

Alternative ORFs and small ORFs: shedding light on the dark proteome

Mona Wu Orr¹, Yuanhui Mao², Gisela Storz^{1,*} and Shu-Bing Qian^{2,*}

¹Division of Molecular and Cellular Biology, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, MD 20892, USA and ²Division of Nutritional Sciences, Cornell University, Ithaca, NY 14853, USA

Received July 04, 2019; Revised August 03, 2019; Editorial Decision August 08, 2019; Accepted August 15, 2019

ABSTRACT

Traditional annotation of protein-encoding genes relied on assumptions, such as one open reading frame (ORF) encodes one protein and minimal lengths for translated proteins. With the serendipitous discoveries of translated ORFs encoded upstream and downstream of annotated ORFs, from alternative start sites nested within annotated ORFs and from RNAs previously considered noncoding, it is becoming clear that these initial assumptions are incorrect. The findings have led to the realization that genetic information is more densely coded and that the proteome is more complex than previously anticipated. As such, interest in the identification and characterization of the previously ignored 'dark proteome' is increasing, though we note that research in eukaryotes and bacteria has largely progressed in isolation. To bridge this gap and illustrate exciting findings emerging from studies of the dark proteome, we highlight recent advances in both eukaryotic and bacterial cells. We discuss progress in the detection of alternative ORFs as well as in the understanding of functions and the regulation of their expression and posit questions for future work.

INTRODUCTION

Most genome annotation is based on assumptions about what is being translated in the cell. For example, minimal length cutoffs commonly are utilized to prevent spurious annotation of open reading frames (ORFs). Additionally, smaller ORFs nested within larger ORFs generally are not annotated. However, it is becoming clear that these initial assumptions are incorrect. With the serendipitous discoveries of additional translated ORFs upstream and downstream of annotated ORFs, protein variants expressed from alternative start codons both in-frame and out-of-frame of annotated ORFs, and functional small proteins encoded by

RNAs previously considered noncoding, such as long non-coding RNAs (lncRNAs) and circular RNAs (circRNAs) (1–6), we now know that many previously ignored ORFs are translated as regulatory mechanisms or to give bioactive proteins (Figure 1). These discoveries are revealing that dense gene architectures previously associated with viruses and bacteriophages may be present in all organisms. The findings are also uncovering new ways proteome complexity is generated, beyond the known mechanisms of mRNA editing, alternative splicing, ribosomal frameshifting, stop codon readthrough or use of non-canonical amino acids. Aspects of the discovery and functions of these ignored ORFs have previously been reviewed (7–10). Here we seek to provide an overview of recent advances, compare and contrast what is known in eukaryotic and bacterial cells, and point out unanswered questions.

We note that nomenclature to denote these newly discovered ORFs has been inconsistent. For instance, regulatory ORFs upstream of annotated ORFs are termed upstream ORFs 'uORFs' or upstream conserved coding regions 'uCCs' in eukaryotes and 'leader peptides' in bacteria. ORFs overlapping annotated ORFs have been denoted 'alt-ORFs' in some species, and the gene products of previously unannotated small ORFs 'sORFs' have been referred to as 'small proteins', 'micropeptides' and 'miniproteins', as just a few examples. The field would benefit from adoption of a unified terminology. In this review, we denote the ORF encoding the largest, previously annotated protein in one region of a transcript the 'main-ORF' and all other ORFs derived from alternative start codons in the same region as 'alt-ORFs' (Figure 1). We refer to previously unannotated ORFs of short length (corresponding to the only coding sequence for a particular region of a transcript) as 'sORFs', and regulatory upstream ORFs as 'uORFs' for both eukaryotes and bacteria. The protein products of sORFs and uORFs will be called 'small proteins'. A few exceptions are examples where we keep the original designations for consistency.

In thinking about these previously ignored ORFs, it is worth considering some fundamental differences in the

*To whom correspondence should be addressed. Tel: +1 301 402 0968; Email: storzg@mail.nih.gov
Correspondence may also be addressed to Shu-Bing Qian. Tel: +1 607 254 3397; Email: sq38@cornell.edu

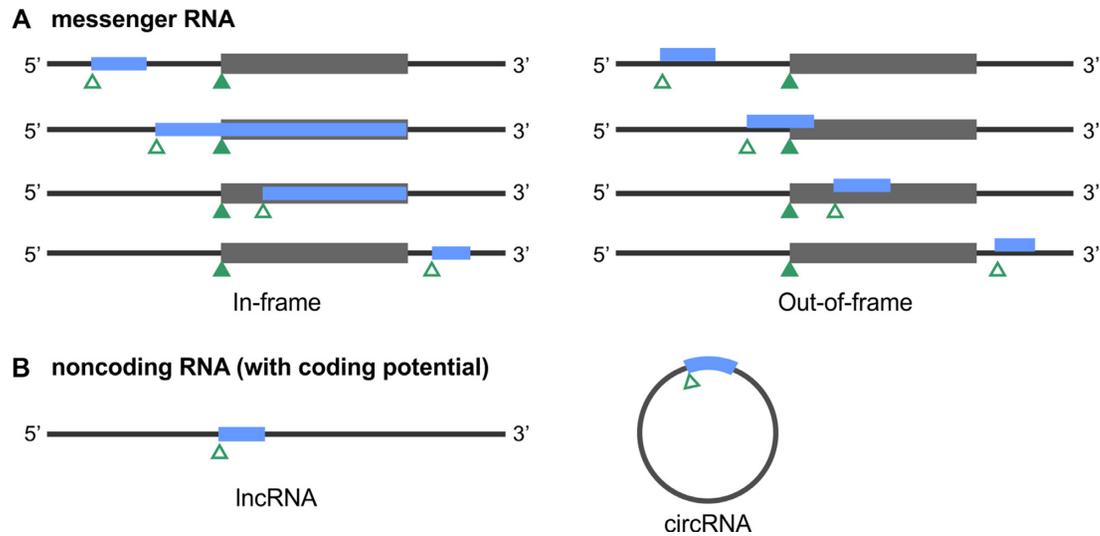


Figure 1. Architecture of alt-ORFs discovered on messenger RNAs (A) and sORFs encoded on transcripts denoted ‘noncoding’ RNAs (B). In (A), the organization of the alt-ORFs (blue) relative to the main ORFs (dark gray) is shown for mature mRNAs with alt-ORFs in the same frame (left), which can result in N-terminal extensions or deletions, and out-of-frame (right) relative to the annotated main-ORF. Annotated start codons for the main-ORF are marked with solid green triangles and alternative start codons are marked with empty green triangles.

