Regulatory start-stop elements in 5' untranslated regions pervasively modulate translation

Authors:

Justin Rendleman[1,2*,§], Solomon Haizel[1,2,§], Shaohuan Wu[1,2], Junjie Liu[1,2], Xinyi Ge[1], Huijing Zou[1], Mahabub Pasha Mohammad[3], Matthew Pressler[1,2], Shuvadeep Maity[1,2,4], Vladislava Hronová[3], Zhaofeng Gao[5], Anna Herrmannová[3], Amy Lei[1,2], Kristina Allgoewer[1,2,6], Daniel Sultanov[1,2], Will Edward Hinckley[1,2], Krzysztof J. Szkop[7], Ivan Topisirovic[8], Ola Larsson[7], Maria Hatzoglou[5], Leoš Shivaya Valášek[3,#,*], Christine Vogel[1,2,#*]

Affiliations:

[1]Department of Biology, New York University; New York, NY, USA.

[2]Center for Genomics and Systems Biology, New York University; New York, NY, USA.

[3]Laboratory of Regulation of Gene Expression, Institute of Microbiology of the Academy of Sciences of Czech Republic; Prague, Czech Republic.

[4]Department of Biological Sciences, BITs-Pilani; Hyderabad, India.

[5]Department of Genetics and Genome Sciences, Case Western Reserve University; Cleveland, OH, USA.

[6]Department of Biology, Humboldt University, Berlin, Germany.

[7]Department of Oncology-Pathology, Science for Life Laboratory, Karolinska Institute; Stockholm, Sweden.

[8]Lady Davis Institute, Gerald Bronfman Department of Oncology, Departments of Biochemistry and Experimental Medicine, McGill University; Montréal, Quebec, Canada.

Co-last authors

* Corresponding authors: cvogel@nyu.edu; valasekl@biomed.cas.cz

§ Co-first authors

Abstract

Various mechanisms controlling ribosome scanning and initiation have been identified within the 5' untranslated region (UTR) of eukaryotic genes, including upstream open reading frames, specific sequence motifs, and secondary structures. We describe the impact and distinct characteristics of a translation regulatory element consisting of a start codon immediately followed by a stop codon. While to date this start-stop element has been described only for specific cases, we identified them in the 5'UTR of hundreds of human genes whereby their general impact on translation control in mammalian cells has been unnoticed. We show that the element is evolutionarily conserved and represses translation of the downstream open reading frames. While having no apparent impact on translation termination or RNA stability, start-stops efficiently halt ribosomes and act as a barrier for the scanning of the small ribosomal subunit. Start-stop element occurs in many transcription factors and key signaling genes, and we highlight routes in which they can affect cellular fate. We investigate the start-stop element in several genes, i.e. *MORF4L1*, *IFNAR1*, *SLC39A1*, and *PSPC1*, and delineate the role of the start-stop in translation regulation of *ATF4* mRNA. Thus, start-stops uniquely modulate translation through elongation-less ribosome stalling.

Introduction

The majority of mRNA translation in eukaryotes occurs in a 5' cap-dependent manner. First, the 43S preinitiation complex (PIC) is recruited to the 5' mRNA cap with the help of the eukaryotic translation initiation factors 4F (eIF4F) and eIF3. Subsequently, these factors together with other eIFs facilitate the scanning of the 48S PIC along the 5' untranslated region (UTR) of the mRNA until the start codon is recognized and translation is initiated [1–3]. Hence, the scanning process naturally compels the 48S PIC to read the genetic information encoded in the 5' UTR, which includes a number of translation control elements that affect protein production, often in a condition-dependent manner [4]. Despite decades of extensive research, understanding of the regulatory code embedded in the 5' UTR is still incomplete, largely due to the sheer number of elements and plausible regulatory outcomes.

Upstream open reading frames (uORFs) are one of the regulatory elements frequently found in 5' UTRs, which occur in approximately half of human mRNAs [5]. Functional uORFs undergo standard translation, including initiation, elongation producing the uORF peptide, and termination. uORFs often act to repress translation of downstream coding regions [5], although exceptions exist, where they act as translation reinitiation-inducible elements [6–12]. A special class of uORFs is an element composed of an AUG immediately followed by a stop codon. Because it inherently excludes an elongation step, it does not represent *bona fide* open reading frames, and we refer to this element as start-stop.

Recent literature has revealed a few examples of regulatory start-stops, such as in the 5' UTR of three *Arabidopsis thaliana* genes, including NIP5;1 [13]. NIP5;1 encodes a boric acid channel, and the start-stop appears to stall 80S ribosomes in a boron-dependent manner, triggering mRNA degradation and thereby preventing expression of the gene. A similar, boron-dependent effect on translation was observed for start-stop sequences in other plant and yeast genes [14–16]. In mammals, a start-stop has been described in the mRNA leader of *HOXA11* encoding an embryogenesis-promoting transcription factor [17]. In this gene, the start-stop interacts with a highly stable stem-loop to inhibit *HOXA11* mRNA translation. However, as the stem-loop is sufficient to stall 80S ribosomes, the precise role of the *HOXA11* start-stop remains unclear. Finally, a start-stop resides upstream of two major regulatory uORFs within the 5'UTR of the master transcription factor ATF4 regulating the integrative stress response [12,18–20]. In this gene, the start-stop's function was shown to depend on the non-canonical initiation factors DENR and MCTS1, as DENR knockout partially reduced the overall stress-induced upregulation of *ATF4* mRNA translation [19]. The same work also suggested that DENR/MCTS1 in general are critical for re-initiation after start-stop elements, i.e. for the translation of the downstream open reading frames.

However, two major aspects of the start-stop element have been overlooked. First, they are distinct from uORFs in terms of their omission of the elongation step, therefore they must affect ribosome processivity in the 5' UTR through a non-canonical mechanism that involves termination without a single elongation cycle. Second, they occur in hundreds of human genes [19], but their impact on translation of these genes is not clear. Therefore, we hypothesized that there are a multitude of ways, beyond specific examples and conditions such as the response to boron, by which start-stop elements augment expression regulation. Indeed, we demonstrate that start-stops affect many signaling molecules and transcription factors in the human genome. They are highly conserved and stall ribosomes more efficiently than uORFs. Unexpectedly, start-stop elements do not show signs of inefficient termination and also do not trigger mRNA degradation. Overall, we propose that they represent an important class of regulatory elements that repress translation of the downstream open reading frames and therefore significantly contribute to regulation of gene expression.

Results

Start-stop elements efficiently stall ribosomes

We defined start-stops as sequence elements in the 5' UTR of eukaryotic genes and that consist of a start codon immediately followed by a stop codon, focusing only on AUG start codons (**Figure 1A**). We defined control sequences as "AUUGGA", "AUUGAA", or "AUUGAG", i.e. start-stop sequences with the inner two nucleotides swapped. We predicted 1,417 and 1,837 human genes with start-stop or control sequences in the 5'UTR, respectively (**Table S1**).

We first examined ribosome coverage of start-stop elements compared to control sequences in a high coverage dataset

collected from HeLa cells, in which ribosome footprinting information had been collected for six replicates [21](**Figure 1B**). We required the length of a ribosome protected fragment (RPF) to be 28-30 nt and any ribosome occupying the start-stop to be positioned with the start and stop codon localizing exactly to the P- and A-site, respectively. We plotted ribosome footprints from these data against the respective number of mRNA reads at the start-stop position. We found that start-stop containing sequences were substantially more often occupied by ribosomes than control sequences. The start-stop in *ATF4* showed the highest occupancy, followed by the start-stops in *UCK2*, *ELP5*, *SLC39A1*, *NFIA*, and *DROSHA*.

Next, we generalized the analysis of the P- and A-site codon frequencies to analysis of all 28-30 nt footprints observed in the 5' UTRs (**Figure 1C, Table S2**). In general, AUG codons were strongly enriched in P-site positions within the 5' UTR, and all three stop codons were enriched in the A-site. Indeed, we observed the overall highest ribosome occupancy for the start-stop sequences, i.e. a combination of an AUG in the P-site and a stop codon in the A-site. In contrast, ribosomes were depleted from combinations with AUG in the A-site, as well as from those with stop codons in the P-site. One hexamer, AUGGCG, exhibited occupancy similar to that of start-stop sequences (**Figure 1C,** black arrow). This hexamer matches the favored context for initiation in mammals [22,23] and most likely represents the initiation site of uORFs in the 5' UTR. Additionally, near cognate start codons (ACG, CUG, GUG, and UUG) also showed high ribosome occupancy when positioned to the P-site of the hexamer. This effect was stronger with a stop codon positioned to the A-site, specifically for CUGUAG, CUGUGA, ACGUGA, ACGUAA, and ACGUAG, suggesting these hexamers, consisting of near cognate start codons followed by a stop codon, might represent non-canonical start-stop elements. Subsequent analyses focussed on start-stop elements starting with an AUG.

Start-stop elements repress translation of many signaling genes

We repeated the analysis in a ribo-seq dataset collected from HEK293 cells [24]. Notably, occupied start-stop sequences were significantly more abundant than occupied control sequences in both the HeLa (n = 188, n = 36) and HEK293 (n = 240, n = 38) datasets (p-value < 0.00001, **Figure 2A**). Intersecting the HeLa and HEK293 datasets we derived a high-confidence core dataset of 88 start-stop and 7 control sequence genes, which had more than three footprint counts on the respective element in both lists (occupied start-stop n = 88, occupied control sequence n = 7, unoccupied start-stop n = 21, unoccupied control sequence n = 85; **Figure 2A, Table S3**). We proceeded to analyze the properties of this core dataset (**Table S4**).

We found that the 88 genes containing occupied start-stops had significantly lower translation efficiency compared to the occupied control sequence genes (p-value < 0.001, **Figure 2B**) and were more evolutionarily conserved (p-value < 0.01, **Figure 2C**). Indeed, we also found many start-stop-containing genes in the single cell eukaryote *S. cerevisiae* (**Table S5**). Occupied start-stops occurred closer to the transcription start site than unoccupied start-stop sequences and tended to involve the first AUG in the 5' UTR, both in the HeLa dataset and independent nanoCAGE data from MCF7 and HEK293T cells [25](**Figure S1**). However, the few occupied control sequences also occurred early within transcripts. Proximity of the start-stop element to the 5' cap can block binding of additional preinitiation complexes through steric hindrance and might also be important for recruiting additional factors [12].

Occupied start-stop sequences had a sequence context consistent with the Kozak consensus sequence surrounding translation initiation sites [22], e.g. with a preference for A or G at the -3 position (**Figure 2D**). Unoccupied start-stops as well as occupied and unoccupied control sequences had mixed preferences at the -3 position (**Figure 2D, Figure S1**). Occupied start-stops also strongly preferred G at the -1 position compared to all other groups (**Figure S1**), again consistent with the Kozak sequence. We also observed a general GC enrichment within +/- 15 nucleotides surrounding the occupied start-stops, contrasting the relative AU rich regions surrounding unoccupied control sequences (**Figure 2D**). The GC enrichment could indicate RNA secondary structures [26] that slow ribosomes as they scan to reinforce initiation immediately upstream of their occurrence in the sequence, or potentially mediate interactions with additional initiation factors.

We manually examined the occupied control sequences for potential reasons for their ribosome binding (**Table S6**). Two occupied control sequences were located within a longer uORF - which would explain the occupancy with ribosomes. In the other occupied control sequences, the AUU in the element could potentially serve as a non-canonical initiation site [27], providing another explanation for ribosomes binding to them. By definition, all control sequences had a G in the +4 position, in support of the Kozak sequence (**Table S6**). All control sequences also had A, G, or C at the

-1 position, and 5 had A or G at the -3 position, further strengthening the Kozak context for possible initiation at the AUU. However, despite these explanations for ribosomes binding to the control sequences, their low abundance (n = 7) compared to that of occupied start-stops (n = 88) illustrates the strong attraction of ribosomes to start-stop elements.

Next, we examined a ribo-seq dataset collected from mouse primary keratinocytes [28] and identified 113 genes with a ribosome positioned to the start-stop at least two RPF counts (**Table S7**). Of these, 23 overlapped with a larger set of human start-stop genes (**Figure 2E**). This overlap is substantial given the stringent cutoffs that we used (e.g. at least three and at least 2 RPF in human and mouse, respectively), the different cell lines considered, and the rapid evolution of untranslated regions [29]. In addition, in all three datasets, start-stops in 5' UTRs were more often occupied by ribosomes than other dicodon sequences: analyzing all possible hexamers (n = 4,096) in comparison to start-stop sequences we identified significantly stronger ribosome binding to start-stop sequences in the 5' UTR compared to the coding region (CDS) or 3' UTR (p-value < 0.001, **Figure 2F**). In all three cell types, start-stops in the 5' UTR were among the top hexamers occupied by ribosomes.

