bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453809; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

**Authors:** Justin Rendleman[1,2*], 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.

*Corresponding author: cvogel@nyu.edu

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453809; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

**Abstract:**

Translation includes initiation, elongation, and termination, followed by ribosome recycling. We characterize a new sequence element in 5' untranslated regions that consists of an adjacent start and stop codon and thereby excludes elongation. In these start-stop elements, an initiating ribosome is simultaneously positioned for termination without having translocated. At the example of activating transcription factor 4 (*ATF4*), we demonstrate that start-stops modify downstream re-initiation, thereby repressing translation of upstream open reading frames and enhancing ATF4's inducibility under stress. Start-stop elements are abundant in both mammals and yeast and affect key regulators such as *DROSHA* and the oncogenic transcription factor *NFIA*. They provide a unique regulatory layer that impedes ribosome scanning without the energy-expensive peptide production that accompanies upstream open reading frames.

**One-Sentence Summary:**

Regulatory start-stop elements in the 5'UTR are hitherto unappreciated contributors to 5' UTR code and alter the translatome.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453809; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

## Main Text:

Most mRNA translation occurs in a 5' cap-dependent manner: the 43S preinitiation complex (PIC) is recruited to the 5' cap via the eukaryotic translation initiation factor 4F (eIF4F), which facilitates subsequent scanning of the PIC across the 5' untranslated region (UTR) of an mRNA until a start codon is recognized and translation is initiated (*1, 2*). As such, genetic information written into the 5' UTR must be scanned through, including translational control elements that influence protein production, oftentimes in a condition dependent manner. However, understanding of the 5' UTR regulatory code is in its infancy (*3*). Upstream open reading frames (uORFs) are one prevalent regulatory element found in 5' UTRs, with approximately half of human mRNAs containing putative uORFs. Functional uORFs undergo standard translation, including initiation, elongation of the uORF peptide, and termination. uORFs often act to repress translation of downstream coding regions (*4*), although more complex cases exist (*5-11*). In addition to uORFs, recent literature has also revealed instances in the 5' UTR in which an AUG is immediately followed by one of the three stop codons (*12-14*); we designate these elements start-stops. Importantly, it has to-date escaped attention that these start-stops do not represent open reading frames like uORFs, as they preclude the elongation step in translation. No translocation can take place at such start-stop elements and ribosomes are in a state of simultaneous initiation and termination, with their *modus operandi* and regulatory potential virtually unknown.

Here, we addressed this pivotal gap in knowledge and show that start-stops are not only uniquely functional, but also highly conserved regulatory elements that affect the expression of pivotal genes within signaling pathways, many whose dysregulation is linked to disease. We first characterized one specific start-stop element in the 5' UTR of the *ATF4* gene, a master regulator of the unfolded protein response (*6-8, 15*). Translation of *ATF4* mRNA is induced in response to a variety of stresses through a system of two coupled uORFs (fig. S1). The existence of a start-stop upstream of uORF1 in the human *ATF4* transcript had been noticed (*14, 16, 17*), but it had been neither recognized as having a distinct nature from classical uORFs, nor investigated with respect to its role in ATF4 regulation. We show that the start-stop expands the ATF4 regulatory circuit to increase sensitivity to, and efficiency of, translation induction under stress, while reducing wasteful production of peptides resulting from uORF translation. We then showed that start-stops are widespread and highly conserved regulatory elements that affect many genes with roles in signaling, through a unique mechanism that impedes ribosomal scanning.

## Results

### *A start-stop element in the 5' UTR of the human ATF4 gene serves as a ribosome anchor*

*ATF4* is the quintessential example of uORF-mediated translation induction in metazoans (*3, 9*), where two sequential uORFs repress translation under normal conditions and induce translation in response to various stresses, through altered bypass of the second uORF. When examining profiles of ribosome occupancy along the *ATF4* transcript in HeLa cells responding to protein folding stress (Fig. 1A), we observed ribosomes at both uORF1 and uORF2 and, expectedly, at elevated levels along the coding sequence (CDS) during stress (fig. S1). Unexpectedly, the highest ribosome coverage localized to a region upstream of uORF1 (Fig. 1A), centered around a six-nucleotide sequence comprised of an adjacent start and stop codon (start-stop). Ribosome binding to this start-stop site was similar in data from HEK293, HAP1, and iPSC cell lines (*18-20*), confirming that the effect was not cell-type specific (fig. S2). It was also independent of the use of cycloheximide during sample collection (fig. S2), and thus does not represent an artifact associated with use of this translation inhibitor (*21*).

Using positional information extracted from ribosome protected fragments (RPFs), we determined that the ribosomes were situated on start-stops with their P-site at the start codon (Fig. 1B). While this is typical of initiating ribosomes, this unique orientation positions the A-site simultaneously at a stop codon, rendering these ribosomes incapable of translocating (Fig. 1B). Instead, the combination of initiation and termination excludes elongation, halts ribosomes at the site, and prevents further scanning of the 5' UTR. This feature is distinct from canonical open reading frames that include elongation and translate a sequence into a peptide. Recent evidence has shown that the small ribosomal subunit (SSU) can linger after termination before continuing to scan downstream (*11, 22*), providing a possible mechanism behind the function of start-stops. In the context of a start-stop, the SSU would be positioned

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453809; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

with the P-site at the start codon even after termination, therefore allowing for multiple rounds of initiation, termination, and 60S recycling.

The *ATF4* start-stop element accounted for the majority of ribosome-RNA interactions compared to other uORFs or the CDS (Fig. 1C), leading us to speculate that it has a role as a ribosome anchor within the mRNA, analogous to poised RNA polymerase II in the promoter regions of many developmental control genes (*23*). Further, ribosome occupancy at the start-stop remained high, regardless of stress (Fig. 1C). This continuous ribosome binding suggested that scanning ribosomes were still impeded by the start-stop even when eIF2α was phosphorylated in response to stress, leading to attenuation of ternary complex recycling. We found negative, but non-significant, correlation between stress and relative ribosome occupancy at the start-stop and uORF2, which was in stark contrast to significant positive correlation for the CDS (Fig. 1D). These findings led us to postulate a model whereby small changes in uORF2 bypass, mediated by the start-stop, would lead to large increases in CDS translation (see below, supplementary text, and fig. S3).

The start-stop in the 5'UTR of *ATF4* mRNA is more conserved across vertebrates than the surrounding sequence, particularly in mammals (Fig. 1E). In some organisms, e.g. dolphin, killer whale, and elephant, the start-stop contains an AUG and a different stop codon, suggesting an evolutionary constraint on the functionality rather than the precise sequence. The start-stop is notably absent from the mouse genome, where the model of ATF4 translation induction was originally discovered (*6, 7*). This absence suggested that the start-stop may not be essential to establish the stress-induced translation of ATF4, but instead provides additional levels of regulation, which we explored below.