translation of eukaryotic and bacterial mRNAs. While eukaryotic mRNAs usually are mono-cistronic, many, but not all, bacterial transcripts are poly-cistronic and encode several proteins. While transcription and translation occur in separate compartments in eukaryotic cells, they can be coupled in bacteria. Additionally, while splicing of mRNAs is common in most eukaryotic species, it is rare in bacteria. The majority of eukaryotic mRNAs are translated by a so-called ‘cap-dependent’ mechanism involving 5'-cap recognition by the eIF4F complex, which associates with eIF3 and the 40S ribosomal subunit to form, with additional factors such as tRNA-bound eIF2, the 43S preinitiation complex (11). The 43S scans through the 5' untranslated region (UTR) until it encounters an initiation codon. It is commonly assumed that the first AUG codon, which the scanning ribosome encounters, serves as the start site for translation. However, one or more potential start codons for the initiation of uORFs can exist upstream of the main start codon. Likewise, AUG codons immediately downstream of the main start codon can potentially serve as initiators. In contrast, in bacteria, a Shine-Dalgarno or ribosome binding sequence in the 5' leader usually helps direct the ribosome to the appropriate start codon. Nevertheless ribosomes can also initiate translation at noncanonical sites, including at the 5' end of ‘leaderless’ transcripts (12). The differences in eukaryotic and bacterial biology result in divergent mechanisms for alt-ORF-mediated translation regulation and for expression control.

IDENTIFICATION AND DETECTION

As more proteins encoded by previously unannotated ORFs were found by chance, it became clear that these represented overlooked translated regions. Thus, systematic approaches for the identification of these ORFs were developed. There are three main approaches: bioinformatic prediction, experimental detection by mass spectrometry, and

extrapolation that a sequence is translated based on ribosome occupancy. As identification of alt-ORFs has been reviewed comprehensively (13,14), we only provide a summary with a focus on recent advancements.

Bioinformatic prediction

Genome annotation initially relied on start codon position and ORF length to differentiate between a protein-encoding ORF from an ORF arising from chance occurrence (15–17). The inclusion of additional information such as the assessment of the conservation of codon usage and/or nucleotide frequencies within ORFs and the conservation in closely related species has improved the bioinformatic identification of alt-ORFs in eukaryotes [reviewed in (14)] and bacteria (15,18,19). For example, work with bacteria led to the development of a bioinformatic program, RanSEPs, which is a random forest-based computational approach that scores sORFs based on coding potential in a species-specific manner (5). Another recent bioinformatic study analyzed existing metagenomic sequencing datasets from the human microbiome to identify thousands of putative sORFs (20). Pioneering studies in mammals revealed more than 12 000 potential regulatory uORFs based on sequence analysis (21). This number may even be low since uORF sequences usually are less conserved than canonical protein-coding ORFs. While regulatory uORFs (leader peptides) have been observed and characterized in bacteria, to our knowledge there have been no systematic computational studies aimed at identification of bacterial uORFs. Given that little is known about nested, internal ORFs, to date there also are no algorithms aimed at detecting this type of gene in any species. Overall, despite significant efforts to improve bioinformatic prediction, the computational methods are usually subjected to a high rates of false positives. Many of these new ORF annotation programs also rely on previously identified alt-ORFs for training and val-

idation. Thus, their ability to identify alt-ORFs is limited to genes similar to the ones that have already been found. Experimental validation of many of the predictions also is missing. Nevertheless, these programs are valuable tools for improved genome annotation and to provide candidates for future validation.

Mass spectrometry

Mass spectrometry has been adapted for detecting previously unannotated proteins in organisms from all kingdoms. Rather than comparing spectrum matches to databases of previously annotated proteins, the mass spectrometric data must be compared to custom databases generated with all possible translations of a genome. Standard mass spectrometry requires proteins to be supported by multiple spectrum matches. However, small proteins often will have only one putative spectrum match, which may be difficult to detect among all other signals. Thus, biochemical enrichment, less stringent cutoffs or other specific approaches need to be taken to address these limitations. For example, peptidomics approaches inhibiting proteolysis and using electrostatic repulsion hydrophilic interaction chromatography to separate peptides prior to HPLC-MS/MS identified 90 new proteins in human cells, many matching proteins encoded by alt-ORFs (22). This method also was used in *Escherichia coli* to identify stress-induced small proteins (23,24). A significant drawback of mass spectrometry is that fragmentation of proteins prior to detection means that the true start codon of a gene must be inferred. In bacteria, translation is initiated with N-formylated methionine tRNA, which is deformylated shortly after translation begins. Inhibition of the deformylase and enrichment for formylated N-terminal peptides allows detection of translation initiation start codons by a process termed 'N-terminomics' (25). This approach was used to globally map the translation initiation sites of *Listeria monocytogenes*, revealing 6 putative sORFs and 19 putative alt-ORFs with translation initiation sites internal to an annotated ORF (26). Future advances in the isolation of the small protein fractions and mass spectrometric detection undoubtedly will have a large impact on the field.

Ribosome profiling

Thus far, the most effective identification of translation initiation sites for alt-ORFs has been the mapping of actively translated mRNAs by a technique alternately known as ribosome profiling, Ribo-seq or ribosome foot-printing. The actively translated regions of mRNA bound by ribosomes are protected from nucleolytic digestion, whereupon sequencing of the protected regions gives a snapshot of which sequences are translated. Ribosome profiling was first developed in eukaryotes (27), where strong start and stop codon peaks as well as clear 3 nt periodicity can be observed allowing unambiguous identification of the reading frame. Consequently, ribosome profiling has successfully mapped thousands of ORFs in mammals (28,29), fungi (30) and plants (2,31).

Nevertheless, standard ribosome profiling predictions have some limitations for predicting translated alt-ORFs.

The quality of ribosome profiling data and the prediction model strongly affect the number of alt-ORFs identified (2). Furthermore, given that many are short, the mapping of alt-ORFs is more strongly affected by fluctuations in periodicity. Alternative splicing in 5' UTRs can disrupt codon periodicity as can overlapping uORFs (32). To overcome these problems, several strategies have been developed to directly capture translation initiation sites by stalling initiating ribosomes (TIS-seq) (3,33,34). Translation inhibitors such as harringtonine and lactimidomycin preferentially arrest the initiating ribosomes at the start codon (3,34). The elongating ribosomes are subsequently run off (GTI-seq) (33) or depleted by another translation inhibitor such as puromycin (QTI-seq) (34), and the enriched sequences protected by initiating ribosomes are subjected to high-throughput sequencing. QTI-seq revealed that ~50% of mouse mRNAs contain at least one uORF occupied by ribosomes (34). Combined with standard ribosome profiling, these modified approaches substantially improve the sensitivity and specificity of alt-ORF mapping.