Furthermore, start-stop-containing genes from all three datasets were significantly biased in their functions (**Figure 2G**, **Table S8**). They participate in various signaling pathways, e.g. ERK5, TGF- β, and PFKFB4, while being implicated in cancer (p-value < 0.001). Start-stop containing genes were also significantly enriched in transcription factors (p-value < 0.001), such as *ATF4*, *NFIA*, *JARID2*, and *SMAD6/7*. *ATF4* is discussed in more detail below. NFIA is an important transcription factor involved in gliogenesis with some links to the stress response [30,31]. *JARID2* is a transcription repressor with roles in development [32]. Smad proteins play a key role in TGF-β signaling regulation [33]. The start-stop sequence is conserved in one transcript isoform of the *Drosophila melanogaster* homologous gene Mothers against dpp (Mad). When expressed in S2 cells, the start-stop-containing isoform is occupied by ribosomes [34]**(Figure S1)** supporting the likely functionality of the element and suggesting that expression of transcript variants with and without the start-stop element in diseased tissue might contribute to the dysregulation of SMAD7 protein levels frequently observed in cancers [34].

Start-stop elements are often part of complex regulatory scenarios

Next, we examined typical start-stop gene architectures to illustrate their modus operandi (**Figure 3A**). We found start-stops occur in a variety of contexts, indicating highly transcript specific roles. For instance, in some genes such as *MORF4L1, UCK2, MEN1, NFIA, CALCUL1,* and *CTXN1*, the start-stop appeared to be the only detectable sequence element in the 5' UTR. Many other genes with start-stops had more complex architectures that included additional uORFs, e.g. *ATF4, GPR37, DROSHA, ELP5, IFNAR1, SLC39A1,* and *PSPC1*(**Figure 3A**).

In some genes, e.g. *ATF4* and *ELP5*, the start-stop occurred close to the 5'cap (**Figure 3A**), potentially obstructing scanning by the small subunit. In other genes, e.g. *DROSHA* and *IFNAR1*, the start-stop was located in close proximity to the CDS, likely both reducing ribosome scanning and repressing reinitiation downstream of the start-stop by impairing the ribosome's ability to reinitiate fast enough for efficient translation. This effect is due to an insufficient time/distance for the post-termination ribosome to reacquire the ternary complex (TC) whose concentration is frequently limiting.

Further, we examined examples of start-stop genes for the positioning of the element with respect to the exon-intron structure of the 5' UTR. We identified several cases in which either alternative transcription start sites or alternative splicing produces transcript variants for the same gene that either included or excluded the start-stop (**Figure 3B**). For example, in several genes the start-stop was located in the first exon, but the different transcript variants used different first exons (**Figure 3B**, left). Examples of these genes include *MEN1, ELP5, IL1R1, SLC39A1,* and *USP1*. We observed high coverage of the *MEN1* start-stop in both human and mouse (**Table S7**). *ELP5* and *SLC39A1* are highly occupied in HeLa and covered in HEK293 (**Table S7**); the mouse orthologs have no start-stop, but short uORFs instead. The start-stops in *IL1R1* and *USP1* were highly occupied in HeLa, but less so in the other datasets. In other examples, e.g. *SLC38A1* and *ZNF711*, the start-stop was located in an internal exon in the 5' UTR, and the exon was either present or absent in alternative variants (**Figure 3B**, right). In both genes, the start-stop showed high ribosome occupancy in the human datasets. We hypothesized that transcript variants including the start-stop would be translationally repressed, while transcript variants excluding the start-stop would be translated - providing a potential mechanism of expression regulation at the level of translation.

Next, we demonstrated the impact of the start-stop on translation of several example genes, i.e. *MORF4L1, IFNAR1, SLC39A1,* and *PSPC1* (**Figure 3C-F**). Mutating the start-stop significantly de-repressed translation in all these genes (p-value < 0.001), supporting our hypothesis that start-stops impede ribosomes as they scan through the 5' UTR, while the mutation removes this blockade. Notably, the mutated start-stop did not alter transcript stability in any of the four genes (**Figure S2**).

MORF4L1 is a little-characterized gene involved in cancer cell migration [35], and conserved in mouse and human. Mutating its start-stop derepressed translation in both HeLa and HEK293T cells (**Figure 3C**). For IFNAR1, which functions in JAK-STAT interferon signaling [36], mutating the start-stop in HeLa cells lead to an almost 8-fold increase in translation of the reporter, despite the presence of an additional uORF (**Figure 3D**). Mutating the start-stop in SLC939A1 and PSPC1 increased translation of the reporter ~1.5-fold (**Figure 3E, F**), suggesting that the start-stop is functional. SLC39A1 is a member of the zinc-iron permease family and linked to several cancers [37,38]. Its 5' UTR variants can include or exclude the start-stop (**Figure 3B**), while encoding the same protein (Zhang et al. 2021; Desouki et al. 2015; Dong et al. 2014). A large-scale screen identified an internal ribosome entry site also present in SLC39A1's 5' UTR suggesting a complex regulatory model which might explain the comparatively low impact of the start-stop (Yang et al. 2021). Paraspeckle component protein 1 (PSPC1) marks mammal-specific RNA–protein nuclear bodies that regulate gene expression. The protein is linked to several cancers [39].

Start-stop elements repress translation more so than uORFs

Given that start-stops are related to uORFs, but critically function without an elongation step, we investigated characteristics of start-stop elements in comparison to uORFs. First, to compare the overall impact of start-stops and uORFs on transcript stability and translation efficiency, we conducted a series of gain-of-function experiments in which we inserted start-stop elements as well as a uORF into the 5' UTR sequence of *ACTB* (**Figure 4A**). *ACTB's* 5' UTR has average length (84 nucleotides) and to the best of our knowledge does not contain any known translation regulatory elements. To mimic scenarios we had observed in start-stop-containing genes (**Figure 3A**), we inserted the start-stop or uORF either near the 5' cap or the coding region (**Figure 4A**). Using quantitative RT-PCR and luciferase reporters, we estimated the impact of the elements on transcript stability and translation. We focused on a comparison with a nine-nucleotide uORF which differed from start-stop elements through the addition of a single codon. We chose the additional codon to be GCG as it had one of the highest average ribosome occupancies amongst all 'middle' codons in human nine-nucleotide uORFs (**Figure 1C, Figure S3**). In addition, the GCG codon provided a strong Kozak sequence for the uORF, further ensuring its ability to attract ribosomes.

Indeed, both the start-stop and the uORF repressed translation of the reporter compared to the wild-type *ACTB* 5' UTR (**Figure 4B**). This repression could be attributed to their negative impact on ribosome scanning. Further, for both the start-stop and the uORF, the effect was about two times stronger when positioned close to the CDS (**Figure 4B**), suggesting that not only was ribosome scanning affected, but reinitiation was decreased if the downstream AUG was near the end of the respective element, as expected based on insufficient time to recruit a new TC.

Notably, we observed significantly stronger translation repression when inserting a start-stop compared to when inserting an uORF (p-value < 0.001), regardless of the position within the 5' UTR (**Figure 4B**). Inserting a start-stop led to a 5- to 10-fold reduction in translation, while we observed only a 3- to 6-fold reduction for inserted uORFs. The finding indicates that it is more difficult for the translation machinery to overcome the elongation-less start-stop than a uORF. Further, start-stops and uORFs differed in their effect on mRNA levels. While the *ACTB* 5' UTR transcript with an inserted uORFs remained at wild-type levels, inserting a start-stop element slightly stabilized the mRNA (up to 1.3-fold, **Figure 4C**). This stabilizing effect was unexpected, as it opposes what we had observed at the level of translation. However, the finding was consistent with what we had observed for the individual start-stop genes *MORF4L1, IFNAR1, SLC39A1,* and *PSPC1*, where mutating the start-stop resulted in decreased transcript levels (**Figure S2**).

Secondly, we further explored differences between start-stop elements and uORFs by examining large-scale studies. **Figure 4D** shows the analysis of data from a screen measuring the impact of the 5' UTR on translation efficiency, as estimated from the distribution of the reporter sequences across different fractions of a polysome profile [40]. The sequences comprised randomized stretches fused to a shared 5' UTR, and we then examined these randomized sequence stretches for the presence of start-stops or uORFs. Similar to the results from the gain-of-function

experiments (**Figure 4B**), both start-stops and uORFs repressed translation compared to the control, i.e. shifting sequences to the monosome fraction (**Figure 4D**). However, start-stops had a significantly stronger effect than uORFs (p-value < 0.001). We confirmed these findings with data from another large-scale screen with randomized sequences, demonstrating a similar trend (**Figure S4**): start-stops were present in monosomes significantly more often than uORFs (p-value < 0.01). Further, start-stops stabilized mRNA slightly more than uORFs (p-value < 0.05), confirming the above findings.

Start-stop elements show stronger ribosome binding than uORFs

Next, we tested start-stops for characteristics that could indicate impaired termination (due to the lack of elongation) that could perhaps delay ribosome release from the element. To do so, we exploited findings from an analysis which had shown that the ribosome footprint length is representative of the presence or absence of tRNAs or initiation/release of factors in the A-site: ribosomes with empty A-sites produced shorter reads (21-22 nucleotides) than ribosomes with occupied A-sites (28-30 nucleotides)[41]. As translation inhibitors can severely bias the read length distribution, we used a ribosome footprinting dataset acquired for HeLa cells without any inhibitors [41].

Using this dataset, we first showed that start-stops indeed attracted ribosomes significantly more often than uORFs or control sequences (p-value < 0.001, **Figure 4E**). This finding is consistent with their stronger effect on translation of the downstream open reading frame, as observed in **Figures 4B** and **D**: ribosomes appear to be substantially delayed or even completely halted at start-stop elements, as well as the findings from the initial analysis (**Figure 1B**).

Secondly, we showed that ribosomes on start codons at uORFs or main open reading frames (CDS) had empty A-sites more often than ribosomes on stop codons at these sequences, indicating that ribosomes waiting at start codons for aminoacylated tRNAs are more common than ribosomes waiting for release factors at the stop codon (**Figure 4F**). In other words it seems that the conversion rate of the just assembled 80S initiation complex into an elongation competent 80S ribosome is slower than that of the latter into a termination complex. Ribosomes stalled at the start-stop elements had read length distributions significantly different from any of the other distributions with values between those observed for the start and stop codon at uORFs/CDSs (p-value < 0.01, **Figure 4F**).

A start-stop element in the human ATF4 gene supports the uORF-driven delayed reinitiation

Finally, we examined the role of the start-stop element in *ATF4* in more detail. *ATF4* is the quintessential example of uORF-mediated translation induction in metazoans [4,10]: two sequential uORFs repress translation under normal conditions and induce translation during the integrated stress response, through a mechanism called delayed reinitiation. In this mechanism, the preinitiation complex scans the 5' UTR until it recognizes the start codon of the first uORF (uORF1), and the short ORF is translated. Following termination at this uORF, the small ribosomal subunit remains associated with the mRNA and continues scanning downstream. Downstream reinitiation is dependent on ternary complex abundance: when TC levels are high, i.e. under normal conditions, TC recruitment occurs rapidly and reinitiation is favored at the next actionable start codon, which corresponds to the second uORF (uORF2). This second uORF overlaps with the main ORF. Therefore, termination of uORF2 occurs downstream of the main *ATF4* start codon, suppressing ATF4 protein synthesis. During stress conditions, when ternary complex abundance is lower, reinitiation rates at the second uORF are reduced, promoting bypass of its start codon. However, as the uORF is very long (59 amino acids in total), the small subunit eventually recruits a ternary complex and the scanning ribosomes can reinitiate further downstream at the main ORF, thereby inducing ATF4 translation [7,8]. While this model has been well-established, several observations contrast the original model's predictions [21,42,43] and have triggered additional investigation [44].

The human *ATF4* gene contains a start-stop element located upstream of uORF1, with very high ribosome occupancy in the HeLa data under both normal and stress conditions (**Figure 5A**)[21]. While the start-stop sequence is highly conserved across vertebrates (**Figure S5**), it does not occur in the mouse Atf4 sequence where the original delayed reinitiation model had been discovered [7,8]. Prior work examined the start-stop in the human *ATF4* with respect to factors that impact reinitiation *downstream* of the element [19,45,46]. Specifically, it showed that reinitiation after the start-stop depended on the function and expression of the reinitiation factors eIF2D and DENR/MCTS1 [19,45,46]. The studies proposed a mechanism in which DENR/MCTS1 and/or eIF2D are required for release of the deacylated tRNA from the ribosome to enable recycling of the large subunit, continued scanning of the small subunit, and downstream reinitiation [19].

While these findings explain how translation can resume after the start-stop in the human *ATF4* gene, they do not explain the specific impact of the start-stop on the ATF4 translation system itself. Given that the ribosome read counts at the start-stop are higher than anywhere else in the 5' UTR (**Figure 5A**), independent of conditions (**Figure S5**) , we hypothesized that the *ATF4*'s start-stop has a significant impact on the overall ribosome flow and translation of both the upstream and main open reading frames. Specifically, we hypothesized that the start-stop would support the delayed reinitiation model, i.e. ensure translation suppression of the main open reading frame under normal conditions, but reduce wasteful production of the long peptide encoded by the second uORF and therefore improve cellular economy. To test this hypothesis, we examined a series of luciferase reporter constructs, as well as CRISPR-Cas9 edited Hap1 cell lines.

Indeed, removing the start-stop by CRISPR-Cas9 (6 nucleotide deletion) resulted in little observable changes in basal ATF4 levels as measured by Western blotting, consistent with the established model in which the two uORFs are sufficient to repress ATF4 translation under normal conditions (**Figure 5B**). The results were the same when modifying the start-stop to a stop-stop (UAGUAG) sequence (**Figure S5**). The result was also confirmed by a reporter construct which contained the start-stop but no other functional uORF (**Figure S5**): the start-stop reduced the overall ribosome flow and hence translation of downstream elements.