***ATF4's start-stop represses basal translation of uORFs and affects main ORF inducibility under stress***

We hypothesized that the presence of the start-stop in human 5' UTR impacts the *ATF4* system in two ways: 1) reducing translation of uORF1 and, consequently, reinitiation at uORF2 during normal conditions, while maintaining a high degree of suppression at the CDS, and 2) increasing the sensitivity of *ATF4* mRNA translation induction during stress. We propose that the former improves cellular economy, as constitutive synthesis of the uORF2 peptide is costly, while we anticipate the latter to enhance stress-adaptability of the cell. Based on recent studies, we speculated that these functions are realized through noncanonical initiation factors eIF2D, DENR, and MCTS1 which enable reinitiation after the uORFs and the start-stop (*14, 24*).

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453809; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

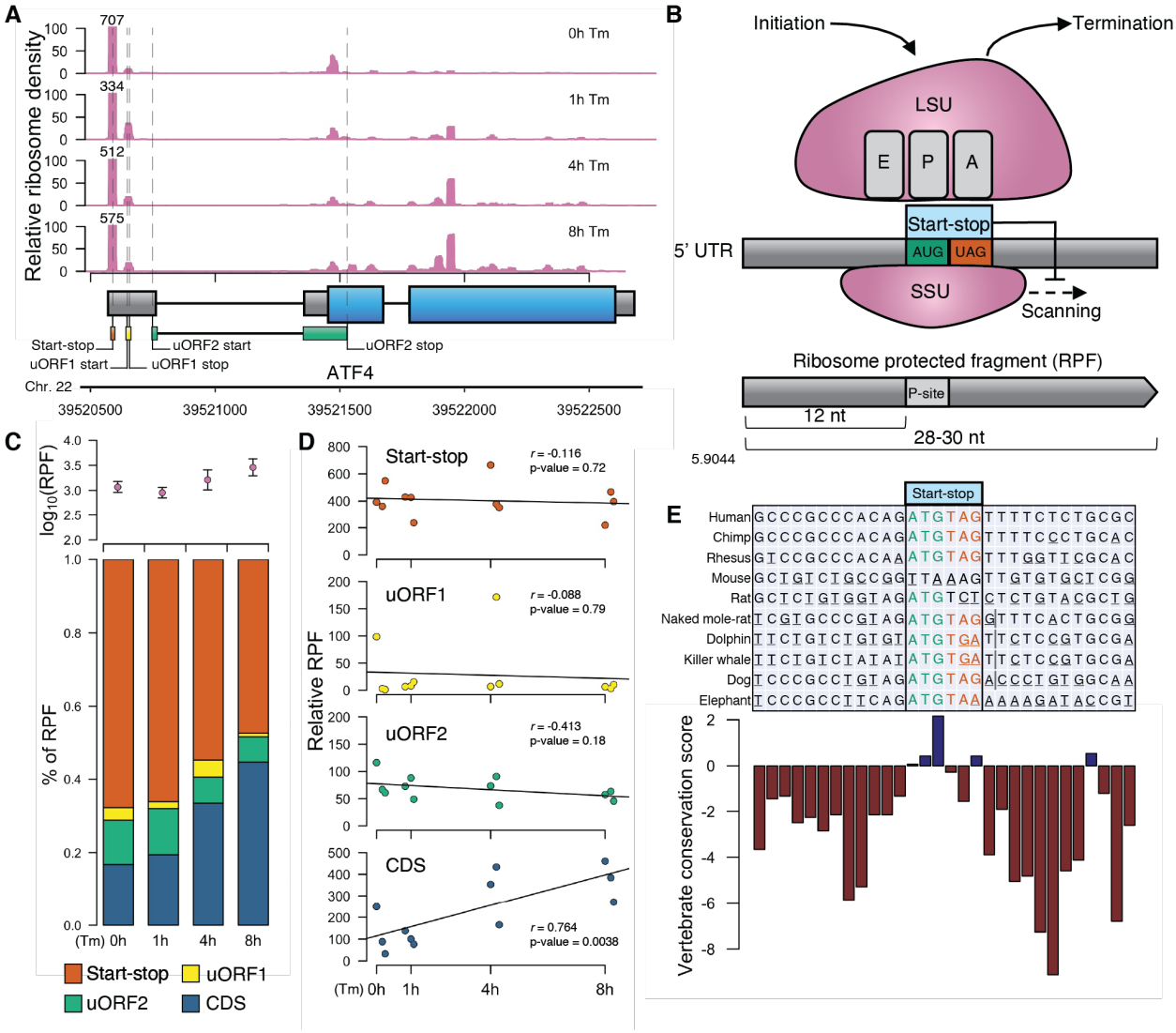



**Fig. 1. Adjacent start and stop codons block ribosomes in the 5' UTR of *ATF4*.**
(**A**) Relative ribosome density displayed as a function of position along *ATF4*. Profiles correspond to 0, 1, 4, and 8 hours post treatment with the ER stressor tunicamycin (Tm) in HeLa cells (*25*). The coordinates (GRCh38/hg38) along chromosome 22 are labeled; gene schematic indicates UTR (grey boxes), CDS (blue boxes), and introns (black lines). The start-stop and uORFs are annotated. (**B**) Schematic of ribosomes at the start-stop element. Ribosome protected fragments (RPFs) align with the P-site at the start codon. Since the A-site occupies a stop codon, elongation is blocked and termination follows. (**C**) Top panel shows RPF coverage for 0, 1, 4, and 8 hours post Tm treatment. Means ± SEM; **P* < 0.05; t-test. Lower panel shows fraction of RPFs according to the region of transcript where the P-site mapped. (**D**) Normalized RPF values for each region were used to fit linear models; Pearson correlation coefficients (*r*) and correlation test p-values are shown. (**E**) Genomic alignments of ten mammalian species are shown for the start-stop element and surrounding sequence. Sequence differences with the human genome are underlined; gaps are denoted by a vertical line. Below each nucleotide we display the phyloP-based vertebrate conservation scores. Abbreviation: SSU, small ribosomal subunit; LSU, large ribosomal subunit.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453809; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To determine the start-stop's role in human *ATF4* mRNA translation, we used CRISPR-Cas9 genome editing to generate HAP1 cell lines that lacked the start-stop by deleting the entire sequence (ΔStart-stop) or by mutating the start codon (Start-stop$^{TAGTAG}$) and compared these to a line with an intact start-stop (Start-stop$^{WT}$)(Fig. 2A)(*26*). We subjected CRISPR-edited cells to a mild dose (1μM) of thapsigargin, an inhibitor of the sarco/endoplasmic reticulum $Ca^{2+}$ ATPase, to induce acute ER stress and activate the integrated stress response. This treatment was sufficient to trigger a stress response, as determined by measurements of spliced *XBP1* (*27*)(fig. S4). *XBP1* splicing was similar between ΔStart-stop and Start-stop$^{WT}$ cells, indicating the IRE1 branch of the unfolded protein response (*28*) was unaffected by mutations at the *ATF4* start-stop.