In contrast to eukaryotes, ribosome profiling in bacteria has not resulted in clear codon resolution; the signatures are less strong and more variable. Recently however, the use of the ribosome-stalling antibiotics tetracycline (35), retapamulin (1) and Onc112 (36) for ribosome profiling of *E. coli* has allowed identification of novel alt-ORFs. The treatment with tetracycline revealed 28 genes with multiple start codons and 312 sORFs <50 amino acids long (35). Treatment with retapamulin led to the identification of 124 internal translation initiation sites, 42 in-frame and 78 out-of-frame to the annotated gene (1). Analysis of datasets from both retapamulin and Onc112-treated cells led to the identification of 68 putative sORFs; of the 41 tested for expression by chromosomal tagging, 38 were shown to be translated (36). The treatment with ribosome inhibitors has been a powerful approach, but again some caution is warranted. The accumulation of ribosomes on start codons as well as the elevated levels of free ribosomes that are the result of some treatments can lead to spurious ribosome binding to nonoptimal start codons (37).

The results of these studies underscore the extent to which genome annotation has underestimated the translated proteome. However, given the caveats of each detection method, we stress the importance of additional validation of protein production. Furthermore, given the observation that alt-ORF translation can change based on media type and stress conditions [reviewed in (9,38)], proteomic- and transcriptomic-based studies are limited by the growth condition in which the cells were sampled. It is likely that as detection approaches improve and additional growth conditions are tested, even more previously unannotated proteins will be identified.

ALT-ORF TRANSLATION AS A REGULATORY FUNCTION

The translation of an ORF can itself be a regulatory event. For most studied examples, this is due to translation of an ORF upstream of or overlapping the main-ORF (uORF or leader peptide), an event that can have both positive and negative consequences (Figure 2). However, the translation

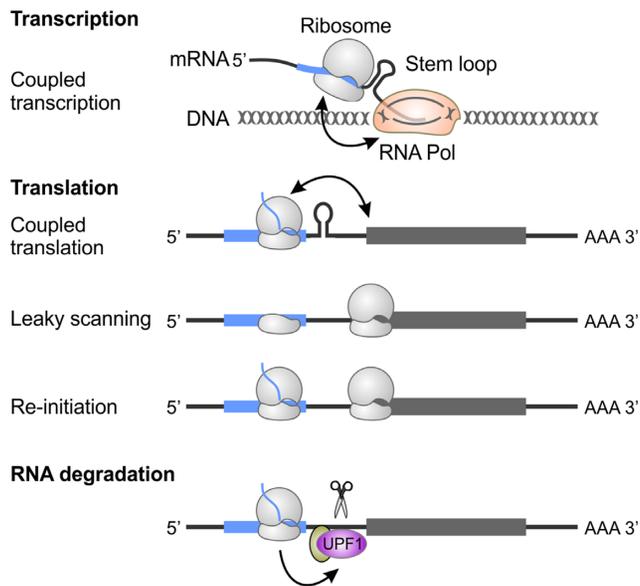


Figure 2. Regulatory roles of alt-ORFs in gene expression. The translation of alt-ORFs has wide-ranging impacts on gene expression. In bacteria, due to coupled transcription-translation, uORF translation may alter downstream RNA structures that impact transcription. A typical example is the *trp* operon. Uni-directional mRNA translation also can result in a direct effect of uORF translation on downstream main-ORF translation. Possible mechanisms include altering mRNA structures coupled to downstream translation, as well as regulation via leaky scanning and translation re-initiation. uORF translation also can affect mRNA stability throughout the recruitment of components of mRNA surveillance pathways such as UPF1.

of internal ORFs or downstream ORFs also can impact synthesis of the protein encoded by the main-ORF.

Regulation of main-ORF translation initiation in eukaryotes

About 50% of mammalian transcripts contain at least one potential uORF (33). The uORFs were hypothesized to primarily suppress the translation of main-ORFs. This notion has been supported by reporter assays and mutational studies (30,39). Translation inhibition can be achieved by uORF translation blocking scanning 43S ribosomes or sequestering 80S ribosomes. Consequently, uORFs may act as a physical barrier to prevent ribosomal access to the downstream main-ORF (40–42). Translational arrest also can be induced by interaction between the uORF-encoded polypeptides and translating ribosomes (43,44). Additionally, the presence of non-optimal codons slows down the translating ribosomes, as was found for the *Cx41* transcript in *Xenopus laevis* (45) and *ATB2/AtbZIP11* in plants (46).

A few eukaryotic uORFs stimulate the translation of mRNAs encoding stress responsive proteins. The best characterized examples are *GCN4* in yeast and *ATF4* in mammals. The 5' UTR of *ATF4* contains two uORFs: one near the 5' terminus and the other overlapping the main-ORF but in a different reading frame (47). During normal growth, the eIF2-GTP-tRNA_i^{Met} ternary complex is abundant and ribosomes translate the first uORF. The small subunit remains associated with the mRNA, resumes scanning and rapidly reacquires the readily available ternary complex,

thus allowing it to reinitiate translation at the second uORF. Termination at uORF2 prevents initiation of the main-ORF due to the overlap in sequence. Under stress conditions that trigger eIF2 α phosphorylation, the scanning ribosomes take longer to acquire a ternary complex after reinitiating downstream of the first uORF. As a result, more ribosomes bypass the second uORF and become available to translate the downstream main-ORF. It is perplexing to find that uORFs play both stimulatory and inhibitory roles in the translation of main-ORFs. Molecular mechanisms controlling leaky scanning (start codon skipping) versus reinitiation remain incompletely understood. uORF number, length, position and other features might be critical for the overall regulatory effects, and the study of uORFs is an important area for future work given how much remains to be learned.