Further, similar to what we observed for the start-stops in general (**Figure 4C, Figure S2**), the start-stop in *ATF4* appeared to stabilize the transcript compared to the wild-type (**Figure 5C**, **Figure S5**). This result also aligns with previous work that has shown the translation of second uORF triggers *ATF4* mRNA degradation via nonsense-mediated mRNA decay (NMD)[47]. Thus, reducing translation of the second uORF through ribosomes stalling at the upstream start-stop under normal conditions would lower NMD-related degradation and consequently increase *ATF4* mRNA half-lives - as we observed.

Unexpectedly, we observed a second role of the start-stop: mutating it resulted in reduced ATF4 induction under stress (**Figure 5B**), independent of the changes in transcript levels (**Figure 5C**). We hypothesized that the reduced inducibility in our system was due to the start-stop supporting stress-induced bypass of the second uORF. To test this hypothesis, we designed reporters that fused the human *ATF4* 5' UTR up to the second uORF's start codon to the firefly luciferase coding sequence (**Figure 5D**). The reporter also included a renilla luciferase expressed from the same plasmid as firefly luciferase. The renilla luciferase intensity served as an internal control to monitor overall translation levels under normal conditions and under stress (**Figure S5**).

This construct allowed us to quantify translation reinitiation at uORF2 (through the firefly readout). We hypothesized that the start-stop would not affect reinitiation rates *after* uORF1. However, the start-stop as a road block would lower the total number of initiation-competent ribosomes reaching uORF1 under both normal or stress conditions, and thus affect the number of post-termination 40S ribosomes scanning downstream of the uORF1's stop codon. Further, reinitiation at uORF1 would be reduced under stress, as low TC levels hinder TC recruitment and production of a scanning competent PIC after termination from the start-stop. Initiation-incompetent ribosomes skipping the AUG of uORF1 would then be given more time to become initiation-competent en route to the AUG of uORF2, explaining its observed translation even under stress (**Figure 5A**).

Translation of the wild-type reporter - i.e. the construct including the start-stop, uORF1, and the uORF2 fused to the firefly reporter - decreased under thapsigargin compared to normal conditions, but was still translated, albeit at lower levels (**Figure 5D**). Mutating the start-stop resulted in higher expression of the firefly reporter compared to the wild-type sequence, under both normal and stress conditions, again with reduced levels under stress (**Figure 5D**). The result suggests that the start-stop acts as a road-block which reduces ribosome flow along the 5' UTR and lowers translation of uORF2 under both normal and stress conditions.

Examining these data more closely, we noticed a significant difference in the stress-induced drop in translation of uORF2, i.e. its bypass, as measured by the reduction of firefly signal under normal to stress conditions: stress-induced bypass of the second uORF was ~4 times larger in the wild-type construct compared to the start-stop mutant (~12% vs. ~3%, p-value < 0.001, right panel, **Figure 5D**), after accounting for stress induced overall changes in translation as estimated by the renilla luciferase control. The renilla luciferase control decreased with stress due to the global decline in initiation during the stress response, but was not affected by the presence or absence of the start-stop (**Figure S5**).

These results suggested a model in which, under normal conditions and in the absence of the start-stop, like in the mouse *Atf4* 5' UTR, more ribosomes reach uORF1 than in the presence of the start-stop (**Figure 5E**). Both in the presence and in the absence of the start-stop, most ribosomes will reinitiate at uORF2 upon termination from uORF1 due to high TC abundance. However, as the start-stop lowers ribosome flow, it would also lower translation of uORF2, i.e. render the system more economical, in particular under normal conditions. Under stress, the start-stop promotes by-pass of uORF2 while still allowing its translation as observed by us and others [21,42,43], therefore supporting ATF4 induction. The precise interplay of the start-stop with the uORFs and other cis-acting elements is discussed elsewhere [44].

Finally, we tested whether our interpretation of the experimental observations could be matched by a simple mathematical model that reflected different regulatory scenarios (**Figure S6**). To do so, we incorporated the quantitative measurements into the model and conducted deterministic simulations to examine the impact of the start-stop on ATF4 translation. In the model, we based a sigmoidal dependence of reinitiation at uORF2 on the levels of phosphorylated eIF2alpha, according to a Hill-Langmuir relationship of cooperative binding. We then examined the increases in reinitiation at the CDS as a result from lowering translation of uORF2. The simulation showed that even only 10% less overall reinitiation at uORF2 due to a start-stop reducing ribosome flow - but at the same overall availability of the ternary complex - would increase ATF4 translation under stress several fold compared to a model without the start-stop (**Figure S6**). The model also confirmed that within biologically relevant ranges of phospho-eIF2alpha and *ATF4* mRNA levels, the start-stop is capable of reducing costly translation of uORF2, while ensuring even higher ATF4's induction rate under stress.

Discussion and conclusion

We present an in-depth characterization of start-stop elements, a unique group of translation regulatory sequences that occur in the 5' UTR, halt the scanning ribosome and therefore repress translation of the downstream open reading frames. Start-stop elements do not involve elongation, and are therefore distinct from open reading frames. Start-stops occur in >1,400 human genes and are evolutionarily conserved. While start-stop elements have been previously described for a small number of individual genes [17,19], we demonstrate here that they represent an important group of regulators that most likely affect translation of hundreds of genes.

Start-stops often function in conjunction with other elements, e.g. uORFs, and share some of their characteristics. However, we demonstrate that, in comparison to uORFs, start-stops are more efficient in retaining the ribosome - which is remarkable given that start-stop elements have suboptimal nucleotides at the +4 and +5 initiation sites. Ribosome stalling at a start-stop has two main effects: it reduces the number of 40S subunits scanning downstream of the element ("ribosome flow") and second, it directly impedes downstream re-initiation if the start-stop is located close to the next open reading frame. While these properties are similar to those of uORFs, start-stops stall ribosomes more efficiently and they do so without the need for elongation. In addition and in contrast to uORFs, the ribosome halted on start-stops slightly stabilizes the transcript. We do not yet understand the reasons for this effect on mRNA stability.

Ribosomes initiating on the AUG for the start-stop are uniquely positioned with their A-site at a stop codon, rendering these ribosomes incapable of translocating and elongating. This feature is distinct from canonical open reading frames - even minimal uORFs with only one additional codon - that include elongation and translate a sequence into a peptide. As the 40S subunit has been shown to linger upon termination [12,48], reinitiation at the same site is possible as the ribosome is already positioned with the P-site at the start codon even after termination. Therefore, ribosomes could be stalled on a start-stop element either through repeated cycles of initiation and termination, or through impaired termination and ribosome release.

We searched for evidence indicating altered initiation or termination at start-stop elements. Recently, *in vitro* single-molecule analysis suggested that recruitment of release factors to ribosomes on a start-stop element follows the same kinetics as recruitment to stop codons of longer open reading frames [49]. Consistent with these findings, our analysis of length distributions of ribosome footprints indicated no apparent bias at start-stop elements as a result of A-site occupancy of the ribosome. However, neither our analysis nor the other study [49] quantified ribosome release (recycling), which can also impact dynamics of ribosome stalling at start-stops. Ribosome release from start-stops has been indirectly linked to the presence of the reinitiation factors DENR/MCTS1[50,51]: the complex is thought to trigger

the release of deacylated tRNA from post-termination 40S ribosomes, therefore supporting their continuous scanning. Indeed, targets of the DENR/MCTS1 system are present in our set of start-stop containing genes (**Table S9, Figure S7**), but this overlap does not explain the prevalence of footprints on start-stops compared to other DENR/MCTS1 targets.

Further, we investigated the possible modes by which start-stop elements can contribute to regulation of gene expression. As start-stop containing genes frequently occur in transcription factors and signaling molecules, one function might be to prevent harmful overproduction of these genes which could interfere with their proper functioning. Further, suppression at the level of translation also allows for precise subcellular localization of the mRNA prior to translation, as well as an increased speed of expression induction. In addition, as start-stop elements are very efficient in suppressing downstream translation, start-stops (rather than uORFs) might allow for the non-translated transcript to be stored, e.g. in stress granules. Supporting this idea, we observed a significant overlap between start-stop containing genes and transcripts enriched in stress granules (p-value < 0.01, **Table S9**)[52], e.g. *NFIA* and *DROSHA* shown in **Figure 1B**. *DROSHA* has diverse roles in miRNA biogenesis and RNA processing in general, some of them linked to the stress response [53]. *DROSHA*'s start-stop has been described in the context of DENR/MCTS1-dependent reinitiation [19]. Finally, we illustrated how a start-stop element expands the repertoire of translation regulation in the cell at the example of *ATF4*. We demonstrated that ribosome stalling at the start-stop element can reduce wasteful production of the uORF2 peptide and, at the same time, support ATF4 inducibility under stress.

Next, the dynamic inclusion and exclusion of start-stop elements in transcripts can modulate translation across tissues and conditions. For example, we identified genes with transcript variants that include or exclude the start-stop element and where the transcript variants show tissue specific expression patterns, i.e. *USP1* and *ZNF711* (**Figure S8**). *USP1* is a deubiquitination enzyme with diverse roles, e.g. in the regulation of DNA damage repair [54]. It has four transcript variants with several translation regulatory elements, but only one contains the start-stop bearing exon. As USP1 protein is overabundant in cancer [54], we hypothesize that dysregulation of its expression might be, in part, due to switching to transcript variants without the translation repressive start-stop. Another example arises from *ZNF711,* which is of unknown function but with similarity to transcription factors. It is linked to brain development and its mutation causes intellectual disability [55,56]. *ZNF711* has a complex 5' UTR with several exons and three start-stop elements, creating five distinct transcript variants with three different combinations of start-stop inclusion or exclusion (**Figure S8**). All five transcript variants are expressed, with distinct tissue-specific patterns.

Importantly, start-stop mutations might offer a new mechanistic explanation for translation-based pathological phenotypes. In one scenario, gain-of-function mutations could create start-stops in the 5' UTR of a transcript, subsequently repressing its translation. Alternatively, a loss-of-function mutation in the start-stop would de-repress translation of a gene, leading to aberrant expression levels. Indeed, we identified example genes with naturally occurring sequence variants, e.g. single nucleotide polymorphisms (SNPs), located within or next to the start-stop. For example, we identified a SNP disrupting the start-stop of *PTPN11* (**Figure S9**). Its 5' UTR contains a start-stop but no other predictable translation regulatory sequences. *PTPN11* encodes the SHP-2 protein which participates in RAS/MAPK and PI3K signaling. PTPN11/SHP-2 overexpression enhances, amongst others, progression of liver cancer [57]. The start-stop disrupting SNP primarily occurs in the East Asian subpopulation where liver cancer is frequent [58].

Another example is represented by *UHMK1*, a serine/threonine RNA-binding protein kinase that promotes cell cycle progression and has been implicated in perturbations of mRNA export and translation leading to metabolic reprogramming and drug resistance in melanoma [59]. We identified a SNP changing the guanidine immediately upstream of the start-stop in *UHMK1*'s 5' UTR and therefore disrupting its Kozak sequence (**Figure S9**). Both the start-stop and an upstream uORF bind ribosomes more so than the AUG of the *UHMK1*'s coding region, suggesting translation repression of the gene. As *UHMK1* overexpression promotes colorectal cancer cell growth and drug resistance [60], derepressed *UHMK1* translation might contribute to the disease. Consistent with this hypothesis, the SNP is prevalent amongst African Americans and colorectal cancer is more common in this subpopulation [61].

In sum, start-stops represent a hitherto underappreciated element in the 5' UTR that seems to control translation of many key regulators in humans. The lack of the elongation step in start-stops seems to delay or even stall ribosome progression much more so than in the conceptually related uORFs. Therefore, start-stop elements drastically reduce the

ribosome flow along the 5' UTR without production of a peptide, and if positioned near the next downstream initiation codon, reduce the probability of reinitiation. In contrast to ribosome stalling within a coding region, which often leads to transcript turnover [62], start-stop elements slightly stabilize the transcript. This property might support localization and/or storage of the transcript until its translation is desired. Since start-stops often occur in combination with uORFs and other elements in the same 5' UTR, they seem to represent an important part of complex regulatory systems that fine-tune translational control under various environmental conditions.

Methods

Start-stop and control sequences and their ribosome occupancy

We downloaded GENCODE v43 human gene annotation file from the UCSC Table Browser (https://genome.ucsc.edu/cgi-bin/hgTables). When a gene had multiple principal transcripts, we selected the transcript with the longest 5' UTR. To extract the 5' UTRs, which include start-stop sequences ("AUGUGA", "AUGUAA", or "AUGUAG"), as well as control sequences ("AUUGGA", "AUUGAA", or "AUUGAG"), we used a customized R script. In total, we obtained 1,417 and 1,837 genes containing start-stop and control sequences in their 5' UTRs, respectively (**Table S1**).

