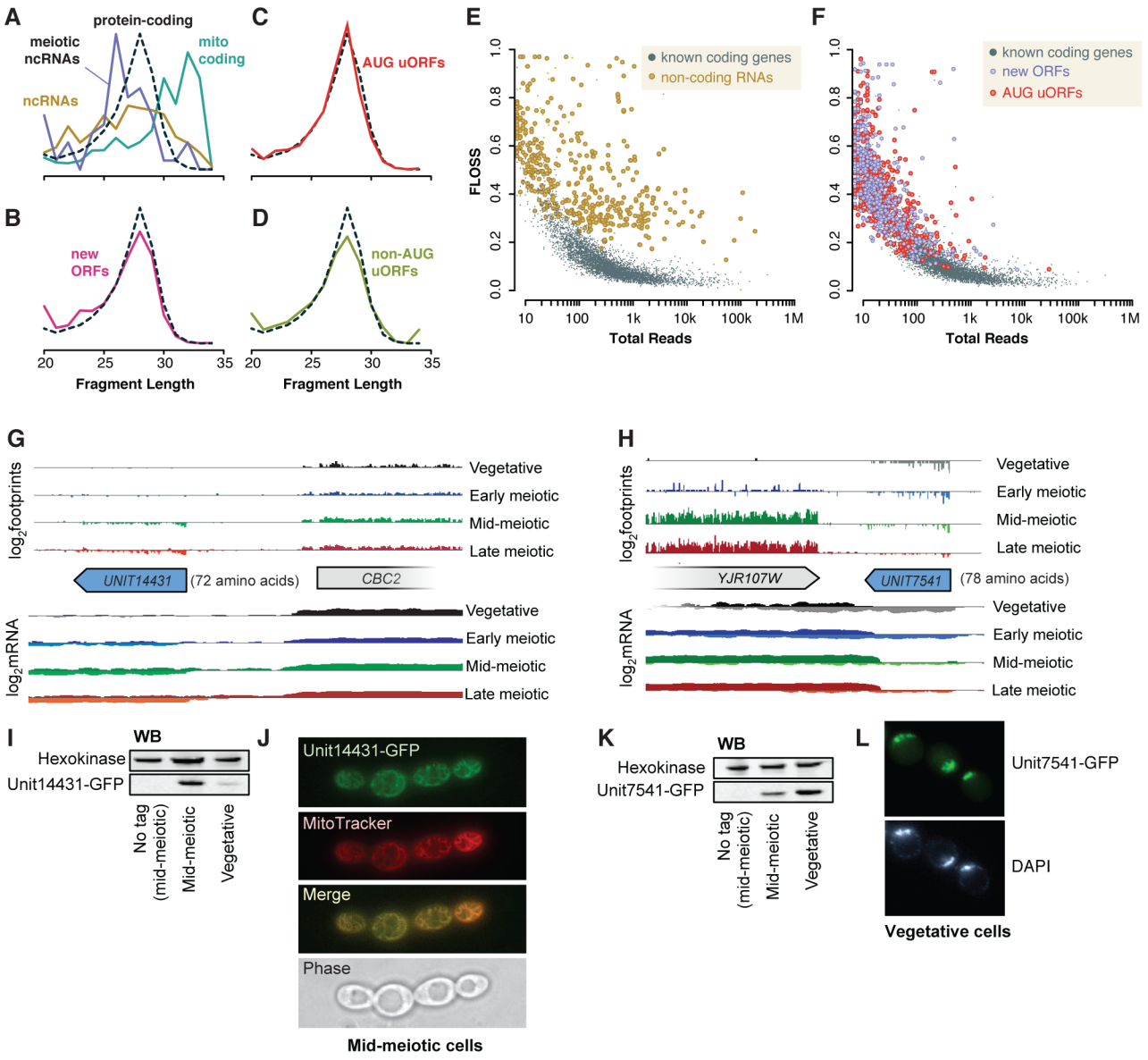

Figure 5. Novel meiotic reading frames based on true ribosome footprints yield protein products

(A – D) Distribution of fragment lengths mapping to nuclear coding sequences compared to (A) classical non-coding RNAs, meiotic lncRNAs, and mitochondrial transcripts, (B) novel independent ORFs, (C) translated AUG uORFs, and (D) translated non-AUG uORFs. (E, F) Fragment length analysis of yeast coding sequences compared to (E) classical non-coding RNAs and (F) novel independent ORFs and AUG uORFs. (G, H) Ribosome profiling and mRNA-Seq data for novel reading frames showing meiotic induction (G) or repression (H) of a ~75 codon ORF on an independent transcript (Brar et al., 2012). (I) Western blot confirming meiotic expression of the Unit14431-GFP fusion. (J) Microscopy on meiotic yeast reveals mitochondrial targeting of the Unit14431-GFP fusion. (K) Western blot confirming vegetative expression of the Unit7541-GFP fusion. (L) Microscopy demonstrating nuclear localization of Unit7541-GFP.