Next, we evaluated the impact of the start-stop on the sensitivity of translation induction and measured ATF4 protein levels at 0, 2, 4, and 8 hours after treatment with thapsigargin. As shown by Western blotting, ATF4 protein levels were low in control conditions (0 hours), independently of the start-stop status (Fig. 2B). All three cell lines displayed an acute response to thapsigargin treatment, with ATF4 protein levels increasing 2 and 4 hours post exposure, followed by a decrease at the 8-hour timepoint. Notably, ATF4 induction in both mutant lines was dramatically attenuated (ΔStart-stop and Start-stop$^{TAGTAG}$), as compared to Start-stop$^{WT}$ cells. These results suggested that the start-stop cooperates with uORFs to confer greater control of ATF4 protein production.

Further, we also found the start-stop element rendered ATF4 translation more efficient with respect to stabilizing the *ATF4* mRNA and reducing translation of uORF2's peptide. Constitutive translation of the uORF2 peptide (59 amino acids in length) is a wasteful byproduct of the *ATF4* system. We hypothesized that by anchoring ribosomes early on as they scan the transcript, the start-stop could improve the cellular economy through reduced initiation at uORF1 and thus attenuated reinitiation at uORF2 under normal conditions. Previous work has shown that translation of uORF2 triggers *ATF4* mRNA degradation via nonsense-mediated mRNA decay (NMD)(*29*); therefore, if the start-stop lowered uORF2 translation under normal conditions, it should also attenuate NMD-related degradation resulting in higher *ATF4* mRNA levels. Indeed, we found that *ATF4* mRNA expression levels were elevated in Start-stop$^{WT}$ cells compared to ΔStart-stop (Fig. 2C). Control measurements of 18S rRNA levels showed no differences (fig. S5). Both Start-stop$^{WT}$ and ΔStart-stop cells exhibited increased *ATF4* mRNA levels after treatment with thapsigargin, indicating that transcriptional induction was unaffected by the start-stop mutation. Furthermore, following treatment with the RNA polymerase II inhibitor 5,6-dichloro-1-β-ribofuranosyl-benzimidazole, *ATF4* mRNA levels declined less rapidly in Start-stop$^{WT}$ compared to ΔStart-stop cells (Fig. 2D)(*29*). Using these results, we calculated mRNA half-lives, and found the *ATF4* mRNA with the Start-stop$^{WT}$ to be significantly longer lived than the mRNA with the ΔStart-stop (Fig. 2D), consistent with our analysis of steady state mRNA levels.

Next, we established a direct link between the start-stop and downstream reinitiation at uORF2, underlining the start-stop's role as an additional layer of uORF-based regulation. We designed reporters with human *ATF4* 5' UTR up to the uORF2 start codon fused to firefly luciferase coding sequence; in this way, we used firefly luciferase activity as a readout of reinitiation at uORF2 (Fig. 2E). Treatment with thapsigargin lowered reporter activity, consistent with increased bypass of uORF2 during ER stress (Fig. 2E). Mutating the AUG of the start-stop to a UUG resulted in higher basal expression of the reporter compared to the WT sequence, with lower expression of the reporter under stress (Fig. 2E). Activity of Renilla luciferase, which served as an internal control and was expressed from the same plasmid as firefly luciferase, did not differ between cells transfected with either construct, and decreased with stress due to the global decline in initiation during the stress response, though to a lesser degree (fig. S6).

***Quantitative modeling illustrates the start-stop's role in regulatory efficiency and sensitivity***

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453809; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

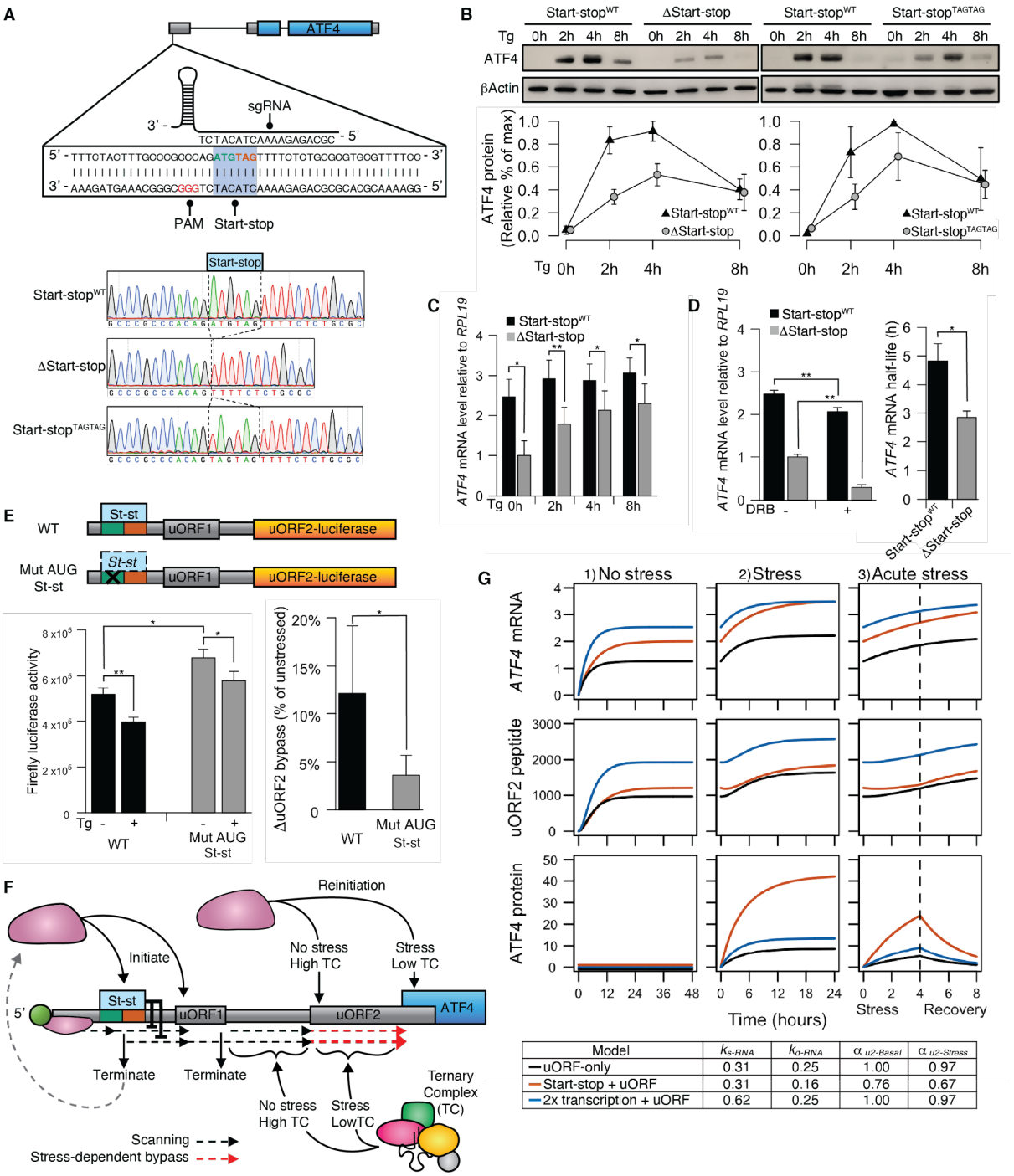
To quantify the impact of the start-stop on the reinitiation of uORF2 during stress independently of global changes in initiation associated with eIF2α phosphorylation, we calculated the change in uORF2 bypass as:

$$\Delta uORF2\ bypass = \Delta uORF2\ initiation - \Delta global\ initiation$$

We used the pre- and post-stress firefly luciferase values to calculate the change in uORF2 initiation (ΔuORF2 initiation), and Renilla luciferase values to estimate the expected change in global initiation rates due to eIF2α phosphorylation (Δglobal initiation). This analysis revealed significantly more bypass of uORF2 during stress for the reporter with wild-type *ATF4* 5' UTR compared to the reporter with the mutated start-stop. These findings corroborated the differences in induction observed in the CRISPR-edited cell lines. Taken together, the data supported two important roles for the start-stop in human ATF4 translational control: 1) reducing the flow of ribosomes through the 5' UTR during normal conditions thus reducing uORF2 translation and stabilizing the mRNA, which allows for complete translation suppression of the CDS at a lower cost to the cell, and 2) higher bypass probability of uORF2 during stress, leading to greater ATF4 induction (Fig. 2F).

Next, we used mathematical modeling incorporating parameters measured in our experiments to determine whether the proposed model was true. We compared the start-stop system to a system that only relied on uORFs for translational control and to a system that simply relied on doubled transcription (supplementary text). We simulated values for *ATF4* mRNA, uORF2 peptide, and ATF4 protein levels for each scenario under three conditions: 1) cells reaching steady state under control conditions, 2) cells transitioning from control conditions to stress, and 3) cells experiencing acute stress followed by recovery. Under control conditions, none of the models produced ATF4 protein, as is observed in cultured cells (Fig. 2G). The model in which transcription was doubled reached the highest mRNA levels at steady state, but also produced twice as much uORF2 peptide. The model including both the start-stop and uORFs exhibited an increase in *ATF4* mRNA levels relative to the uORF-only model, due to decreased mRNA decay. Importantly, this start-stop mediated mRNA stabilization did not increase wasteful uORF2 translation as the start-stop also reduced the flow of scanning ribosomes within the 5' UTR. However, the most striking difference between the models was the greater stress-inducibility of ATF4 protein of the model including uORFs and a start-stop compared to other models (Fig. 2G): a small decrease (~10%) in the probability of initiating at uORF2 resulted in >400% increase in induction of the CDS. This finding aligned with our observations in ribosome profiles of cells undergoing ER stress, where small decreases in ribosome occupancy at uORF2 coincided with large increases in CDS occupancy (Fig. 1D). Additionally, under acute stress, the uORF-only and the uORF + start-stop models produced ATF4 protein levels that were similar to those observed for ΔStart-stop and Start-stop$^{WT}$ cell lines, respectively, indicating that indeed, the uORFs are essential and sufficient for ATF4 translation under stress, while the start-stop increases the sensitivity of induction (Fig. 2, B and G).

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453809; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.




| Model | $k_{s\text{-}RNA}$ | $k_{d\text{-}RNA}$ | $\alpha_{u2\text{-}Basal}$ | $\alpha_{u2\text{-}Stress}$ |
|---|---|---|---|---|
| uORF-only | 0.31 | 0.25 | 1.00 | 0.97 |
| Start-stop + uORF | 0.31 | 0.16 | 0.76 | 0.67 |
| 2x transcription + uORF | 0.62 | 0.25 | 1.00 | 0.97 |

**Fig. 2. The start-stop represses basal downstream translation while enhancing inducibility of ATF4 during stress.**

(**A**) Schematic of CRISPR-Cas9 design to disrupt start-top in *ATF4* 5' UTR. Chromatograms for HAP1 clones are shown. (**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 (left $n = 3$, right $n = 2$) normalized to the loading control (% of max). (**C**) *ATF4* mRNA levels measured by qPCR ($n = 3$ for each time point). (**D**) *ATF4* mRNA levels after treatment with RNA polymerase II inhibitor (DRB). Right panel shows calculated half-lives in hours (h). (**E**) Diagrams of luciferase constructs. Bar plots on left show luciferase activity; bar plots on right show change in uORF2 bypass

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453809; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

after treatment with Tg. (**F**) Diagram for human *ATF4* system incorporating a start-stop element. Grey dashed lines indicate potential subsequent initiation following termination. (**G**) Simulated data for mathematical models of *ATF4* system: uORF-only (black lines), start-stop + uORF (orange lines), and doubled (2x) transcription + uORF (blue lines). We modeled three scenarios: 1) initialization in unstressed conditions to reach steady state, 2) transition to stress conditions over 24 hours, 3) acute stress for four hours, followed by return to unstressed conditions (recovery). Table displays variable parameters for each model. $*P < 0.05$, $**P < 0.01$; paired t-tests. Means ± SEM. Abbreviation: DRB, 5,6-dichloro-1-β-ribofuranosyl-benzimidazole; St-St, start-stop.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453809; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

***Start-stops are widespread conserved elements affecting many signaling genes***

We next demonstrated that start-stops represent a broader mechanism of translational control far beyond *ATF4*, affecting hundreds of human genes. To do so, we analyzed ribosome profiling data for start-stop sequences (AUGUAA, AUGUAG, AUGUGA; Fig. 3A) within 5' UTRs with ribosomes stalled in the initiation-termination state, as determined by the P-site/A-site positioning within the RPF. We compared these data to control sequences consisting of the same nucleotide makeup, but lacking potential for initiation or termination (AUUGAA, AUUGAG, AUUGGA). Across six replicate HeLa samples, only eight control sequences displayed an average ribosome coverage of three or more ribosome footprints, whereas 88 start-stop sequences surpassed this threshold (Fig. 3A, Table S1). We further expanded our analysis, in total identifying 402 genes containing start-stop sequences across HeLa (*25*) and HEK293 (*18*) cells, and mouse primary keratinocytes (*30*)(Fig. 3B). Twenty-three start-stops were conserved between human and mouse cells (Fig. 3C, Table S2). *ATF4* showed generally the highest occupancy in HeLa and HEK293 cells, followed in HeLa by *UCK2*, *ELP5*, *SLC39A1*, *NFIA*, and *DROSHA* (Table S2). The start-stop containing genes functioned in different signaling pathways, e.g. ERK5, TGF- β, and PFKFB4, as well as in cancer (Fig. 3C, Table S3), and were significantly enriched in transcription factors (Benjamini-Hochberg adjusted $P = 1x10^{-6}$), such as ATF4, NFIA, JARID2, and SMAD6/7.

We next showed that start-stops were more often occupied by ribosomes than other dicodon hexamers. As a measure for how likely a given hexamer was to be engaged by a ribosome, we tabulated total occurrences for hexamers within transcripts using transcriptomic data (binned into 5' UTR, CDS, and 3' UTR), then calculated the fraction occupied by ribosomes using RPFs from the same study. We performed this analysis on all hexamers ($n = 4{,}096$) from data generated in HeLa (*25*), HEK293 (*18*), and mouse primary keratinocytes (*30*). In all three cell types, start-stops were among the top hexamers occupied by ribosomes in the 5' UTR (Fig. 3B); this effect was not observed in the CDS or 3' UTR. As above, the same results were observed from samples collected without the use of cycloheximide (fig. S7).