We downloaded ribosome profiling and RNA-seq data collected in HeLa cells from the Gene Expression Omnibus (GSE113171), i.e. SRA files (study ID SRP140470). In this experiment, HeLa cells were treated with 0.5ug/mL tunicamycin for 0, 1, 4, or 8 hours, and prior to harvest cell growth media was removed and cells were washed with fresh media supplemented with 0.1 mg/mL cycloheximide for five minutes [21]. We converted the SRA files to fastq format and performed trimming using the cutadapt algorithm (version 1.12) to remove low quality bases at the 3' and 5' ends of reads (-q 30,30), and to discard reads shorter than 25 nt (-m 25)[63]. Trimmed fastq files were then aligned to the human genome (NCBI GRCh38) using the HISAT2 aligner (version 2.0.5), with a seed value set to 23 (--seed 23), and no soft clipping of reads (--no-softclip)[64].

We examined the quality of the ribosome footprint data using the RiboTaper analysis pipeline [24,65]. The analysis was based on the observation that genuine ribosome footprint reads typically exhibit a distinct 3-nt periodicity within coding regions. We confirmed this pattern in reads with length ranging from 28-30 nt in the Hela dataset. Therefore, we counted ribosome footprints when the length of the read fell within the 28-30 nt range and its starting position was exactly 12 nt upstream of the AUG or AUU in start-stop or control sequences, respectively (P-site position). We excluded reads with other lengths or positions from our analysis. We used htseq-count to assign reads to genes and to start-stops [66]. GTF files were modified to count reads mapping to start-stops in the 5' UTR by excluding any overlap with the coding sequence (CDS) and requiring a 12 nt distance from the start codon. We required the start-stop's P-site to be located completely within the 5' UTR, but allowed for partial overlap of the 3' end of reads with the CDS to capture ribosomes approaching the start site.

We repeated the analysis with other published datasets, i.e. ribosome profiling and RNA-seq data from HEK293 cells [24], mouse primary keratinocytes [28], and *Saccharomyces cerevisiae* [67](**Table S5**). For the HEK293 data, we followed the same processing steps as described above. Mouse sequencing data was first trimmed for the adapter sequence AGATCGGAAGAGCACACGTCT and then aligned to the NCBI GRCm38 reference genome. Yeast sequencing reads were trimmed with cutadapt to remove polyA adapter sequences, and subsequently aligned to the NCBI *S. cerevisiae* build3.1 reference genome.

To generate the heatmap displaying fractional occupancies of all P-site/A-site combinations, we compiled ribosome footprint data for all ribosome reads in the 5'UTR (**Table S2**). The resulting data was then clustered using the Perseus software (version 1.5.0.15), using the Euclidean clustering method [68].

To determine the fraction of start-stop and control sequences that were occupied in each mammalian dataset compared to all hexamer sequences, we mapped RNA-seq reads containing a start-stop or control sequence anywhere within the read to the 5' UTR; this analysis provided the total number of possible positions where a ribosome could engage with the given sequence (the denominator). We repeated the analysis for reads mapping to the CDS and 3' UTR regions to demonstrate that ribosome occupancy at start-stop is specific to the 5' UTR region. We extended this analysis to include all possible hexamer sequences (AAAAAA, AAAAAC… UUUUUA, UUUUUU; n = 4,096), and calculated

the fractional occupancy at each P-site/A-site combination using the number of reads mapped to that specific combination.

Characteristics of start-stop-containing genes

We compiled a high-confidence core set for human data using the following criteria. Occupied start-stop and control sequences were required to have 3 or more RPF reads in both the HeLa and HEK293 datasets; unoccupied start-stop and control sequences were required to have 0 RPF reads, but at least 6 RNA reads in both HeLa and HEK293 datasets. In addition, we manually quality controlled the resulting sets (**Table S3**). These stringent criteria resulted in 88 and 7 occupied start-stop and control sequences, respectively. Additionally, the core set comprised 21 and 85 unoccupied start-stop and control sequences.

To estimate the number of ribosomes involved in translating the mRNA transcript of each corresponding gene, we calculated the translation efficiency as $\log_2$(RPF/RNA) within the CDS region of the gene. Statistical analysis was performed using a two-sample t-test.

To calculate evolutionary conservation, we downloaded conservation scores from the UCSC Genome Browser (http://genome.ucsc.edu). For each element (start-stop-occupied, start-stop-unoccupied, control-occupied, control-unoccupied), phyloP evolutionary conservation scores from 100 vertebrate species were averaged across the hexamer sequence and statistical comparison of the start-stop-occupied elements with the other three groups was performed by two-sample t-tests, after exclusion of outliers (**Table S4**).

To calculate nucleotide frequencies surrounding specific elements, we extracted sequences surrounding start-stops from transcript annotations. Relative nucleotide frequencies were then calculated for ±15 nt from the beginning and end of each start-stop for comparison.

For analysis of enriched functions, we compiled lists of genes that contained occupied start-stop sequences in their 5' UTRs for HeLa, HEK293, and mouse primary keratinocytes, using less stringent cutoffs. For occupancy, we required at least three RPFs and at least two RPFs on the start-stop in the human and mouse data, respectively (**Table S7**). Gene lists were analyzed through the use of IPA (QIAGEN Inc., https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis)[69](**Table S8**). P-values were adjusted for multiple hypothesis testing using the Benjamini-Hochberg correction [70].

Reporter constructs to assess the effect of start-stop genes on translation and mRNA stability

We used the dual-luciferase psiCHECK2 system (Promega) following the manufacturer's protocol. The psiCHECK2 vector expresses both firefly luciferase and renilla luciferase, allowing for an internal control within the same plasmid. HEK293T or HeLa cells were seeded in a 24-well plate at a density of 4 x 10^4 cells in 1 mL of Dulbecco's Modified Eagle Medium (DMEM) growth media supplemented with 10% fetal bovine serum (FBS, Thermo Fisher) and penicillin/streptomycin (Gibco). The cells were then incubated at 37 ºC and 5% CO_2 until they reached approximately 70% confluency. We transfected cells using 1 ug of reporter plasmid and TurboFect™ (Thermo Fisher) transfection reagent, following the manufacturer's protocol. Subsequently, we used the Dual-Glo® luciferase assay kit (Promega) to measure the activities of both firefly and renilla luciferase in the cell lysate harvested from each well of the 24-well plate, as per the manufacturer's instructions.

To test the function of the start-stops in MORF4L1(NCBI Reference Sequence: NM_006791.4), IFNAR1(NCBI Reference Sequence: NM_000629.3), SLC39A1(NCBI Reference Sequence: NM_001271958.2), and PSPC1(NCBI Reference Sequence: NM_001354909.2), we inserted the 5' UTRs of these respective human genes upstream of the renilla reporter gene (hRluc) at the NheI site of the psiCHECK2 plasmid. To examine the role of the start-stop in a gain-of-function construct, we used the 5' UTR of β-Actin (ACTB, NCBI Reference Sequence: NM_001101.5).

Sequences were mutated using site-directed mutagenesis with primers as shown in **Table S10.** All transfections were performed in triplicate as biological replicate, and both firefly and renilla luciferase activities were measured in triplicate as well. We then calculated relative luciferase activity by dividing the averaged reporter luciferase activity by the control luciferase activity. This value was further normalized by relating all measurements to that of the respective wild-type 5' UTR. Statistical analysis was conducted using two-tailed, paired t-tests, where samples were paired by the day of the biological replicate.

To estimate mRNA stability, HEK293T or Hela cells were transfected with 1ug of the respective plasmid. Total RNA was isolated from the cells using the RNeasy kits (Qiagen) according to the manufacturer's instructions. Reverse Transcription of cDNA was performed using SuperScript III Reverse Transcriptase (Thermo Fisher Scientific). Quantitative RT-PCR was carried out to measure firefly (hluc) and renilla (hRluc) RNA levels using KAPA SYBR FAST qPCR master mix (Kapa Biosystems) with the following program: preincubation at 95 ºC for 3 minutes followed by 40 cycles of 95 ºC for 10 seconds, 60 ºC for 30 seconds, and 72 ºC for 30 second. We then analyzed the samples with the Roche LightCycler 480 system. **Table S10** lists the primer sequences for the RT-PCR.

Comparing start-stops to minimal upstream open reading frames

To compare the impact of start-stop and upstream open reading frame (uORF) sequences on translation, we examined a large-scale dataset published by Seelig et al. [40]. In their analysis, the authors constructed a library of randomized 50-nucleotide 5' UTR oligomers which were fused to the eGFP CDS. They then analyzed the effect on translation by separating sequences by polysome profiling and sequencing the respective fractions. From this dataset, we specifically extracted sequences with either a start-stop, minimal uORF, i.e. sequences which started with "AUG" and ended with "UGA", "UAA", or "UAG", and had one of the 59 sense codons in the middle (excluding AUG as a middle codon), or a control sequence (AUUGAA, AUUGAG, AUUGGA).

To compare the differences between start-stops and uORFs in terms of ribosome binding and ribosome A-site occupation, we analyzed published ribosome footprinting data from HeLa cells [41]. Importantly, these data had been generated without the use of translation inhibitors that could potentially interfere with ribosome positioning and footprint length. We removed the adaptor and random sequences from the raw reads with Cutadapt v3.1 [63]. Subsequently, we aligned the trimmed reads against the human genome GRCh38.p10 with STAR v2.7.6a [63,71], with similar parameter settings to those described in reference [41]: --outFilterIntronMotifs RemoveNoncanonicalUnannotated --outFilterMultimapNmax 1 --outFilterMismatchNoverLmax 0.1 --outSAMtype BAM SortedByCoordinate --outSAMunmapped Within --outSAMattributes Standard. We constructed lists of start-stop, control, and uORF containing genes as described above, using GENCODE v43 human gene annotation file (https://genome.ucsc.edu/cgi-bin/hgTables).

To examine the overall ribosome occupancy, we used a customized R script to count reads across all positions relative to the start and stop sites of the start-stop sequences and of the uORFs, and the first and last codons in the control sequence. This analysis was conducted with a 40 nt window surrounding each element.

To examine the ribosome footprint lengths as a measure of A-site occupancy, we extracted reads positioned to the AUG or stop codon of the start stop, the uORF, or CDS, or the first and last codon of control sequences. We then counted short (20-22 nt) and long (27-30 nt) footprint reads separately. We repeated this analysis for ribosomes at the stop codon of the uORF (A-site) and for the AUG and stop codons of the main CDS. By calculating the average ratios of the short read counts to long read counts across different positions, we obtained an estimate of the occupancy status of the A-site-containing ribosomes.

Relative ribosome density profiles along the ATF4 transcript

To examine ribosome density along the ATF4 5' UTR and beginning of the CDS, we used samtools -depth algorithm [72] and published Hela data [21]. Values were normalized based on library size, using size factors from the DESeq2 R package [73]. To account for variation in ATF4 mRNA levels, we calculated the relative ribosome density for each sample as the ratio of normalized ribosome density divided by the normalized ATF4 mRNA abundance. These relative ribosome densities were then plotted along the ATF4 genomic coordinates [21].

Crispr-Cas9 genome editing of ATF4

We used Crispr-Cas9 based genome editing to modify the start-stop in ATF4 in the human HAP1 cell line. The HAP1 cell line is near-haploid which facilitates generation of a homozygous mutant at the start-stop site. To design Cas9 guide RNA targeting this site, we identified a PAM sequence (NGG) in the ATF4 5' UTR region surrounding the start-stop site. To assess potential target sequences, we employed the Benchling CRISPR Guide RNA Design software (https://benchling.com). Among the options, we selected the guide target sequence CGCAGAGAAAACTACATCTG, which possessed an on-target score of 57.1 and an off-target score of 59.9 [74,75].

We then used the lentiCRISPRv2 plasmid (AddGene 52961) as a one vector system for co-delivery of the hSpCas9 protein and the sgRNA; cloning of the target sequence into the plasmid was performed as previously described [75]. Two oligonucleotides with overhangs were synthesized for cloning the selected sgRNA into the lentiCRISPRv2 plasmid (**Table S10**). The oligos were phosphorylated and annealed with T4 PNK (NEB) supplemented with 10mM ATP at 37 ºC for 30 minutes, followed by 95 ºC for 5 minutes, and a ramp from 95 ºC to 25 ºC at 5 ºC/minute. Next, the plasmid backbone was digested and dephosphorylated using BsmBI restriction enzyme (NEB) and Quick CIP (NEB), with incubation at 55 ºC for 1 hour and heat inactivation at 80 ºC for 20 minutes. We performed gel purification using the QIAquick Gel Extraction kit (Qiagen), and ligated the linearized backbone to annealed oligos using Quick Ligase (NEB). Transformation into NEB Stable Competent *E. coli* bacteria (NEB) was performed according to manufacturer instructions. Single colonies were selected, and plasmid DNA was purified from overnight cultures using the QIAprep Spin Miniprep Kit (Qiagen). Proper annealing of sgRNA into the lentiCRISPRv2 backbone was confirmed by Sanger sequencing (Azenta) using the appropriate primer (**Table S10**).