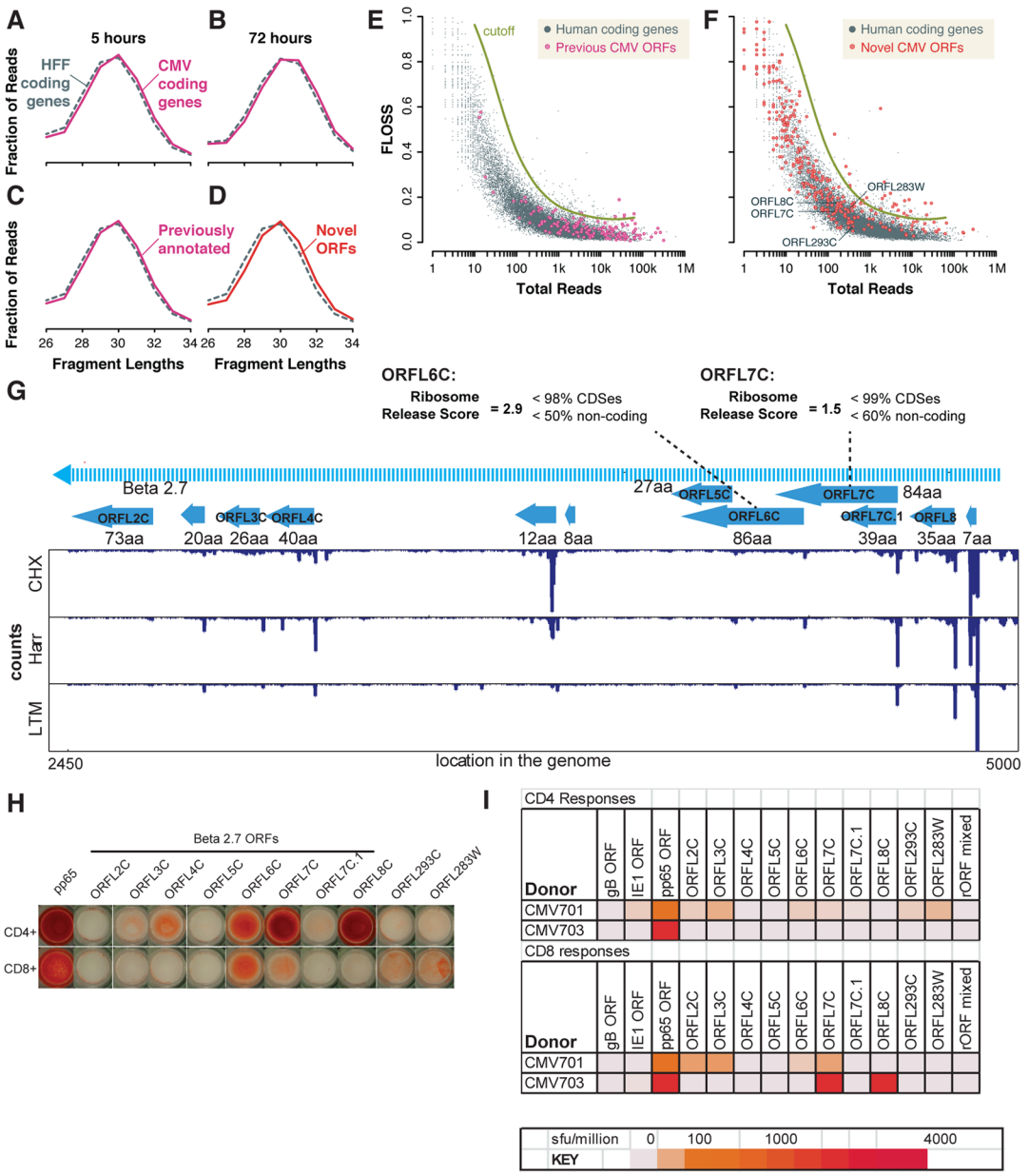

Figure 6. Novel human cytomegalovirus reading frames based on true ribosome footprints lead to antigens in humans

(A – D) Distribution of fragment lengths mapping to human nuclear CDSes compared to all annotated CMV coding sequences after (A) 5 hours or (B) 72 hours of infection, and of specifically the (C) previously annotated and (D) novel CMV coding sequences after 5 hours of infection. (E, F) Fragment length analysis of human coding sequences compared to (E) previously annotated CMV reading frames and (F) novel CMV annotations. (G) Ribosome fooptirnt organization on beta 2.7 transcript (Stern-Ginossar et al., 2012). (H) ELISPOT assay of human donor T cell responses to novel CMV reading frames along with controls. (I) Quantitation of ELISPOT data.