In general, AUG codons were enriched in P-site positions in the 5' UTR, and all three stop codons were enriched in the A-site (Fig. 3D). The combination of these two (start-stop), exhibited the highest ribosome occupancy across all P-site/A-site combinations. Ribosomes were depleted from combinations with AUG in the A-site, as well as those with stop codons in the P-site. One hexamer exhibited occupancy similar to that of start-stops, AUGGCG, corresponding to the favored context for initiation in mammals (*31, 32*). Near cognate start codons (ACG, CUG, GUG, and UUG) also showed increased ribosome occupancy when positioned in the P-site of the hexamer. This effect was often stronger when a stop codon was positioned in the A-site, particularly for CUGUAG, ACGUAA, ACGUAG, and ACGUGA, suggesting these hexamers, consisting of near cognate start codons followed by a stop codon, may function similarly to canonical start-stops.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453809; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

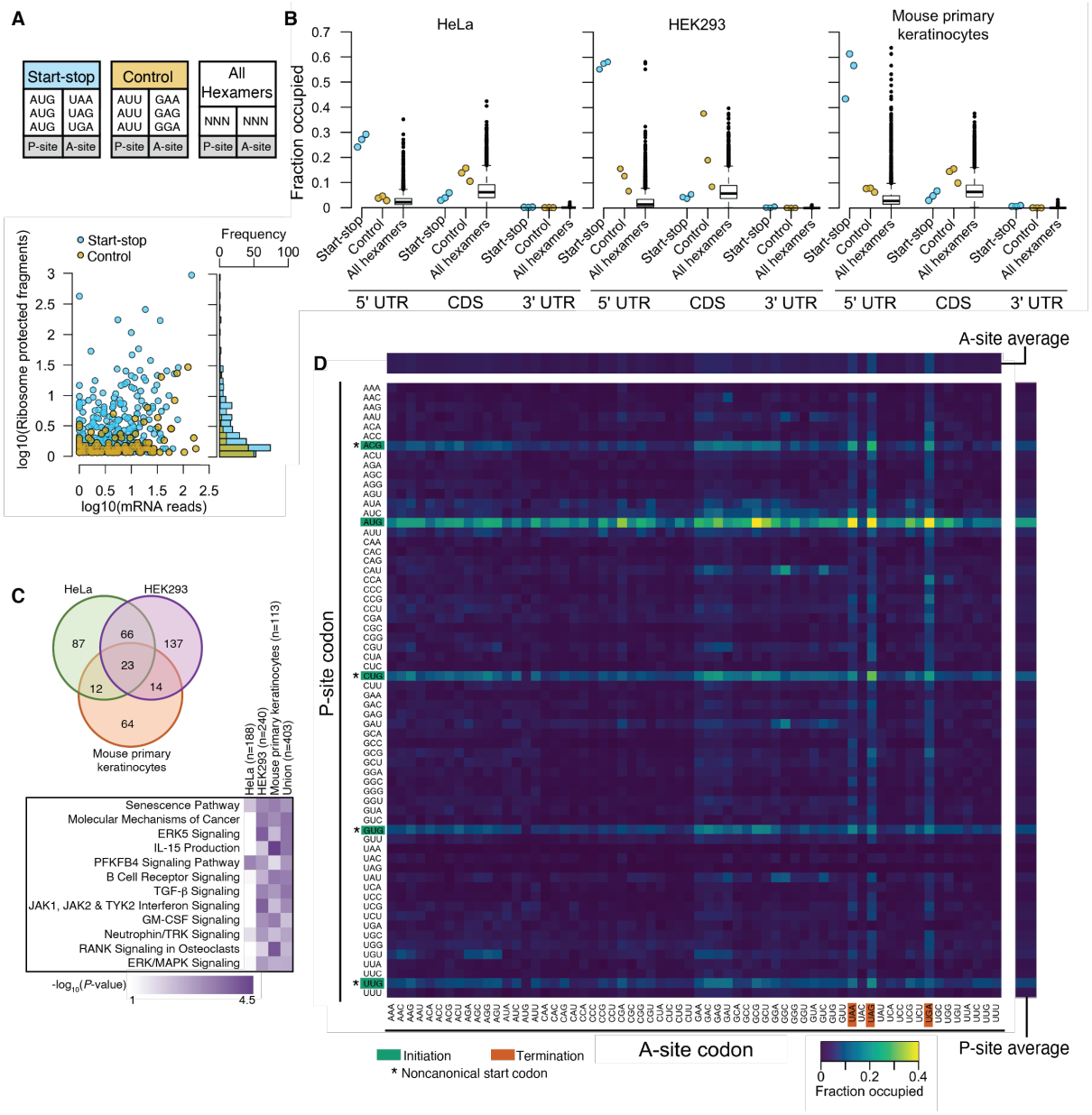



**Fig. 3. Start-stops are favored sites for ribosome occupancy in 5' UTRs.**
(**A**) Table shows sequences for start-stops, control sequences, and all hexamers that were queried for Fig. 3A and 3B. Scatter plot shows ribosome occupancy (RPF) vs. mRNA read depth for individual start-stops or control sequences in 5' UTRs (*25*); histogram summarizes frequencies across RPF coverages. (**B**) Data points in blue show fraction of start-stops occupied by ribosomes based on positional information of P- and A-sites in RPFs for HeLa (*25*), HEK293 (*18*), and mouse primary keratinocytes (*30*). 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). (**C**) The Venn diagram shows overlap of start-stop containing genes in HeLa (*25*), HEK293 (*18*), and mouse primary keratinocytes (*30*). Heatmap displays enriched pathways for each cell type and the union. (**D**) The heatmap dissects the 5' UTR further to display the fraction occupied for all P-site/A-site combinations (64 x 64 = 4,096). The average occupancy for each P-site codon is displayed on the right, and the average occupancy for each A-site codon is displayed on the top. Initiation P-sites are highlighted in green including noncanonical start codons (*). Termination A-sites are highlighted in orange.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453809; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Next, we identified typical gene architectures that illustrate the *modus operandi* of start-stop containing genes and their general functionality. While there was no architecture like that of *ATF4*, start-stops occurred in a variety of contexts, indicating transcript specific roles (Fig. 4A). Some transcripts had predicted splice isoforms including/excluding the start-stop and thereby suggesting a mode of regulation, e.g. *ELP5*, *DROSHA*. Other transcripts suggested a similar route by via alternative transcription start sites, e.g. *HMBOX1, MAP2K5*. We found several transcripts where the start-stop was the only identifiable translational control element within the 5' UTR, e.g. in *MORF4L1*, *UCK2*, *NFIA* (Fig. 4A); for these we hypothesized the start-stop would repress downstream translation at the CDS.