HAP1 cells were cultured in Iscove's Modified Dulbecco's Medium (IMDM, Corning) supplemented with 10% fetal bovine serum (FBS, Thermo Fisher) at 37 ºC and 5% CO_2 until they reached approximately 40% confluency. The lentiCRISPRv2 plasmid containing the start-stop target sequence was transiently transfected into HAP1 cells using Lipofectamine 3000 (Thermo Fisher) and incubated for 24 hours. Cells were then passaged into a fresh growth medium supplemented with 0.4 ug/mL of puromycin for selection of positive transfection. After three days of growth in the selection medium, individual cells were isolated into separate wells of a 96-well plate using fluorescence-activated cell sorting (FACS). Clonal populations were subsequently grown out and PCR screening was conducted to confirm the presence of the desired band size within the ATF4 region before being sent for Sanger sequencing (Azenta) using primers as shown in **Table S10.** Three lines were selected for follow up based on the sequences at the start-stop in ATF4 5' UTR: Start-stop^{WT} cells, which had no differences in sequence compared to the wild-type human genome, ΔStart-stop cells, in which the start-stop sequence was completely deleted, and Start-stop^{TAGTAG} cells, where the start codon was mutated to a stop codon (AUG->UAG). For stress induction experiments, cells were incubated with a final concentration of 1 uM thapsigargin.

We used Western blotting to measure ATF4 protein levels in the different HAP1 cell lines. To do so, we diluted ATF4 Rabbit monoclonal antibody (Cell Signaling) in a ratio of 1:1000 into a Clean Western Buffer with 5% BSA. Next, we washed the membranes three times in Clean Western Buffer for 5 minutes each and then incubated for 1 hour with Digital Anti-Rabbit secondary antibody (Kindle Biosciences) at 1:2000 in Clean Western Buffer with 5% BSA. Proteins were detected with the KwikQuant Imaging System (Kindle Biosciences). As a loading control detection, we used the same Western protocol, replacing the primary antibody with the β-Actin Rabbit antibody (Cell Signaling) after incubating the blots with stripping buffer (Thermo Fisher).Western blot images were processed with the ImageJ software to quantify protein levels for ATF4 and loading controls. We subtracted background values for each sample and divided ATF4 values by loading control values to account for variability in sample loading between wells. To allow for comparisons across the three replicate blots, these normalized ATF4 values were then transformed based on the percent of the maximum value from each blot (relative % of max).

To measure ATF4 transcript levels in the different HAP1 cell lines, we isolated total RNA from cells using the RNeasy kits (Qiagen) following the manufacturer's protocol. cDNA was reverse-transcribed from RNA samples using SuperScript III Reverse Transcriptase (Thermo Scientific).

To estimate the decay of the ATF4 mRNA, we treated Start-stop^{WT} and ΔStart-stop cells with either RNA polymerase II inhibitor 5,6-dichloro-1-β-ribofuranosyl-benzimidazole (+) or DMSO (-) and isolated RNA after 2 hours. Quantitative RT-PCR was carried out using the KAPA SYBR FAST qPCR master mix (Kapa Biosystems) with samples pre-incubated at 95 ºC for 3 minutes then 40 cycles of 95 ºC for 10 seconds, 60 ºC for 30 seconds, and 72 ºC for 1 second, and analyzed with the Roche LightCycler 480 system. The RT-PCR primer sequences are listed in **Table S10**.

Primer efficiencies were calculated with standard curves generated by qPCR of serially diluted cDNA. ATF4 mRNA levels were quantified relative to RPL19 levels using the following equation:

$$\frac{ATF4\ primer\ efficiency^{-Ct_ATF4}}{RPL19\ primer\ efficiency^{-Ct_RPL19}}$$

We similarly used this formula to calculate 18S rRNA values relative to RPL19 for the control comparison. Statistical analyses were carried out using paired t-tests, where paired samples correspond to RNA samples from Start-stop^{WT} and ΔStart-stop collected on the same day and qPCR reaction done in the same plate.

uORF2 bypass under stress with and without an upstream start-stop

To test for the effect of the start-stop sequence in the human ATF4 5' UTR on translation of uORF2 and uORF2 bypass under stress, we cloned the 5' UTR sequence of human ATF4 from the transcription start site to the start codon of uORF2 into the dual-luciferase psiCHECK2 plasmid. This was done to ensure that uORF2 was in the same reading frame with the reporter gene hluc+ (firefly luciferase), allowing us to use firefly luciferase activity as a measure of translation initiation at the uORF2 start codon. uORF2 bypass under stress could now be measured as a decrease in translation of the uORF2-firefly fusion construct. We used renilla luciferase activity as a readout of global translation levels which we expected to be reduced under stress.

We also used the same 5' UTR sequence, but mutated the AUG start codon of the start-stop to UUG to analyze the initiation activity at the uORF2 start codon in the absence of a start-stop. Lastly, we constructed reporter plasmids for the human ATF4 5' UTR with uORF1 and uORF2 start codons mutated to AGG in a similar manner, with the ATF4 5' UTR fused to the firefly reporter gene hluc+.

To determine bypass, we first normalized firefly luciferase and renilla luciferase activities by the transfection efficiency correction factors, and then calculated the change in normalized firefly in response to stress (i.e. bypass). We also corrected for the background change as measured by the normalized renilla luciferase activity.

Acknowledgments:

We would like to thank Dr. Neville Sanjana for providing HAP1 cells and lentiCRISPRv2 plasmid and helpful discussions regarding CRISPR experimental design.

This work was supported by the US National Institutes of Health 5R35GM127089 and Chan Zuckerberg Initiative (to CV); the Czech Science Foundation Grant of Excellence in Basic Research 19-25821X (to LSV); the FRQS (Fonds de Recherche du Québec – Santé) Senior Award (283444) (to IT); and The Swedish Research Council and the Wallenberg Academy Fellow program (to OL).

Author contributions:

Conceptualization: JR, SH, MH, IT, LSV, CV

Methodology: JR, SH, SW, MPM, JL, XG, HZ, SM, VH, AH, LSV, CV

Analysis: JR, SH, MPM, MP, SM, VH, ZG, AH, AL, KA, DS, WH, KS, OL

Visualization: JR, SH, SW, MPM, MP, ZG, AL, KS, CV

Funding: LSV, CV

Writing – original draft: JR, LSV, CV

Writing – review & editing: JR, SH, SW, MH, MP, IT, LSV, CV

Competing interests:

Authors declare no competing interests.

Supplementary Notes

Supplementary figures and notes

Supplementary Tables

Tables S1 to S10

Figures

Higher resolution figures are provided separately

Figure 1. Start-stop elements are widespread with distinct characteristics

A. The schematic illustrates the start-stop element in relationship to the bound ribosome. Ribosome footprints starting at the -12 position (P-site) have a typical length of 28-30 nt (nt). **B.** Scatter plot showing ribosome occupancy (RPF) vs. mRNA read depth for individual start-stops and control sequences in 5' UTRs. **C.** The heatmap dissects the 5' UTR to display the fraction occupied for all P-site/A-site combinations (64 x 64 = 4,096). Initiation P-sites, including non-canonical start codons, are highlighted with green arrows,while termination A-sites are denoted with red arrows. The AUGGCG hexamer is represented by the black box.

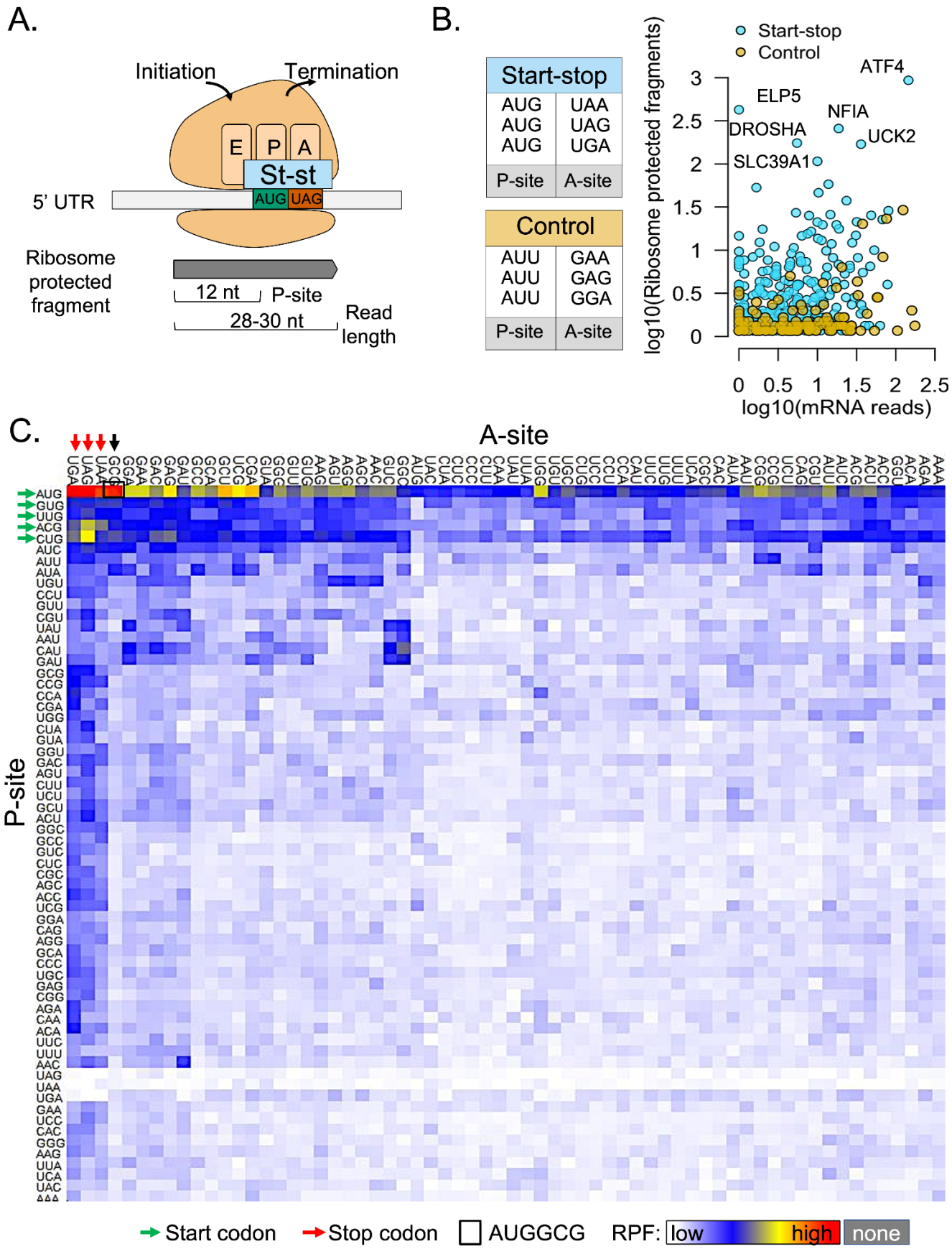

A.
Initiation
Termination
E
P
A
St-st
5' UTR
AUG
UAG
Ribosome protected fragment
12 nt
P-site
28-30 nt
Read length
B.
Start-stop
AUG UAA
AUG UAG
AUG UGA
P-site A-site
Control
AUU GAA
AUU GAG
AUU GGA
P-site A-site
Start-stop
Control
ATF4
ELP5
NFIA
DROSHA
UCK2
SLC39A1
log10(Ribosome protected fragments)
log10(mRNA reads)
C.
A-site
P-site
Start codon
Stop codon
AUGGCG
RPF:
low
high
none

Figure 2. Start-stop elements are commonly present with unique features

A. Flow chart showing the number of genes with start-stop or control sequences predicted across the entire human genome ('Genome wide'), with more than 3 ribosome footprint reads ('Ribosome occupied'), and with three or more ribosome footprint reads in both HEK293 and HeLa data ('Core set'). **B.** Translation efficiencies of genes containing a highly occupied start-stop or control sequences. Translation efficiency of CDS was calculated from ribosome profiling and RNA-seq data in HeLa cells [21]. **C**. Sequence conservation in vertebrates compared among highly occupied start-stops and controls, based on occupancy with ribosomes. Phylop-based conservation scorers were obtained using 100 vertebrate species, downloaded from the UCSC browser (https://genome.ucsc.edu). **D.** Relative frequencies of each nucleotide and combinations at positions ± 15 upstream and downstream of start-stops and controls. **E** Venn diagram showing overlap of start-stop-containing genes in HeLa [21], HEK293 [24], and mouse primary keratinocytes [28]. Human and mouse data were filtered for start-stop sequences with at least three and two RPF reads, respectively **F.** Data points in blue show the fraction of start-stops occupied by ribosomes based on positional information of P- and A-sites in RPFs for HeLa [21,24], HEK293[24], and mouse primary keratinocytes [28]. Occupancy is binned by 5' UTR, CDS, and 3' UTR. For comparison, control sequences are shown in yellow, and box plots show distribution of all possible hexamers (n = 4,096). **G.** Heatmap displaying enriched pathways for each cell type and the union.