Control of mRNA stability in eukaryotes

Nonsense-mediated decay (NMD) is an RNA quality control mechanism that eliminates mRNAs with premature translation termination codons (48,49). If the uORF stop codon is recognized as premature, it can trigger NMD, thereby decreasing transcript abundance and protein synthesis. Indeed, transcriptome-wide analysis of NMD in yeast, plant and human cells supports the idea that uORF-containing transcripts are more likely to be targeted by NMD (50–53). Intriguingly, uORFs in plants can trigger NMD in a size-dependent manner (54). The uORFs encoding proteins of >50 amino acids are sensitive to NMD, whereas short uORFs fail to activate NMD response. In yeast, the uORF in the *CPA1* mRNA encodes an arginine attenuator peptide. When arginine is abundant, translation of the arginine attenuator peptide causes ribosomal pausing at the uORF termination codon, repressing *CPA1* translation (40) and leading to NMD. Nevertheless, not all the uORF-containing mRNAs are subjected to NMD (42,55). For example, although *GCN4* contains four uORFs, it is resistant to NMD due to the presence of a stabilizer element (56,57). By contrast, *ATF4* contains two uORFs and appears to be targeted by NMD (58,59). The mechanistic correlations between NMD and uORF translation remain to be determined.

Regulation of transcription and translation initiation in bacteria

The tryptophan operon of *E. coli* is the classic bacterial example of translation of a uORF (traditionally referred to as a leader peptide in bacteria, but perhaps better designated uORF to unify nomenclature moving forward) affecting transcription of the downstream main-ORF. The 5' UTR of the *trp* operon encodes a 14 amino acid ORF, *trpL*, with two codons for tryptophan. Downstream of the tryptophan codons is an RNA sequence that can fold into either a transcription terminator or anti-terminator structure. When the activated tryptophan tRNAs are abundant, the ribosome is able to translate *trpL* and the terminator structure forms, attenuating transcription. However, when the ribosome pauses on the codons due to a lack of charged tRNA, this stalling allows formation of the anti-terminator

and permits transcription of the entire operon. Examples of similar attenuation mechanisms have been documented in other amino acid biosynthetic operons and in other bacteria [reviewed in (60,61)]. Transcription attenuation influenced by uORF translation also has been described for genes impacting other aspects of metabolism such as *pyrL* regulating pyrimidine biosynthesis (62) and *mgtL* regulating the expression of the magnesium transporter MgtA (63).

Like in eukaryotes, bacterial uORFs also regulate translation of a downstream gene. The uORF-encoded protein can stall the ribosome due to interactions between the nascent peptides and the ribosome exit channel, resulting in the formation of a secondary structure that frees the ribosome binding site of the downstream gene for translation initiation. For example, the short uORFs upstream of chloramphenicol resistance *cat* and erythromycin resistance *erm* genes encode proteins that arrest the ribosome allowing translation of the downstream antibiotic resistance gene [reviewed in (64,65)]. Translation of the uORF can affect translation of downstream genes via translational coupling. This effect is seen in operons, where the rate of translation of downstream (often overlapping) ORFs on the same polycistronic mRNA is affected by translation of the upstream ORF [reviewed in (66)]. The translation of the uORF can also affect binding of regulatory proteins, indirectly impacting the translation of the main-ORF. This mechanism is illustrated by the *iraD* gene in *E. coli* (67). The CsrA RNA-binding protein was shown to modulate translation of the *iraD* mRNA, but this regulation was found to occur via binding to the ribosome binding site of an upstream *ORF27* (also designated *idIP*), which overlaps the start codon of *iraD* (67). The recent search for sORFs in *E. coli* identified four uORFs whose translation affects expression of the downstream main-ORF: translation of *pssL* and *yoaL* enhanced levels of the main-ORFs *pssA* and *yoeE*, while translation of *baxL* and *argL* reduced levels of the main-ORFs *baxA* and *argF* (36). The regulatory mechanisms for these newly identified uORFs, and likely many others, have yet to be characterized.

Regulation of main ORF translation by internal ORF translation in bacteria

Translation from an internal alternative start codon can affect translation of the main-ORF in bacteria. For example, mutation of the out-of-frame AUG start codon of a 12 amino acid ORF internal to the *E. coli sfsA* gene resulted in loss of expression of the *sfsA* main-ORF (1). The smallest possible open reading frame is a start codon followed by a stop codon, which have been referred to as ‘minimal ORFs’ or ‘start-stops’. Ribosome profiling revealed out-of-frame internal minimal ORFs in 13 *E. coli* genes (1). Mutation of the start codon of a minimal ORF in one gene, *yecJ*, resulted in an increase in translation of the main-ORF, suggesting that these minimal ORFs also can modulate translation of the main-ORF (1). How commonly internal alt-ORFs regulate translation elongation in bacteria, whether the same regulation occurs in eukaryotic cells, and the exact mechanisms of regulation are currently unknown.

INTRINSIC FUNCTIONS AS PROTEINS

In addition to regulatory roles of uORF and alt-ORF translation mentioned above, the translational products of some these ORFs have cellular functions. For instance, small proteins encoded by uORFs in fruit flies exert critical functions in development (68), and the uORF-encoded AtCDC26 is part of the plant anaphase promoting complex/cyclosome (APC/C), regulating the accumulation of APC/C target proteins (69). Nevertheless, only a small fraction of uORF-encoding proteins are evolutionarily conserved and the codon usage in uORFs generally is similar to the random triplet frequency found in 5' UTRs (28,70). These observations suggest that most uORFs are selected not to encode bioactive proteins, but to regulate main-ORF translation. An alt-ORF in-frame with and upstream of main-ORF can encode a signal sequence that relocalizes the main-ORF encoded protein (71,72). For example, PTEN is primarily localized to the cytoplasm and the nucleus, where it acts as a tumor suppressor by negatively regulating PI3K/AKT-mediated cell survival and proliferation (72). eIF2A-mediated translation of an alt-ORF leads to an N-terminally extended form of PTEN, which is found in mitochondria and is required for mitochondrial structure and function. Other alt-ORFs, in contrast, may just be translated spuriously. Detailed mutational studies are needed to differentiate between eukaryotic uORFs and altORFs that produce functional proteins and uORFs that simply provide upstream initiation and termination sites or alt-ORFs corresponding to spurious translation.

In bacteria, independent functions have not been demonstrated for uORFs (leader peptides), but may exist in some cases. In contrast, internal alt-ORFs have been shown to have functions both related and unrelated to the main-ORF [reviewed in (13)]. As one example, ribosome profiling identified an in-frame alt-ORF in the C-terminal region of the *E. coli* ArcB sensor kinase coding region, resulting in the production of just the C-terminus of ArcB (ArcB-C) (1). ArcB is a part of the two-component regulatory system ArcAB that controls responses to changes in oxygen (73). Under microaerobic conditions, wild-type *E. coli* outcompeted the *E. coli* strain in which the *acrB-C* start codon was mutated, indicating that the AcrB-C alt-ORF is important for survival in low oxygen environments (1). Whether production of AcrB-C changes in response to aeration or other stress conditions, how translation of this alt-ORF is regulated and how AcrB-C acts remain to be identified. The frequency of alt-ORF translation indicated by ribosome profiling suggests there is much to be learned about the physiological roles of bacterial alt-ORFs.