We demonstrated that this is true for the start-stop in *MORF4L1*, which is conserved in mouse and human with high occupancy in all three cell types (fig. S9). We constructed a reporter assay for MORF4L1 translation by fusing the 5' UTR of the *MORF4L1* gene to the luciferase coding sequence and transfected HeLa and HEK293T cells with the plasmid to measure luciferase activity. We found that mutating the start codon of the start-stop from AUG to AGG significantly increased luciferase activity in both cell lines, supporting our hypothesis that start-stops impede ribosomes as they scan through the 5' UTR, while the mutation removes this blockade (Fig. 4B). This general role in suppression of downstream translation was confirmed by analysis of ribosome occupancy data (*25*) which showed that the translation efficiency of the CDS was significantly lower on average across genes with an occupied start-stop, compared to genes without a start-stop sequence ($P < 0.001$; Fig. 4C).

Further underlining this result, we then showed that the start-stop in human *ATF4* mRNA was capable of suppressing translation of the CDS. We reasoned that this effect would normally be masked by uORF1 and uORF2, as we know these severely reduce ATF4 translation under basal conditions. Therefore, we created a luciferase reporter using the 5' UTR of human *ATF4* with uORF1 and uORF2 start codons mutated to AGG. In this way, we tested the start-stop's impact on CDS translation independently of the uORFs. For this experiment, we mutated the start codon of the start-stop (AUG to AGG) and the stop codon (UAG to CAG) and found that doing so significantly increased luciferase activity, again consistent with inhibitory effects on downstream translation (Fig. 4D).

Additionally, we found that the short, six-nucleotide start-stop was able to suppress downstream CDS translation to the same level as a long inhibitory uORF, if it were to initiate from the same position. To do so, we generated a construct where only the stop codon of the start-stop was mutated, thereby creating a long overlapping uORF that redirects any ribosomes initiating at this start codon to terminate downstream of the CDS initiation site, abolishing luciferase activity. The overlapping uORF construct displayed translation suppression comparable to the intact start-stop (Fig. 4D), indicating that the stop codon is not necessary for initiation to occur. However, the overlapping uORF relies on diverting a steady flow of ribosomes past the start codon of CDS through the synthesis of a large polypeptide, whereas the start-stop obstructs the flow of ribosomes at the 5' end of the transcript without elongation. Therefore, the start-stop uses far fewer resources to achieve the same effect. Absence of uORF1 and uORF2 in all three constructs expectedly ablated stress-inducibility of the CDS (Fig. 4D). In comparison, however, the constructs with either a start-stop or an overlapping uORF were resistant to translation inhibition of the CDS during stress, which could be due to decreased initiation rates (leaky scanning) at the upstream AUG of the start-stop or overlapping uORF, leading steady initiation at the main AUG downstream.

Next, we characterized ribosome-occupied start-stop elements in support of their functionality in gene expression regulation (Table S4). First, we showed that occupied start-stops occurred closer to the transcription start site than control sequences (Fig. 4E), supporting the role of a ribosome anchor early in the 5'UTR. We confirmed this result with 5' end distances that were experimentally measured using nanoCAGE data in MCF7 and HEK293T cells (fig. S10)(*33*). The control sequences did not have a significant difference in 5' end distances between those that were occupied or unoccupied by ribosomes, although the few control sequences that displayed ribosomal occupancy also occurred early within

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453809; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

transcripts. Proximity to the 5' cap could potentially block binding of additional preinitiation complexes through steric hinderance and might also be important for recruiting co-factors that mediate start-stop regulation, if any exist.

We found that start-stop elements whose AUG was the first in the transcript were almost always engaged with ribosomes (Fig. 4F), while unoccupied start-stops were frequently preceded by multiple AUGs (Fig. 4F). This finding is consistent with the scanning nature of ribosomes (*34*) that would place functional, ribosome-occupied start-stops in close proximity to the 5' cap. Most control sequences were not preceded by AUGs (Fig. 4F), as anticipated given these are not translational control elements.

In addition, ribosome-occupied start-stops were under greater selective pressure than unoccupied sequences and controls, indicated by a higher sequence conservation among vertebrates and supporting functionality of the sequence (Fig. 4G). Indeed, start-stop containing genes also existed in the single cell eukaryote *S. cerevisiae* and possibly represent translational regulators during meiosis (supplementary text, fig. S11, and Table S5). Furthermore, start-stop sequences had a sequence context consistent with the Kozak consensus sequence surrounding translation initiation sites (*31*), with strong preference for A or G at the -3 position when occupied (Fig. 4H). In comparison, unoccupied start-stops frequently had a T at the -3 position (Fig. 4H), likely contributing to the lack of initiation at these sites. There was also a preference for occupancy when a G or C was at the -1 position, again consistent with the Kozak sequence. We found a general GC enrichment within +/- 15 nucleotides (Fig. 4H), indicative of RNA secondary structure that could contribute to start-stop regulation (*35*), either through slowing ribosomes as they scan to reinforce initiation, or potentially mediating interactions with additional reinitiation factors.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453809; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