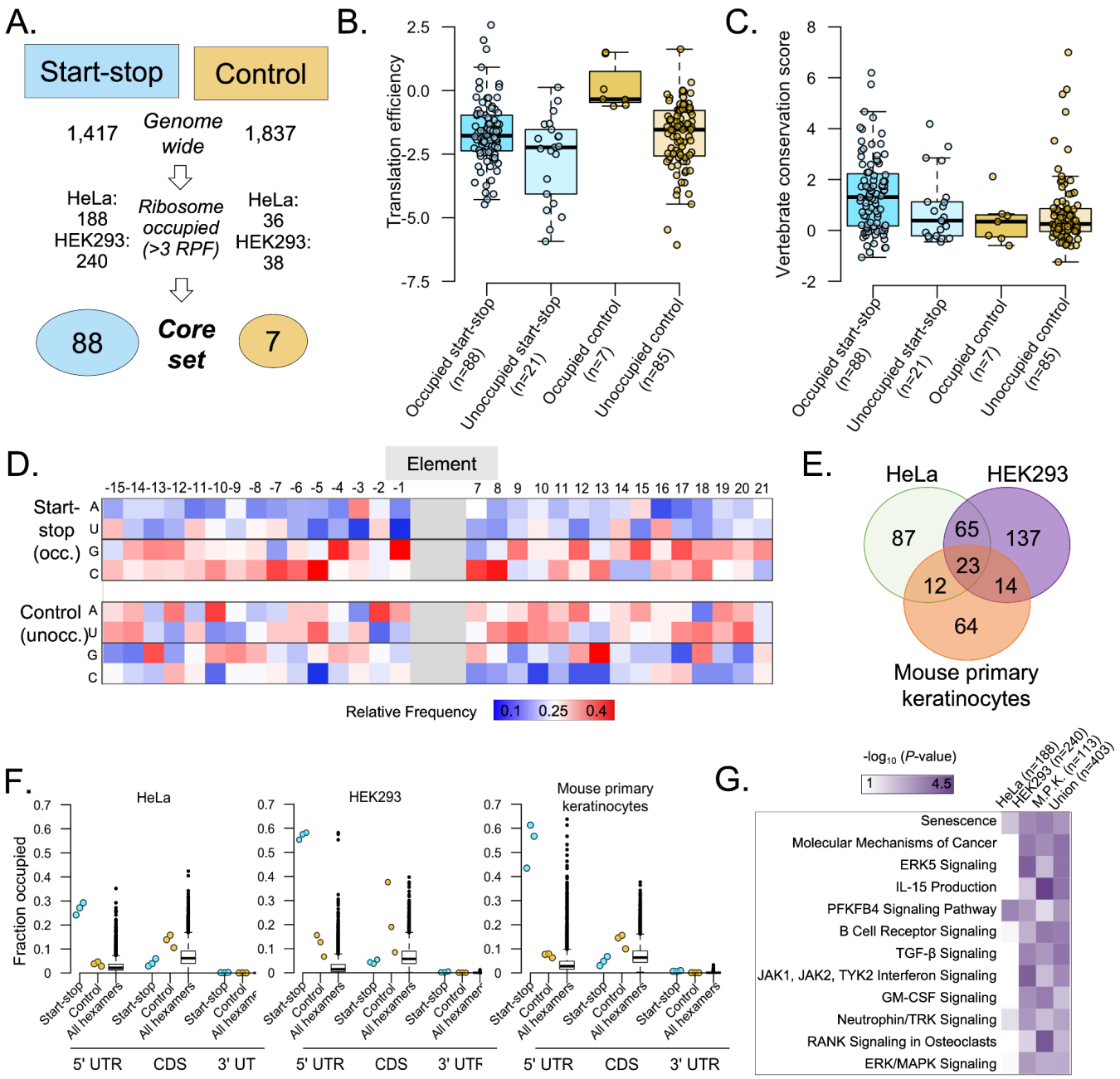

A.
Start-stop
Control
1,417
Genome wide
1,837
HeLa: 188
HEK293: 240
Ribosome occupied (>3 RPF)
HeLa: 36
HEK293: 38
88
Core set
7
B.
Translation efficiency
Occupied start-stop (n=88)
Unoccupied start-stop (n=21)
Occupied control (n=7)
Unoccupied control (n=85)
C.
Vertebrate conservation score
Occupied start-stop (n=88)
Unoccupied start-stop (n=21)
Occupied control (n=7)
Unoccupied control (n=85)
D.
Element
Start-stop (occ.)
Control (unocc.)
Relative Frequency
0.1
0.25
0.4
E.
HeLa
HEK293
87
65
137
23
12
14
64
Mouse primary keratinocytes
F.
HeLa
HEK293
Mouse primary keratinocytes
Fraction occupied
Start-stop
Control
All hexamers
5' UTR
CDS
3' UTR
G.
-log10 (P-value)
1
4.5
HeLa (n=188)
HEK293 (n=240)
M.P.K. (n=113)
Union (n=403)
Senescence
Molecular Mechanisms of Cancer
ERK5 Signaling
IL-15 Production
PFKFB4 Signaling Pathway
B Cell Receptor Signaling
TGF-β Signaling
JAK1, JAK2, TYK2 Interferon Signaling
GM-CSF Signaling
Neutrophin/TRK Signaling
RANK Signaling in Osteoclasts
ERK/MAPK Signaling

Figure 3. Start-stop elements occur in diverse 5' UTRs

A. Diagrams showing the relative t positioning (not to scale) of start-stops within 5' UTRs relative to uORFs for sample genes. **B.** Diagrams showcase sample genes with highly occupied start-stop sequences that only occur in some specific transcript variants, but not others. **C.** Schematic of MORF4L1 5' UTR luciferase constructs. Relative luciferase activity was measured in HEK293T and HeLa cells transfected with plasmids expressing luciferase fused to MORF4L1 5' UTR with wild-type start-stop sequence (WT) and AUGUAG->UUGUAG, respectively (St-st Mut). Measurements were based on three replicates. **D.** As panel **C.**, with IFNAR1 WT and AUGUAG->AUUGAG (St-st Mut) in HEK293T. **E.** As panel **C.**, with SLC39A1 WT and AUGUAG->UUGUAG (St-st Mut). **F.** As panel **C.**, with PSPC1 WT and AUGUAG->AUUGAG (St-st Mut). Diagrams are not drawn to scale. Plots show means ± standard error of the mean. * - *P-value* < 0.1; ** - *P-value* < 0.01; *** - *P-value* < 0.001. CDS - coding region; Lucif. - luciferase reporter.

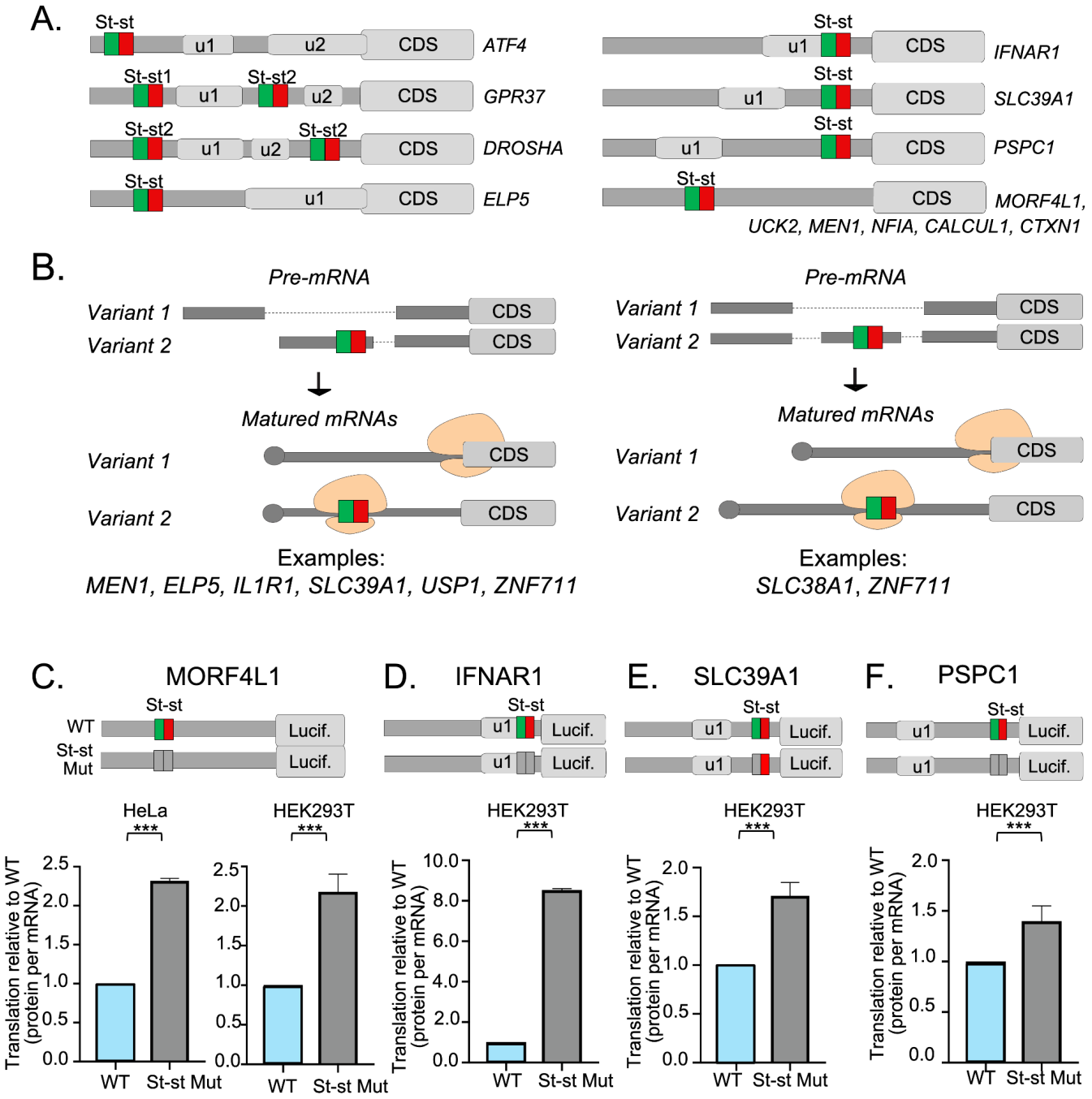

A.
St-st
u1
u2
CDS
ATF4
St-st1
St-st2
GPR37
DROSHA
ELP5
IFNAR1
SLC39A1
PSPC1
MORF4L1,
UCK2, MEN1, NFIA, CALCUL1, CTXN1
B.
Pre-mRNA
Variant 1
Variant 2
Matured mRNAs
Examples:
MEN1, ELP5, IL1R1, SLC39A1, USP1, ZNF711
SLC38A1, ZNF711
C.
MORF4L1
D.
IFNAR1
E.
SLC39A1
F.
PSPC1
WT
St-st Mut
Lucif.
HeLa
HEK293T
***
Translation relative to WT (protein per mRNA)
WT
St-st Mut

Figure 4. A Gain-of-function model provides insights into mechanistic aspects

A. Diagram explaining the gain-of-function construct used for the results shown in **B.** and **C.** We inserted either start-stop or 9 nt uORF (uORF) sequences into the ACTB 5' UTR used for the luciferase reporter, while the unmodified ACTB 5' UTR represents the wild-type. **B.** Translation efficiency for the different constructs shown in **A.** as calculated from the ratio of the protein level measured by luciferase assays and RNA level measured by RT-qPCR. All results are relative to wild-type (set to 1). **Figure S10** shows the protein measurements. **C.** RNA levels as measured by RT-qPCR for the different constructs shown in **A.** All results are relative to wild-type (set to 1). **D**. Average ribosome footprint counts at -12 position relative to AUG of start-stop and uORF, and AUU of the control in the 5' UTRs for the respective genes. We performed a t-test to determine the significance of the differences. **E**. Ratio of short vs. long read counts for ribosome footprints, corresponding to ribosomes with an empty and occupied A-site, respectively. Reads are counts at the -12 position relative to AUG for the start-stop, at the -12 and -9 positions for the initiation and termination codon of uORF, respectively, and at the -12 and -15 positions for the initiation and termination codon of main ORF (CDS, coding region), respectively, and at the -12 position relative to AUU for the control. We performed Fisher's exact test to determine the significance of the differences. **Figure S11** shows the underlying footprint data. Plots in **B.-F.** show means ± standard error of the mean. * - *P-value* < 0.1; ** - *P-value* < 0.01; *** - *P-value* < 0.001.