In addition to functional proteins encoded by uORFs and alt-ORFs, another major group of previously overlooked proteins are those encoded by sORFs found on independent ‘noncoding’ transcripts such as lncRNAs and circRNAs or in operons with other main-ORFs. Only a limited number of sORF-encoded proteins have been characterized but already it is clear that they have a broad range of functions [reviewed in (8,13,38,74)]. However, a few general categories observed for small membrane proteins are coming into focus (Figure 3).

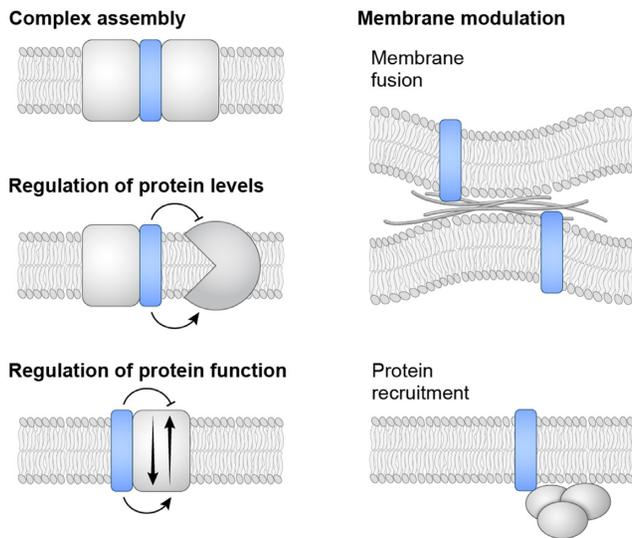


Figure 3. Biological functions of sORF products. The general functions of several small membrane proteins are illustrated. The small proteins can be subunits of a larger protein complexes (such as the CydX, CcoQ and CcoM protein components of cytochrome oxidase complexes in bacteria), affect protein stability (such as the low-magnesium induced MgtS protein that stabilizes the *E. coli* MgtA magnesium importer) or function (such as the Sarcolipin, Phospholamban, Myoregulin and DWORF modulators of the mammalian CIRCA calcium pump) that mediates membrane reorganization (such as the Minion-Myomerger-Myomixer small protein in skeletal muscle cells) and recruit complexes to cell membranes (such as the Prl42 and SpoVM proteins in *L. monocytogenes* and *B. subtilis*, respectively).

Components of larger protein complexes

A number of small proteins have been found as part of larger protein complexes, often in the membrane. Among the best characterized examples are the small proteins associated with photosystems I and II as well as cytochrome oxidases [reviewed in (75)]. Some of the associated small proteins were found as unexpected density in structural studies of these complexes (76). CydX, a small transmembrane α -helical protein associated with the *E. coli* cytochrome *bd-I* oxidase, is required for the function of this complex (77–79). However, very little is known about what CydX and other small proteins found in similar complexes are doing. One can imagine that they might be required for complex assembly as reported for the CcoQ protein (80) or for stability as suggested for the CcoM protein (81), both associated with a cytochrome *c* oxidases in *Pseudomonas stutzeri*.

Interestingly, CydX homologs have a wide diversity of sequences and a range of protein lengths. Furthermore, CydX activity withstands extensive mutagenesis and highly divergent CydX homologs can complement the *E. coli* *cydX* mutant (82,83). This heterogeneity makes predictions of small protein homologs across species difficult, particularly if genomic context is not conserved. The observations also suggest that evolution could occur rapidly in this type of small protein, as mutations in residues other than a few specific binding determinants are well-tolerated.

Modulating membrane protein activity and levels

In addition to assisting in the assembly or function of a membrane protein complex, small proteins have been found to act as regulators of membrane proteins. In eukaryotes, the small single-transmembrane α -helical proteins Sarcolipin (84), Phospholamban (85), Sarcolamban (86), Myoregulin (87) and DWORF (88,89) have been shown, in different species, to positively or negatively control the activity of the SERCA calcium pump. Sarcolipin, Phospholamban, Sarcolamban and Myoregulin share very few conserved amino acids, but structural analyses indicate that they bind SERCA in the same transmembrane groove with a similar orientation.

In *E. coli*, the small membrane protein AcrZ binds AcrB, the inner membrane resistance-nodulation-division (RND) family component of the AcrAB-TolC multidrug efflux pump (90). AcrZ modulates the pump substrate specificity, as loss of AcrZ results in increased sensitivity to a subset of the antibiotics transported by AcrB (90). AcrZ binds AcrB in a hydrophobic groove conserved in the transmembrane domain of RND pump proteins (91). While RND pumps with this conserved groove are found broadly across bacterial classes, AcrZ is only conserved in enterobacterial species, raising the question whether other small proteins of dissimilar sequence bind and regulate RND pumps in species lacking AcrZ.

Several bacterial small proteins have also been found to affect the activity or levels of membrane-embedded transporters or signaling proteins related to magnesium homeostasis. In *E. coli*, the small proteins MgtS and MgrB are induced by PhoPQ, a two-component system activated by signals such as low magnesium (92). MgtS expression increases intracellular magnesium by binding two membrane transporters, the magnesium importer MgtA (93) and the phosphate-cation symporter PitA (94). MgtS appears to prevent MgtA degradation by a membrane protease while modulating PitA activity. The determinants governing the specific MgtS interaction with two different transporters are not fully understood. Conversely, in *Salmonella enterica*, the PhoPQ-regulated MgtR small protein enhances membrane protease degradation of MgtC (95) and MgtA (96). Furthermore, the MgrB small protein interacts with the sensor kinase PhoQ itself to repress autophosphorylation, resulting in a negative feedback loop to control expression of the PhoPQ regulon (97,98). The example small proteins described here fine-tune the cellular responses to environmental changes. Since many small proteins are induced under stress conditions or are tissue or species specific, they could have arisen as a rapidly evolving means to regulate protein activity. It is likely that many more examples of these modulators will be discovered as additional small proteins are characterized.