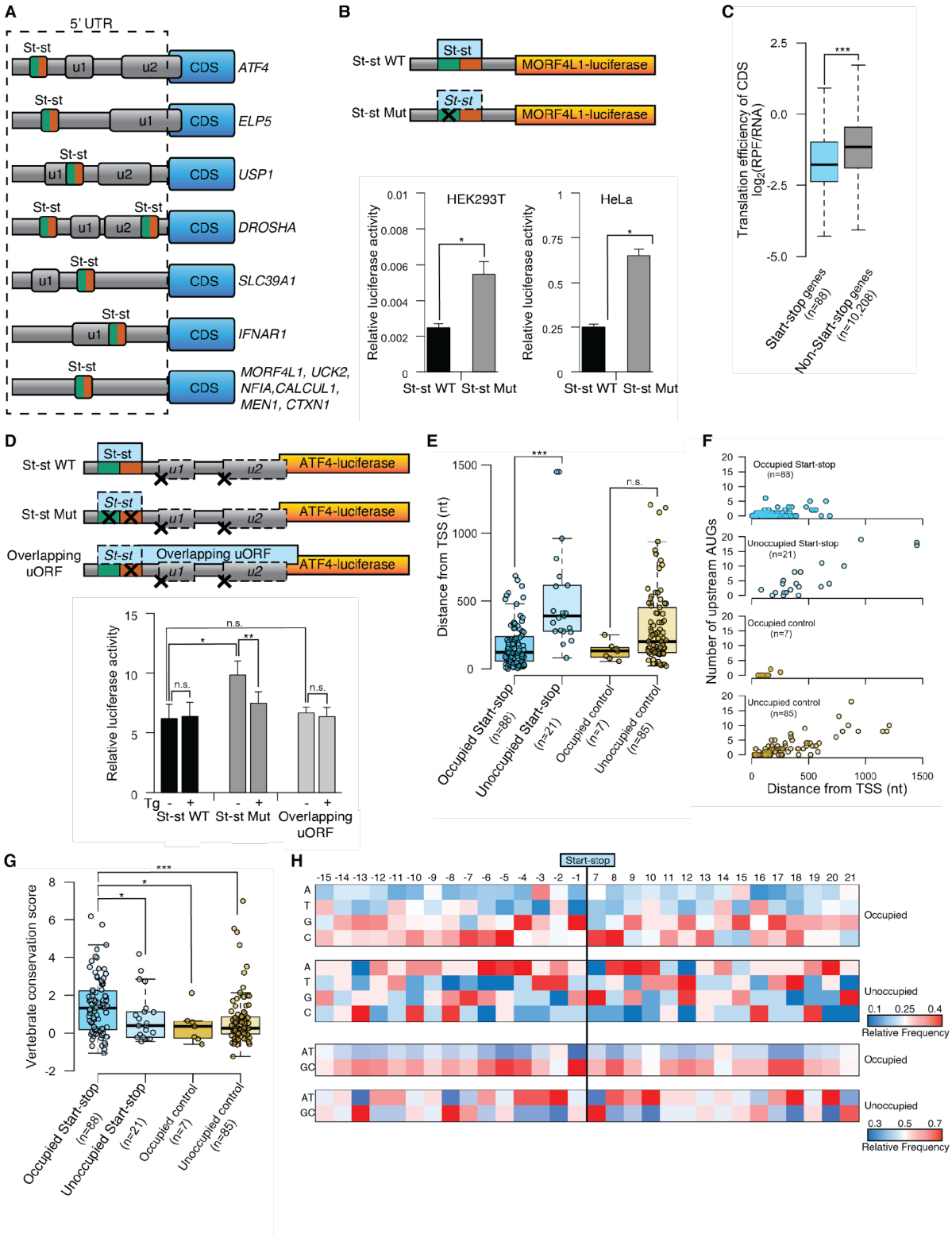



**Fig. 4. 5' UTR features contribute to start-stop function.**
(**A**) Diagrams showing the relative placement (not to scale) of start-stops within 5' UTRs relative to uORFs. (B) Schematic of *MORF4L1* 5' UTR luciferase constructs. Relative luciferase activity was measured in HEK293T and HeLa cells transfected with plasmids expressing luciferase fused to *MORF4L1* 5' UTR with wild-type start-stop sequence (St-st WT) and a mutated AUG in the start-stop (St-st Mut). (**C**) Translation efficiency of CDS was calculated from ribosome profiling and RNA-seq data in HeLa cells and compared between genes containing an occupied start-stop and those not containing a start-stop. (**D**) Schematic of *ATF4* 5' UTR luciferase constructs devoid of uORFs. Relative luciferase activity was measured in HEK293T cells transfected with plasmids expressing luciferase fused to *ATF4* 5' UTR with WT start-stop sequence (St-st WT), a mutated start-stop (St-st Mut), and a mutated stop

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453809; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

codon in the start-stop creating a long uORF overlapping with the CDS (Overlapping uORF). Luciferase activity was also measured after treatment with thapsigargin (Tg). (**E**) Distances from TSS to start-stops or control sequences were compared based on occupancy with ribosomes. (**F**) Number of upstream AUGs as a function of distance to the TSS for start-stops and controls based on occupancy with ribosomes. (**G**) Sequence conservation in vertebrates compared among start-stops and controls based on occupancy with ribosomes. (**H**) Relative frequencies of each nucleotide and dinucleotide combinations at positions ± 15 upstream and downstream of start-stops and controls. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$; paired t-tests for luciferase assays, two-sample t-tests for others. Means ± SEM. Abbreviation: St-st, start-stop; u1, uORF1; u2, uORF2; CDS, coding sequence; RPF, ribosome protected fragments; TSS, transcription start site; n.s., not significant; nt, nucleotide.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453809; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

## Discussion

The complexity of life depends on the ability to engage intricate systems that respond to many situations (*36, 37*), either intrinsic, e.g. through development, or externally inflicted, e.g. by environmental stress. Much has been learned about the orchestration of gene expression in response to a stimulus at the level of transcription (*38*), but translational control remains far less understood (*39*). Just as the upstream and downstream genomic regions of a gene inform on proper mRNA expression, UTRs flanking the CDS include the instructions for translation of proteins from each transcript. Here we characterized a hitherto unappreciated translational control element in the 5' UTR that regulates expression of many key genes in human. This element comprises adjacent start and stop codons on which ribosomes are stalled during scanning in a state of simultaneous initiation and termination that excludes elongation. Due to the lack of ribosome movement via elongation in start-stop elements, downstream reinitiation has entirely different kinetics compared to that canonical uORFs, as we demonstrate at the example of ATF4 regulation. While uORFs also impede translation of the CDS (*4*), they include elongation, ribosome movement, (perhaps wasteful) synthesis of the uORF peptide, and canonical termination which allows for return of the post-termination 40S complex to the scanning mode and downstream reinitiation.

We show that start-stop elements have several properties consistent with their role as efficient and economical suppressors of downstream translation reinitiation and that they affect hundreds of genes in the human and mouse genomes. Start-stop containing genes were enriched, for example, in the interleukin-15 (IL-15) production pathway ($P = 2.97\text{x}10^{-5}$, mouse primary keratinocytes). As IL-15 is linked to wound healing in the epidermis, it is tempting to speculate that start-stops play a role in the adaptation of keratinocytes in this process (*40*). Another 17 start-stop containing genes associated with cancer ($P = 2.43\text{x}10^{-4}$, Fig. 3D, Table S3). Notably, this included both genes that encode inhibitory Smad proteins (*SMAD6* and *SMAD7*) that play a key role in TGF-β signaling regulation (*41*). We found the start-stop sequence to be 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 (fig. S8)(*42*) supporting the likely functionality of the element. SMAD7 protein levels are frequently dysregulated in cancers (*43*), and alternative isoforms of *SMAD7* mRNAs differentially include the start-stop element, which can be altered in diseased tissues. However, the role of the SMAD7 start-stop has yet to be explored. Aberrant translation upregulation of DROSHA has been observed in both cancer and neurodegeneration, but the role of its start-stops in the 5'UTR has gone completely unnoticed. Many more examples like these exist, highlighting the importance of start-stops in affecting translation as a novel mode of regulation of key genes.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453809; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