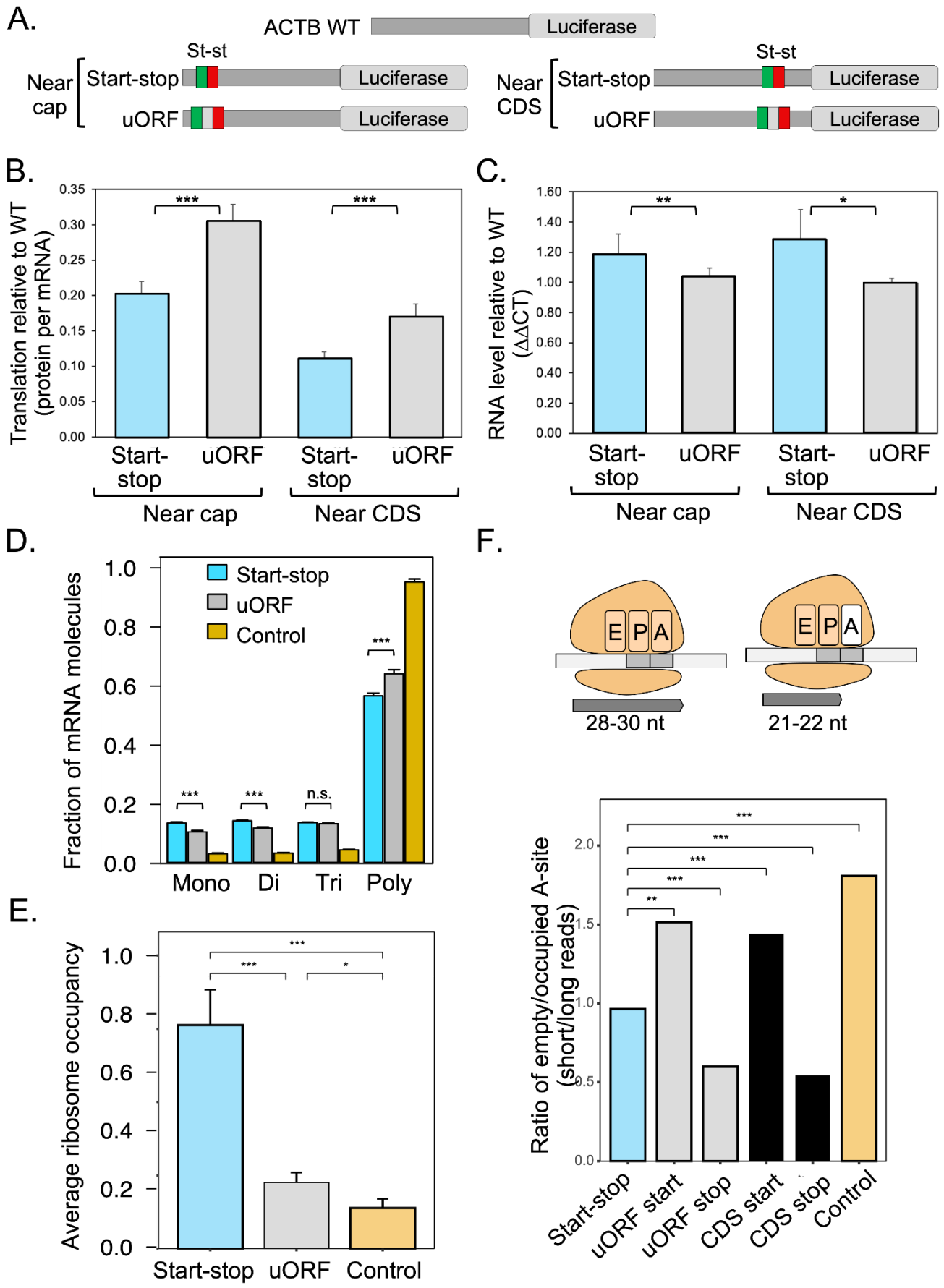
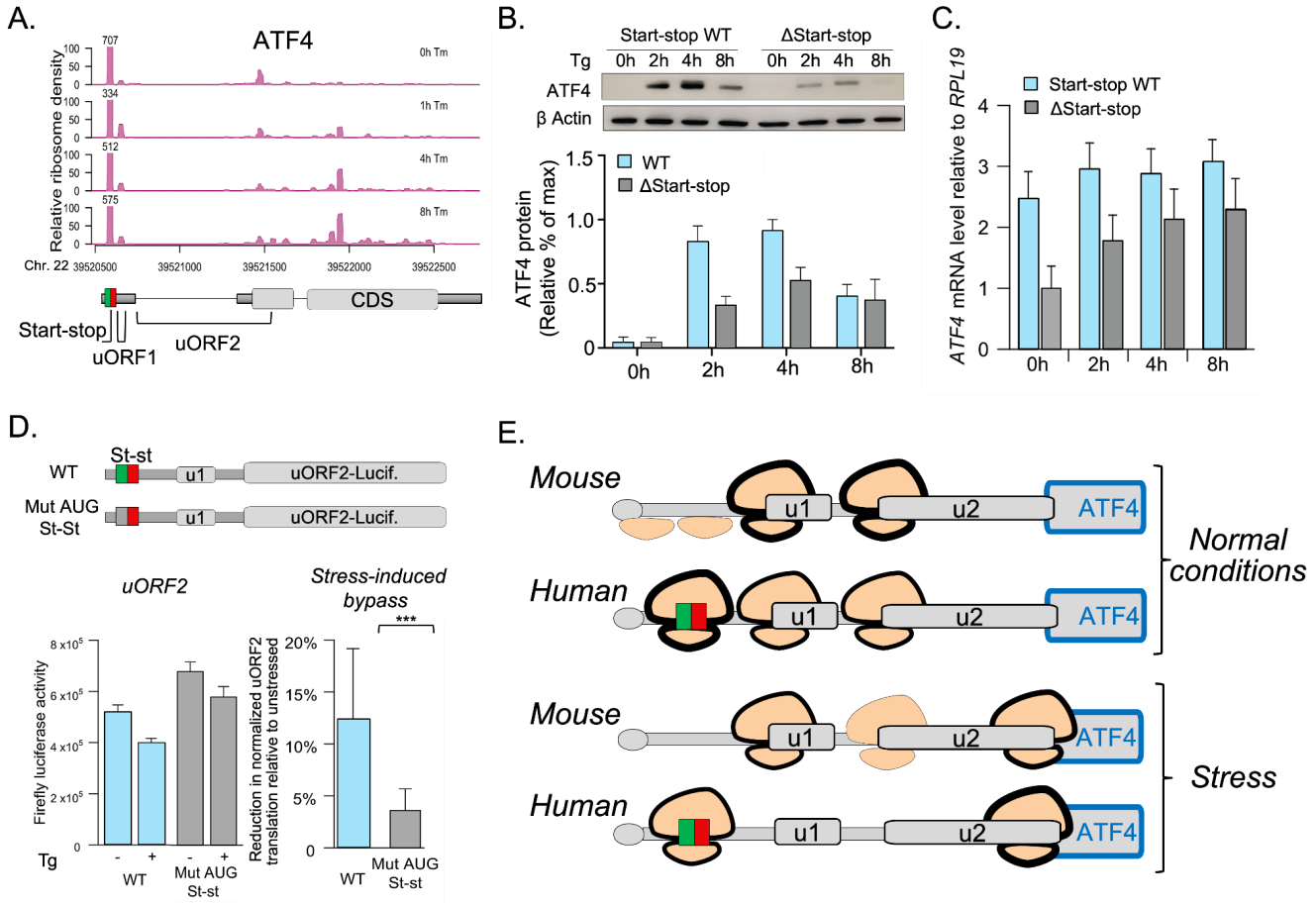

A.
ACTB WT
Luciferase
St-st
Near cap
Start-stop
uORF
Luciferase
Near CDS
B.
Translation relative to WT (protein per mRNA)
Start-stop
uORF
Near cap
Near CDS
C.
RNA level relative to WT (ΔΔCT)
D.
Fraction of mRNA molecules
Start-stop
uORF
Control
Mono
Di
Tri
Poly
n.s.
E.
Average ribosome occupancy
F.
E P A
28-30 nt
21-22 nt
Ratio of empty/occupied A-site (short/long reads)
Start-stop
uORF start
uORF stop
CDS start
CDS stop
Control

Figure 5. The start-stop in the human ATF4 gene supports the uORF-driven model of translation control

A. Relative ribosome density displayed as a function of position along the ATF4 gene. Profiles correspond to 0, 1, 4, and 8 hours post treatment with the stressor tunicamycin (Tm) in HeLa cells (25). The coordinates (GRCh38/hg38) along chromosome 22 are labeled; gene schematic indicates UTR (grey boxes), CDS (grey boxes), and introns (black lines). The start-stop and uORGFs are annotated. **B.** Representative Western blots of ATF4 and loading control for CRISPR-edited HAP1 cells before and after treatment with thapsigargin (Tg, 1uM). Below are ATF4 protein levels normalized to the loading control (% of max) obtained from three replicate measurements. **C.** ATF4 mRNA levels measured by qPCR in three replicate measurements. **D.** Diagrams of luciferase constructs. Bar plot on the left shows the firefly luciferase activity; bar plot on the right shows change in uORF2 bypass after Tg treatment. Values are firefly values of the fusion construct normalized the renilla luciferase measurements. **E.** Diagram for mouse and human ATF4 gene without and with a start-stop element (stst) upstream of the uORFs (u1, u2), supporting the uORF based regulatory model. The width of the ribosome border reflects the relative level of initiation at the respective start codons. CDS - coding region. * - *P-value* < 0.1; ** - *P-value* < 0.01; *** - *P-value* < 0.001.



References

1. Hinnebusch, A.G. (2017). Structural Insights into the Mechanism of Scanning and Start Codon Recognition in Eukaryotic Translation Initiation. Trends Biochem. Sci. *42*, 589–611.

2. Valášek, L.S., Zeman, J., Wagner, S., Beznosková, P., Pavlíková, Z., Mohammad, M.P., Hronová, V., Herrmannová, A., Hashem, Y., and Gunišová, S. (2017). Embraced by eIF3: structural and functional insights into the roles of eIF3 across the translation cycle. Nucleic Acids Res. *45*. 10.1093/nar/gkx805.

3. Valásek, L.S. (2012). "Ribozoomin"--translation initiation from the perspective of the ribosome-bound eukaryotic initiation factors (eIFs). Curr. Protein Pept. Sci. *13*. 10.2174/138920312801619385.

4. Hinnebusch, A.G., Ivanov, I.P., and Sonenberg, N. (2016). Translational control by 5'-untranslated regions of eukaryotic mRNAs. Science *352*, 1413–1416.

5. 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. U. S. A. *106*, 7507–7512.

6. Hinnebusch, A.G. (2005). Translational regulation of GCN4 and the general amino acid control of yeast. Annu. Rev. Microbiol. *59*, 407–450.

7. 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.

8. Vattem, K.M., and Wek, R.C. (2004). Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. Proc. Natl. Acad. Sci. U. S. A. *101*, 11269–11274.

9. Young, S.K., and Wek, R.C. (2016). Upstream Open Reading Frames Differentially Regulate Gene-specific Translation in the Integrated Stress Response. J. Biol. Chem. *291*, 16927–16935.

10. Gunišová, S., Hronová, V., Mohammad, M.P., Hinnebusch, A.G., and Valášek, L.S. (2018). Please do not recycle! Translation reinitiation in microbes and higher eukaryotes. FEMS Microbiol. Rev. *42*, 165–192.

11. Gunišová, S., and Valášek, L.S. (2014). Fail-safe mechanism of GCN4 translational control--uORF2 promotes reinitiation by analogous mechanism to uORF1 and thus secures its key role in GCN4 expression. Nucleic Acids Res. *42*, 5880–5893.

12. Wagner, S., Herrmannová, A., Hronová, V., Gunišová, S., Sen, N.D., Hannan, R.D., Hinnebusch, A.G., Shirokikh, N.E., Preiss, T., and Valášek, L.S. (2020). Selective Translation Complex Profiling Reveals Staged Initiation and Co-translational Assembly of Initiation Factor Complexes. Mol. Cell *79*, 546–560.e7.

13. Tanaka, M., Takano, J., Chiba, Y., Lombardo, F., Ogasawara, Y., Onouchi, H., Naito, S., and Fujiwara, T. (2011). Boron-dependent degradation of NIP5;1 mRNA for acclimation to excess boron conditions in Arabidopsis. Plant Cell *23*, 3547–3559.

14. Sotta, N., Chiba, Y., Miwa, K., Takamatsu, S., Tanaka, M., Yamashita, Y., Naito, S., and Fujiwara, T. (2021). Global analysis of boron-induced ribosome stalling reveals its effects on translation termination and unique regulation by AUG-stops in Arabidopsis shoots. Plant J. *106*. 10.1111/tpj.15248.

15. Tsednee, M., Tanaka, M., Kasai, K., and Fujiwara, T. (2020). Boron-dependent regulation of translation through AUGUAA sequence in yeast. Yeast *37*. 10.1002/yea.3526.

16. Tanaka, M., Sotta, N., Yamazumi, Y., Yamashita, Y., Miwa, K., Murota, K., Chiba, Y., Hirai, M.Y., Akiyama, T., Onouchi, H., et al. (2016). The Minimum Open Reading Frame, AUG-Stop, Induces Boron-Dependent Ribosome Stalling and mRNA Degradation. Plant Cell *28*. 10.1105/tpc.16.00481.

17. Alghoul, F., Laure, S., Eriani, G., and Martin, F. (2021). Translation inhibitory elements from and mRNAs use uORFs for translation inhibition. Elife *10*. 10.7554/eLife.66369.

18. Schleich, S., Acevedo, J.M., Clemm von Hohenberg, K., and Teleman, A.A. (2017). Identification of transcripts with short stuORFs as targets for DENR•MCTS1-dependent translation in human cells. Sci. Rep. *7*, 3722.

19. Bohlen, J., Harbrecht, L., Blanco, S., Clemm von Hohenberg, K., Fenzl, K., Kramer, G., Bukau, B., and Teleman, A.A. (2020). DENR promotes translation reinitiation via ribosome recycling to drive expression of oncogenes including ATF4. Nat. Commun. *11*, 4676.

20. Harding, H.P., Novoa, I., Zhang, Y., Zeng, H., Wek, R., Schapira, M., and Ron, D. (2000). Regulated translation initiation controls stress-induced gene expression in mammalian cells. Mol. Cell *6*. 10.1016/s1097-2765(00)00108-8.

21. Rendleman, J., Cheng, Z., Maity, S., Kastelic, N., Munschauer, M., Allgoewer, K., Teo, G., Zhang, Y.B.M., Lei, A., Parker, B., et al. (2018). New insights into the cellular temporal response to proteostatic stress. Elife *7*. 10.7554/eLife.39054.