Modulating the membrane by membrane fusion and protein recruitment

Small proteins also act at the membrane to affect fusion and protein recruitment. The small membrane protein Minion-Myomerger-Myomixer mediates membrane fusion during skeletal muscle development and regeneration (99–101). The mechanism of action is unknown, but co-

immunoprecipitation showed high enrichment of cytoskeletal proteins, suggesting that this small protein is involved in coordinating cytoskeletal reorganization events involved in fusion. In human and mouse skeletal muscle cells, the lncRNA *LINC00961* produces the SPAR protein (102). SPAR is localized to the late endosomal/lysosomal membranes where it associates with the v-ATPase complex via conserved residues in its N-terminal transmembrane domain (102). SPAR enhances v-ATPase association with the Regulator-Rag complex, which prevents Regulator recruitment of mTORC1 to the membrane, resulting in decreased mTORC1 activation (102). During muscle injury, a reduction in SPAR levels results in enhanced mTORC1 signaling and increased muscle repair (102).

There are also bacterial examples of small protein-mediated recruitment of protein complexes to the membrane. The stressosome in Gram-positive bacteria is a large multi-protein cytoplasmic complex comprised of the RsbR, RsbS and RsbT proteins that sense environmental signals to activate the general stress response sigma factor (103). The Prli42 small protein of the pathogen *L. monocytogenes* was found to co-immunoprecipitate with components of the stressosome (26). Prli42 inserts C-terminally into the cytoplasmic face of the inner membrane and binds RbsR via residues in the N-terminus to tether the stressosome to the inner membrane. Prli42 is upregulated when *L. monocytogenes* grow in blood, and cells lacking the small protein are more sensitive to oxidative stress and show decreased survival in macrophages. Thus, it appears that Prli42 is required for activation of the stressosome-response pathway. The recruitment of a protein complex is reminiscent of the function of *B. subtilis* SpoVM, which is amphipathic helix that embeds horizontally into the membrane at sites of convex curvature and recruits the spore cortex component SpoIVA to the membrane of the nascent spore (104,105). These examples illustrate how small membrane-associated proteins play regulatory roles by recruiting larger protein complexes.

REGULATION OF ALT-ORF AND SORF EXPRESSION

Expression of alt-ORFs and sORFs themselves is often extensively regulated at the transcriptional and post-transcriptional levels, though the impact of various regulatory mechanisms can differ. For instance, alt-ORFs with non-canonical start codon such as CUG or UUG may be more highly dependent on the activity of translation initiation factors or the dynamics of the scanning complex.

Modulation by transcription or alternate splicing

In eukaryotes, alternative splicing or altered transcriptional initiation is quite common at different developmental stages. By generating isoforms with different 5' UTR lengths, it is possible to create new ORFs or abrogate pre-existing ORFs from certain transcripts. During meiotic differentiation of budding yeast, translational reprogramming occurs via a switch between mRNA isoforms that differ in uORF content (106). This feature is also exploited by *Drosophila melanogaster* during development (68). In *Arabidopsis thaliana*, alternative transcription to eliminate uORFs is induced after exposure to blue light (52). Intriguingly, *Caenorhabditis elegans* uses trans-splicing to remove

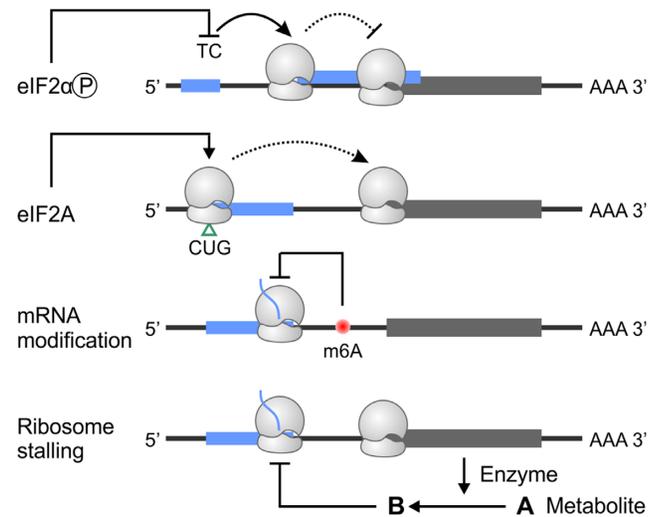


Figure 4. Regulation of alt-ORF translation in eukaryotes. Translation of alt-ORFs is modulated by both *cis*-sequence elements and *trans*-acting factors. Stress signaling pathways lead to phosphorylation of eIF2 α , which triggers translation re-initiation at a second uORF by affecting the eIF2-GTP-tRNA^{Met} ternary complex (TC) availability. This can block translation of the downstream main-ORF. In response to certain stress conditions, eIF2A promotes non-AUG translation (CUG shown here), which appears to stimulate main ORF translation. mRNA modification in the form of m⁶A may influence alternative start codon selection by slowing ribosome scanning. Additionally, translation of some uORFs is sensitive to intracellular metabolites whose levels are controlled by enzymes encoded by the downstream main-ORF.

upstream AUG codons in native 5' UTRs in order to enhance the translation efficiency of a subset of mRNAs (107). Producing transcript isoforms probably is the most straightforward way to achieve alt-ORF diversity without altering the translation machinery.

In bacteria, transcription of the mRNAs encoding the small proteins also has been shown to be strongly regulated. As already mentioned, in *E. coli*, the transcription of the *mgtS* and *mgrB* mRNAs is induced by the response regulator PhoP in low magnesium. Likewise, the transcription of the mRNA encoding AcrZ is induced by the same toxic compounds that activate transcription of the *acrB* mRNA. Global analysis of small protein levels in *E. coli* showed that many are highest under specific environmental conditions (108), most likely regulated at the level of transcription.

Modulation by translation

In eukaryotes, the translation of alt-ORFs can change in response to different developmental stages, environmental cues and stress conditions. For instance, ribosome profiling showed an increase of uORF occupancy under many stress conditions such as starvation, oxidative stress, heat shock and proteotoxic stress (47,109,110). The mechanisms of this dynamic regulation are beginning to be understood [reviewed in (111,112)] and we summarize a few well-characterized pathways below (Figure 4).

It has been well-established that the stringency of start codon recognition can be strengthened or weakened by *cis* elements and *trans*-acting factors. In higher eukaryotes, the optimal context for the initiator AUG triplet is a purine at

position -3 and a guanine at position +4 (113). The presence of mRNA secondary structures at or near the start codon influences the recognition efficiency (114,115). In addition to these *cis* sequence elements, the stringency of start codon selection is subject to regulation by nearly all initiation factors, including eIF1, eIF1A and eIF3 (116–120). Impediments to 43S scanning by secondary structures or RNA-binding proteins enhances the selection of non-canonical start codons that otherwise would be skipped. Inefficient recognition of an initiator codon results in a portion of 43S ribosomes continuing to scan and initiating at a downstream site, in a process known as leaky scanning.