## References and Notes


1. R. J. Jackson, C. U. Hellen, T. V. Pestova, The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat Rev Mol Cell Biol* **11**, 113-127 (2010).
2. T. Kouba, E. Rutkai, M. Karaskova, L. Valasek, The eIF3c/NIP1 PCI domain interacts with RNA and RACK1/ASC1 and promotes assembly of translation preinitiation complexes. *Nucleic Acids Res* **40**, 2683-2699 (2012).
3. A. G. Hinnebusch, I. P. Ivanov, N. Sonenberg, Translational control by 5'-untranslated regions of eukaryotic mRNAs. *Science* **352**, 1413-1416 (2016).
4. S. E. Calvo, D. J. Pagliarini, V. K. Mootha, 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 (2009).
5. A. G. Hinnebusch, Translational regulation of GCN4 and the general amino acid control of yeast. *Annu Rev Microbiol* **59**, 407-450 (2005).
6. P. D. Lu, H. P. Harding, D. Ron, Translation reinitiation at alternative open reading frames regulates gene expression in an integrated stress response. *J Cell Biol* **167**, 27-33 (2004).
7. K. M. Vattem, R. C. Wek, Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. *Proc Natl Acad Sci U S A* **101**, 11269-11274 (2004).
8. S. K. Young, R. C. Wek, Upstream Open Reading Frames Differentially Regulate Gene-specific Translation in the Integrated Stress Response. *J Biol Chem* **291**, 16927-16935 (2016).
9. S. Gunisova, V. Hronova, M. P. Mohammad, A. G. Hinnebusch, L. S. Valasek, Please do not recycle! Translation reinitiation in microbes and higher eukaryotes. *FEMS Microbiol Rev* **42**, 165-192 (2018).
10. S. Gunisova, L. S. Valasek, 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 (2014).
11. S. Wagner *et al.*, Selective Translation Complex Profiling Reveals Staged Initiation and Co-translational Assembly of Initiation Factor Complexes. *Mol Cell* **79**, 546-560 e547 (2020).
12. M. Tanaka *et al.*, The Minimum Open Reading Frame, AUG-Stop, Induces Boron-Dependent Ribosome Stalling and mRNA Degradation. *Plant Cell* **28**, 2830-2849 (2016).
13. S. Schleich, J. M. Acevedo, K. Clemm von Hohenberg, A. A. Teleman, Identification of transcripts with short stuORFs as targets for DENR*MCTS1-dependent translation in human cells. *Sci Rep* **7**, 3722 (2017).
14. J. Bohlen *et al.*, DENR promotes translation reinitiation via ribosome recycling to drive expression of oncogenes including ATF4. *Nat Commun* **11**, 4676 (2020).
15. K. Pakos-Zebrucka *et al.*, The integrated stress response. *EMBO Rep* **17**, 1374-1395 (2016).
16. Y. Park, A. Reyna-Neyra, L. Philippe, C. C. Thoreen, mTORC1 Balances Cellular Amino Acid Supply with Demand for Protein Synthesis through Post-transcriptional Control of ATF4. *Cell Rep* **19**, 1083-1090 (2017).
17. D. E. Andreev *et al.*, Translation of 5' leaders is pervasive in genes resistant to eIF2 repression. *Elife* **4**, e03971 (2015).
18. L. Calviello *et al.*, Detecting actively translated open reading frames in ribosome profiling data. *Nat Methods* **13**, 165-170 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453809; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

19. M. E. Jakobsson *et al.*, Methylation of human eukaryotic elongation factor alpha (eEF1A) by a member of a novel protein lysine methyltransferase family modulates mRNA translation. *Nucleic Acids Res* **45**, 8239-8254 (2017).
20. J. Chen *et al.*, Pervasive functional translation of noncanonical human open reading frames. *Science* **367**, 1140-1146 (2020).
21. M. V. Gerashchenko, V. N. Gladyshev, Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucleic Acids Res* **42**, e134 (2014).
22. S. K. Archer, N. E. Shirokikh, T. H. Beilharz, T. Preiss, Dynamics of ribosome scanning and recycling revealed by translation complex profiling. *Nature* **535**, 570-574 (2016).
23. D. H. Price, Poised polymerases: on your mark...get set...go! *Mol Cell* **30**, 7-10 (2008).
24. D. Vasudevan *et al.*, Translational induction of ATF4 during integrated stress response requires noncanonical initiation factors eIF2D and DENR. *Nat Commun* **11**, 4677 (2020).
25. J. Rendleman *et al.*, New insights into the cellular temporal response to proteostatic stress. *Elife* **7**, (2018).
26. N. E. Sanjana, O. Shalem, F. Zhang, Improved vectors and genome-wide libraries for CRISPR screening. *Nat Methods* **11**, 783-784 (2014).
27. M. Hirota, M. Kitagaki, H. Itagaki, S. Aiba, Quantitative measurement of spliced XBP1 mRNA as an indicator of endoplasmic reticulum stress. *J Toxicol Sci* **31**, 149-156 (2006).
28. C. Hetz, The unfolded protein response: controlling cell fate decisions under ER stress and beyond. *Nat Rev Mol Cell Biol* **13**, 89-102 (2012).
29. L. B. Gardner, Hypoxic inhibition of nonsense-mediated RNA decay regulates gene expression and the integrated stress response. *Mol Cell Biol* **28**, 3729-3741 (2008).
30. A. Sendoel *et al.*, Translation from unconventional 5' start sites drives tumour initiation. *Nature* **541**, 494-499 (2017).
31. S. Nakagawa, Y. Niimura, T. Gojobori, H. Tanaka, K. Miura, Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. *Nucleic Acids Res* **36**, 861-871 (2008).
32. Y. Niimura, M. Terabe, T. Gojobori, K. Miura, Comparative analysis of the base biases at the gene terminal portions in seven eukaryote genomes. *Nucleic Acids Res* **31**, 5195-5201 (2003).
33. V. Gandin *et al.*, nanoCAGE reveals 5' UTR features that define specific modes of translation of functionally related MTOR-sensitive mRNAs. *Genome Res* **26**, 636-648 (2016).
34. A. G. Hinnebusch, Structural Insights into the Mechanism of Scanning and Start Codon Recognition in Eukaryotic Translation Initiation. *Trends Biochem Sci* **42**, 589-611 (2017).
35. C. Y. Chan *et al.*, A structural interpretation of the effect of GC-content on efficiency of RNA interference. *BMC Bioinformatics* **10 Suppl 1**, S33 (2009).
36. G. Weng, U. S. Bhalla, R. Iyengar, Complexity in biological signaling systems. *Science* **284**, 92-96 (1999).
37. J. Rendleman, H. Choi, C. Vogel, Integration of large-scale multi-omic datasets: a protein-centric view. *Curr Opin Syst Biol* **11**, 74-81 (2018).
38. E. P. Consortium *et al.*, Perspectives on ENCODE. *Nature* **583**, 693-698 (2020).
39. B. Vitrinel *et al.*, Exploiting Interdata Relationships in Next-generation Proteomics Analysis. *Mol Cell Proteomics* **18**, S5-S14 (2019).
40. Y. Wang *et al.*, IL-15 Enhances Activation and IGF-1 Production of Dendritic Epidermal T Cells to Promote Wound Healing in Diabetic Mice. *Front Immunol* **8**, 1557 (2017).
41. M. Afrakhte *et al.*, Induction of inhibitory Smad6 and Smad7 mRNA by TGF-beta family members. *Biochem Biophys Res Commun* **249**, 505-511 (1998).
42. J. G. Dunn, C. K. Foo, N. G. Belletier, E. R. Gavis, J. S. Weissman, Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. *Elife* **2**, e01179 (2013).
43. C. Stolfi, I. Marafini, V. De Simone, F. Pallone, G. Monteleone, The dual role of Smad7 in the control of cancer growth and metastasis. *Int J Mol Sci* **14**, 23774-23790 (2013).