22. Nakagawa, S., Niimura, Y., Gojobori, T., Tanaka, H., and Miura, K.-I. (2008). Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. Nucleic Acids Res. *36*, 861–871.

23. Niimura, Y., Terabe, M., Gojobori, T., and Miura, K.-I. (2003). Comparative analysis of the base biases at the gene terminal portions in seven eukaryote genomes. Nucleic Acids Res. *31*, 5195–5201.

24. Calviello, L., Mukherjee, N., Wyler, E., Zauber, H., Hirsekorn, A., Selbach, M., Landthaler, M., Obermayer, B., and Ohler, U. (2016). Detecting actively translated open reading frames in ribosome profiling data. Nat. Methods *13*, 165–170.

25. Gandin, V., Masvidal, L., Hulea, L., Gravel, S.-P., Cargnello, M., McLaughlan, S., Cai, Y., Balanathan, P., Morita, M., Rajakumar, A., et al. (2016). nanoCAGE reveals 5' UTR features that define specific modes of translation of functionally related MTOR-sensitive mRNAs. Genome Res. *26*, 636–648.

26. Chan, C.Y., Carmack, C.S., Long, D.D., Maliyekkel, A., Shao, Y., Roninson, I.B., and Ding, Y. (2009). A structural interpretation of the effect of GC-content on efficiency of RNA interference. BMC Bioinformatics *10 Suppl 1*, S33.

27. Tzani, I., Ivanov, I.P., Andreev, D.E., Dmitriev, R.I., Dean, K.A., Baranov, P.V., Atkins, J.F., and Loughran, G. (2016). Systematic analysis of the PTEN 5' leader identifies a major AUU initiated proteoform. Open Biol. *6*. 10.1098/rsob.150203.

28. Sendoel, A., Dunn, J.G., Rodriguez, E.H., Naik, S., Gomez, N.C., Hurwitz, B., Levorse, J., Dill, B.D., Schramek, D., Molina, H., et al. (2017). Translation from unconventional 5' start sites drives tumour initiation. Nature *541*, 494–499.

29. Resch, A.M., Ogurtsov, A.Y., Rogozin, I.B., Shabalina, S.A., and Koonin, E.V. (2009). Evolution of alternative and constitutive regions of mammalian 5'UTRs. BMC Genomics *10*, 162.

30. Zhou, Z., Chen, Y., Ni, W., and Liu, T. (2019). Upregulation of Nuclear Factor IA Suppresses Oxidized

Low-Density Lipoprotein-Induced Endoplasmic Reticulum Stress and Apoptosis in Human Umbilical Vein Endothelial Cells. Med. Sci. Monit. *25*, 1009.

31. Deneen, B., Ho, R., Lukaszewicz, A., Hochstim, C.J., Gronostajski, R.M., and Anderson, D.J. (2006). The transcription factor NFIA controls the onset of gliogenesis in the developing spinal cord. Neuron *52*. 10.1016/j.neuron.2006.11.019.

32. Bergé-Lefranc, J.L., Jay, P., Massacrier, A., Cau, P., Mattei, M.G., Bauer, S., Marsollier, C., Berta, P., and Fontes, M. (1996). Characterization of the human jumonji gene. Hum. Mol. Genet. *5*. 10.1093/hmg/5.10.1637.

33. Afrakhte, M., Morén, A., Jossan, S., Itoh, S., Sampath, K., Westermark, B., Heldin, C.H., Heldin, N.E., and ten Dijke, P. (1998). Induction of inhibitory Smad6 and Smad7 mRNA by TGF-beta family members. Biochem. Biophys. Res. Commun. *249*. 10.1006/bbrc.1998.9170.

34. Stolfi, C., Marafini, I., De Simone, V., Pallone, F., and Monteleone, G. (2013). The Dual Role of Smad7 in the Control of Cancer Growth and Metastasis. Int. J. Mol. Sci. *14*, 23774.

35. Sang, Y., Zhang, R., Sun, L., Chen, K.K., Li, S.-W., Xiong, L., Peng, Y., Zeng, L., and Huang, G. (2019). MORF4L1 suppresses cell proliferation, migration and invasion by increasing p21 and E-cadherin expression in nasopharyngeal carcinoma. Oncol. Lett. *17*, 294–302.

36. Platanias, L.C. (2005). Mechanisms of type-I- and type-II-interferon-mediated signalling. Nat. Rev. Immunol. *5*, 375–386.

37. Eide, D.J. (2004). The SLC39 family of metal ion transporters. Pflugers Arch. *447*. 10.1007/s00424-003-1074-3.

38. Franklin, R.B., Feng, P., Milon, B., Desouki, M.M., Singh, K.K., Kajdacsy-Balla, A., Bagasra, O., and Costello, L.C. (2005). hZIP1 zinc uptake transporter down regulation and zinc depletion in prostate cancer. Mol. Cancer *4*. 10.1186/1476-4598-4-32.

39. Mohankumar, K., Shrestha, R., and Safe, S. (2022). Nuclear receptor 4A1 (NR4A1) antagonists target paraspeckle component 1 (PSPC1) in cancer cells. Mol. Carcinog. *61*. 10.1002/mc.23362.

40. Sample, P.J., Wang, B., Reid, D.W., Presnyak, V., McFadyen, I.J., Morris, D.R., and Seelig, G. (2019). Human 5' UTR design and variant effect prediction from a massively parallel translation assay. Nat. Biotechnol. *37*, 803–809.

41. Wu, C.C.-C., Zinshteyn, B., Wehner, K.A., and Green, R. (2019). High-Resolution Ribosome Profiling Defines Discrete Ribosome Elongation States and Translational Regulation during Cellular Stress. Mol. Cell *73*, 959–970.e5.

42. Andreev, D.E., O'Connor, P.B.F., Fahey, C., Kenny, E.M., Terenin, I.M., Dmitriev, S.E., Cormican, P., Morris, D.W., Shatsky, I.N., and Baranov, P.V. (2015). Translation of 5′ leaders is pervasive in genes resistant to eIF2 repression. 10.7554/eLife.03971.

43. Ichihara, K., Matsumoto, A., Nishida, H., Kito, Y., Shimizu, H., Shichino, Y., Iwasaki, S., Imami, K., Ishihama, Y., and Nakayama, K.I. (2021). Combinatorial analysis of translation dynamics reveals eIF2 dependence of translation initiation at near-cognate codons. Nucleic Acids Res. *49*. 10.1093/nar/gkab549.

44. Smirnova, A.M., Hronova, V., Gunisova, S., Mohammad, M.P., Petrackova, D., Halada, P., Rendelman, J., Coufal, S., Herrmannova, A., Beznoskova, P., et al. (2023). Stem-loop induced ribosome queuing in the uORF2/ATF4 overlap fine-tunes stress-induced human ATF4 translational control. bioRxiv, 2023.07.12.548609. 10.1101/2023.07.12.548609.

45. Vasudevan, D., Neuman, S.D., Yang, A., Lough, L., Brown, B., Bashirullah, A., Cardozo, T., and Ryoo, H.D.

(2020). Translational induction of ATF4 during integrated stress response requires noncanonical initiation factors eIF2D and DENR. Nat. Commun. *11*, 4677.

46. Castelo-Szekely, V., De Matos, M., Tusup, M., Pascolo, S., Ule, J., and Gatfield, D. (2019). Charting DENR-dependent translation reinitiation uncovers predictive uORF features and links to circadian timekeeping via Clock. Nucleic Acids Res. *47*, 5193–5209.

47. Gardner, L.B. (2008). Hypoxic inhibition of nonsense-mediated RNA decay regulates gene expression and the integrated stress response. Mol. Cell. Biol. *28*, 3729–3741.

48. Archer, S.K., Shirokikh, N.E., Beilharz, T.H., and Preiss, T. (2016). Dynamics of ribosome scanning and recycling revealed by translation complex profiling. Nature *535*. 10.1038/nature18647.

49. Lawson, M.R., Lessen, L.N., Wang, J., Prabhakar, A., Corsepius, N.C., Green, R., and Puglisi, J.D. (2021). Mechanisms that ensure speed and fidelity in eukaryotic translation termination. Science *373*. 10.1126/science.abi7801.

50. Ahmed, Y.L., Schleich, S., Bohlen, J., Mandel, N., Simon, B., Sinning, I., and Teleman, A.A. (2018). DENR–MCTS1 heterodimerization and tRNA recruitment are required for translation reinitiation. PLoS Biol. *16*. 10.1371/journal.pbio.2005160.

51. Skabkin, M.A., Skabkina, O.V., Dhote, V., Komar, A.A., Hellen, C.U.T., and Pestova, T.V. (2010). Activities of Ligatin and MCT-1/DENR in eukaryotic translation initiation and ribosomal recycling. Genes Dev. *24*, 1787–1801.

52. Khong, A., Matheny, T., Jain, S., Mitchell, S.F., Wheeler, J.R., and Parker, R. (2017). The Stress Granule Transcriptome Reveals Principles of mRNA Accumulation in Stress Granules. Mol. Cell *68*. 10.1016/j.molcel.2017.10.015.

53. Lee, D., and Shin, C. (2018). Emerging roles of DROSHA beyond primary microRNA processing. RNA Biol. *15*, 186.

54. García-Santisteban, I., Peters, G.J., Giovannetti, E., and Rodríguez, J.A. (2013). USP1 deubiquitinase: cellular functions, regulatory mechanisms and emerging potential as target in cancer therapy. Mol. Cancer *12*, 1–12.

55. Kleine-Kohlbrecher, D., Christensen, J., Vandamme, J., Abarrategui, I., Bak, M., Tommerup, N., Shi, X., Gozani, O., Rappsilber, J., Salcini, A.E., et al. (2010). A functional link between the histone demethylase PHF8 and the transcription factor ZNF711 in X-linked mental retardation. Mol. Cell *38*. 10.1016/j.molcel.2010.03.002.

56. Wang, J., Foroutan, A., Richardson, E., Skinner, S.A., Reilly, J., Kerkhof, J., Curry, C.J., Tarpey, P.S., Robertson, S.P., Maystadt, I., et al. (2022). Clinical findings and a DNA methylation signature in kindreds with alterations in ZNF711. Eur. J. Hum. Genet. *30*, 420–427.

57. Han, T., Xiang, D.M., Sun, W., Liu, N., Sun, H.L., Wen, W., Shen, W.F., Wang, R.Y., Chen, C., Wang, X., et al. (2015). PTPN11/Shp2 overexpression enhances liver cancer progression and predicts poor prognosis of patients. J. Hepatol. *63*. 10.1016/j.jhep.2015.03.036.

58. Pham, C., Fong, T.-L., Zhang, J., and Liu, L. (2018). Striking Racial/Ethnic Disparities in Liver Cancer Incidence Rates and Temporal Trends in California, 1988–2012. JNCI Journal of the National Cancer Institute *110*, 1259.

59. Smith, L.K., Parmenter, T., Kleinschmidt, M., Kusnadi, E.P., Kang, J., Martin, C.A., Lau, P., Patel, R., Lorent, J., Papadopoli, D., et al. (2022). Adaptive translational reprogramming of metabolism limits the response to targeted therapy in BRAFV600 melanoma. Nat. Commun. *13*. 10.1038/s41467-022-28705-x.

60. Gao, X., Bao, W., Bai, J., Fan, K., Li, L., and Li, Y. (2022). UHMK1 aids colorectal cancer cell proliferation and

chemoresistance through augmenting IL-6/STAT3 signaling. Cell Death Dis. *13*, 1–11.

61. Augustus, G.J., and Ellis, N.A. (2018). Colorectal Cancer Disparity in African Americans: Risk Factors and Carcinogenic Mechanisms. Am. J. Pathol. *188*, 291.

62. Brogna, S., and Wen, J. (2009). Nonsense-mediated mRNA decay (NMD) mechanisms. Nat. Struct. Mol. Biol. *16*, 107–113.

63. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10–12.

64. Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. *37*, 907–915.

65. Harnett, D., Meerdink, E., Calviello, L., Sydow, D., and Ohler, U. (2021). Genome-Wide Analysis of Actively Translated Open Reading Frames Using RiboTaper/ORFquant. Methods Mol. Biol. *2252*, 331–346.

66. Putri, G.H., Anders, S., Pyl, P.T., Pimanda, J.E., and Zanini, F. (2022). Analysing high-throughput sequencing data in Python with HTSeq 2.0. Bioinformatics *38*, 2943–2945.

67. Brar, G.A., Yassour, M., Friedman, N., Regev, A., Ingolia, N.T., and Weissman, J.S. (2012). High-resolution view of the yeast meiotic program revealed by ribosome profiling. Science *335*, 552–557.

68. Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M.Y., Geiger, T., Mann, M., and Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat. Methods *13*, 731–740.

69. Krämer, A., Green, J., Pollard, J., Jr, and Tugendreich, S. (2014). Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics *30*, 523–530.

70. Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Series B Stat. Methodol. *57*, 289–300.

71. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

72. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

73. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

74. Doench, J.G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E.W., Donovan, K.F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., et al. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat. Biotechnol. *34*, 184–191.

75. Sanjana, N.E., Shalem, O., and Zhang, F. (2014). Improved vectors and genome-wide libraries for CRISPR screening. Nat. Methods *11*, 783–784.