In response to stress conditions, phosphorylation of eIF2 α reduces the abundance of eIF2-GTP-tRNA_i^{Met} ternary complex. This reduction leads to a global attenuation of protein synthesis. However, stress genes such as *ATF4*, *GADD34* and *CHOP* are translationally maintained or even induced by a re-initiation mechanism that relies on uORFs. Similarly, UV-induced DNA damage triggers selective translation of mRNAs containing uORFs in the 5' UTR (121), indicating that this regulatory mechanism is broadly prevalent under stress conditions. In mammalian cells, a gene encoding a mitochondrial ribosome protein MRPL18 undergoes alternative translation initiation in response to heat shock stress (122). The selection of a downstream CUG codon gives rise a truncated MRPL18 that is cytoplasmic. Intriguingly, cytoplasmic MRPL18 incorporates into the 80S ribosome in a stress-dependent manner, amplifying the impact of alternative translation.

eIF2A is an alternative tRNA carrier that appears to be responsible for initiation events from non-AUG codons such as CUG. The *BiP* transcript encodes an endoplasmic reticulum (ER) chaperone protein whose translation during ER stress relies on two uORFs with non-AUG codons (-190 UUG and -61 CUG). In response to ER stress, up-regulated eIF2A initiates translation from these non-AUG uORFs, which promotes the downstream translation of the *BiP* main-ORF (123). As another example in yeast, the levels of Ded1p, a DEAD-box RNA helicase, are reduced during meiosis (124). Ded1p catalyzes the forward scanning of ribosomes by interacting with the scanning ribosome at the mRNA entry channel and unwinding mRNA secondary structures. Reduced levels of Ded1p lead to alternative initiation from near-cognate start codons immediately upstream of RNA secondary structures (118). These results suggest that activation of alt-ORF translation is regulated by modulating the ribosome scanning process.

Polyamines such as spermidine are essential for a wide range of cellular functions, and key enzymes in polyamine biogenesis are tightly regulated at the level of translation. The mammalian *AZINI* gene encoding an antizyme inhibitor, a regulator of polyamine synthesis, contains a uORF essential for polyamine-mediated translational repression of this gene (125). A recent study reported that polyamine competitively inhibits elongation factor eIF5A, causing the pausing of elongating ribosomes on a conserved Pro-Pro-Trp (PPW) motif in the uORF (126). The queuing of scanning ribosomes behind a paused ribosome promotes initiation at upstream weak start sites, thereby reducing *AZINI* translation. Likewise, translation of *S*-adenosylmethionine decarboxylase is sensitive to intracel-

lular polyamine levels via uORF translation (127). In this case, a Met-Ala-Gly-Asp-Ile-Ser (MAGDIS) hexapeptide stalls ribosomes at the uORF stop codon and the duration of arrest depends on the concentration of polyamines. It is clear that there is extensive communication between alt-ORF and main-ORF translation to achieve polyamine homeostasis.

In bacteria, the expression of alt-ORFs can also be regulated at the post-transcriptional level. This is illustrated by genome-wide mapping of regulatory small RNA (sRNA) targets in *E. coli*, which showed that many small protein-coding mRNAs are associated with sRNAs, suggesting sRNA-mediated post-transcriptional regulation (128). In the intriguing case of the *gndA* alt-ORF encoded out of frame within the *gndA* main-ORF (6-phosphogluconate dehydrogenase) in *E. coli* (23), GndA protein levels were found to be elevated upon heat shock, while there is no evidence that *gndP* mRNA levels increase under this condition.

Modulation by RNA modification

Nucleotide modifications have been observed in many types of RNA. Some of the RNA modifications serve as sentinels for stress conditions, while others directly affect the decoding process of translation. In eukaryotic mRNAs, different types of methylation have been documented. One abundant and conserved mRNA modification is *N*⁶-methyladenine (m⁶A) (129–132). The abundance of m⁶A has been estimated to be, on average, 3–5 residues per mRNA in human cells. Importantly, the m⁶A modification is dynamic and can be reprogrammed under different conditions. For instance, yeast cells have low levels of m⁶A modification during regular mitotic growth, but 50% of all mRNAs contain m⁶A sites during meiosis (133). In addition, the m⁶A landscape changes in response to various stimuli (134,135). Intriguingly, m⁶A in 5' UTRs affects the ribosome scanning process, thereby influencing uORF translation (136). In the *ATF4* mRNA, the stress-induced removal of m⁶A from uORF2 facilitates uORF2 bypass, thereby enhancing *ATF4* translation. How exactly m⁶A affects the scanning process remains unclear.

ROLES IN DISEASE

A growing body of evidence shows that alternative translation initiation plays a critical role in the maintenance of cellular homeostasis. In humans, mutations that create or delete alt-ORFs are associated with various physiological abnormalities. Examples include craniofrontonasal syndrome (CFNS) and thrombocythaemia (137). CFNS is a rare X-linked dominant disorder caused by loss-of-function mutations in the *EFNB1* transcript. There are at least two CFNS-associated mutations found in the *EFNB1* 5' UTR. Either 95T>C or 95T>G abolishes the stop codon of a uORF, resulting in an overlapped, out-of-frame ORF relative to the main-ORF. Another mutation (411C>G) leads to creation of a new uORF. Translation of these abnormal uORFs leads to repressed translation of main-ORF. In the case of thrombocythaemia (138), a mutation in 5' UTR of thrombopoietin gene causes abnormal translation of the pre-existing uORF, resulting in systemic overexpression of thrombopoietin leading to thrombocythaemia.

A number of neurological disorders are caused by the expansion of nucleotide repeats. Typical examples include Huntington's disease with CAG repeat expansions in the coding sequence of the *Huntingtin* gene (139), and fragile X disorders with CGG repeat expansions in the 5' UTR of the *FMR1* gene (140). Importantly, repeat-associated non-AUG (RAN) translation appears to be the etiology of many neurodegenerative diseases. RAN translation occurs in multiple reading frames and appears to initiate within expanded nucleotide repeats, generating toxic polypeptides. Regulatory mechanisms underlying RAN translation remain incompletely understood.

Both genome-wide and gene-specific analyses have provided evidence for the functional and clinical relevance of 5' UTR mutations in cancer (141–143). Cancer cells exploit multiple mechanisms to modulate main-ORF translation. For example, human epidermal growth factor receptor 2 (*HER2*, also known as *ERBB2*) is overexpressed in ~15–30% of breast cancers (144). The translation of *HER2* is normally repressed by uORF translation. In tumor cells, however, the RNA-binding proteins HUR and hnRNP1 overcome the inhibitory effect of this translation by binding to the 3' UTR of *HER2* (145,146). The deregulation of uORF translation in cancer also can have the opposite effect. The *CDKN2A* gene encodes two proteins p16INK4a and p14ARF, which act as tumor suppressors by regulating the cell cycle (147,148). In addition to mutations in the main-ORF, a single point mutation in the 5' UTR of *CDKN2A* creates a new translation initiation site. Translation of this novel uORF decreases *CDKN2A* protein levels in hereditary melanoma. The repeated observation of uORF mutations associated with disease implies they play a crucial role in pathogenesis.

De-regulation of uORF translation also can occur in a global manner. A recent study reported that alternative translation initiation is globally elevated in embryonic skin cells from a mouse tumorigenesis model (149). It was hypothesized that the phosphorylation of eIF2 α in early stages of tumorigenesis leads to reprogrammed translation via eIF2A-dependent alternative initiation. How differential usage of uORFs genome-wide can drive cancer phenotypes as of yet is unknown.

PERSPECTIVES

As seen from the studies summarized here, the density of information encoded in eukaryotic and bacterial genomes is far greater than previously appreciated, with many more translation events than expected. Translation can serve a regulatory role or result in a functional protein, or possibly both. The tight regulation observed suggests that the translation of these previously ignored regions is important to the cell, a point emphasized by the association of abnormal alt-ORF expression with disease. These studies open up many exciting new avenues of research already mentioned and raise additional questions worth considering.

Standard of evidence

It is unlikely that every possible translated ORF gives rise to a functional protein, and the functional analysis of every predicted translated alt-ORF will take time. This leads

to the question of when to annotate a putative ORF. What level of evidence is necessary? Is evidence of translation by ribosome profiling or matches by mass spectrometry sufficient? The presence of ribosome density does not necessarily indicate translation of a protein product. Similarly, a single spectral match by mass spectrometry does not provide concrete evidence for synthesis. Furthermore, translation of an ORF could be translational noise that neither has a regulatory function nor produces a stable protein. We suggest that multiple lines of evidence must be present before a putative ORF is annotated.

Complication of dual-function

For nested alt-ORFs such as the *gndA-gndP* pair, there are questions about how expression of the two functions encoded by overlapping regions impact each other. Related to this, there are an increasing number of examples where transcripts shown to act as regulatory RNAs also encode small proteins, thus serving a 'dual-function' (150). The regulatory RNA and the small protein could either act in the same pathway or have very different consequences. The plant *Medicago truncatula* regulatory miRNA miR171b encodes the small protein miPEP171, which was shown to increase accumulation of miR171b to enhance lateral root formation (151). In bacteria, the *E. coli* SgrS regulatory sRNA also encodes the SgrT protein; both SgrS and SgrT regulate the same target, the glucose transporter EIICB (152). The *Staphylococcus aureus* RNAlII sRNA, which base pairs with multiple mRNAs to regulate virulence, also encodes a δ -hemolysin protein. While δ -hemolysin is important in *S. aureus* virulence, in this case the regulatory sRNA and the protein it produces do not impact the same target [reviewed in (153,154)]. Very little is known about when and how these dual-function RNAs act as mRNAs versus regulatory RNAs or whether the two activities interfere with each other.

Evolution

Consideration of how alt-ORFs in general, and overlapping functions in particular, evolve also leads to interesting questions. For example, in the case of dual-function RNAs, which activity evolved first? For small, hydrophobic membrane proteins, did these evolve from the duplication of a transmembrane domain? An attractive hypothesis is that alt-ORFs, which are less conserved than main-ORFs (consistent with their regulatory role in gene expression), have evolved more recently. Unlike main-ORFs that primarily employ canonical AUG start codons, alt-ORFs often use near cognate start codons (AUG-like triplets) for initiation. This feature further argues that translation of alt-ORFs is of much lower efficiency when compared to the main-ORF. It was recently shown that in a *D. melanogaster* population, a considerable fraction of alt-ORFs are beneficial and rapidly fixed under positive selection (68). In addition to mutations, the insertion of transposable elements can generate new alt-ORFs. Notably, an estimated ~10% human alt-ORFs are derived from transposable elements (155). The evolutionary forces governing alt-ORF generation in all organisms remain to be explored.

Potential for exploitation

Given their broad roles, it is desirable to engineer alt-ORFs for precise control of protein translation (156–158). Indeed, genome editing of endogenous uORFs in plants enabled translational modulation of main-ORFs encoding proteins involved in development or antioxidant biosynthesis (156). Additionally, inserting pathogen-responsive uORFs into transgenes encoding key immune regulators allows plants to gain broad-spectrum disease resistance without compromising plant fitness (158). More recently, efforts have been undertaken to identify small molecule compounds that target uORF translation (159). Since the majority of uORFs rely on near-cognate start codons, targeting non-AUG translation may be a therapeutic strategy.

Along these same lines, once more is known about small protein mechanisms of action, their functions could be exploited for disease treatment. Both stable synthetic analogs of natural small regulatory proteins and molecules that affect small protein binding could have therapeutic value [reviewed for eukaryotes in (13)]. For instance, the inhibition or mimicry of small proteins involved in infection, such as the stressosome-recruiting protein Prli42, or antibiotic resistance, such as the the AcrZ multidrug efflux pump regulator, could be valuable tools for combating infection or multi-drug resistance, particularly in conjunction with other therapies.

CONCLUSION

Going forward, we suggest that addressing the questions raised in this review will be of great interest and importance. This work will provide novel insights into translational regulation and uncover hidden coding potential for the genomes and additional functions for the proteomes of all organisms.

ACKNOWLEDGEMENTS

We thank Nicholas Guydosh and Allen Buskirk for comments on this review. Research in the G.S. lab is supported by the Intramural Research Program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development.

FUNDING

National Institutes of Health [R01GM1222814, R21CA227917]; HHMI Faculty Scholar Program [55108556] to S.B.-Q. Funding for open access charge: NICHD [ZIA HD008855-12].

Conflict of interest statement. None declared.

REFERENCES

- Meydan,S., Marks,J., Klepacki,D., Sharma,V., Baranov,P.V., Firth,A.E., Margus,T., Kefi,A., Vázquez-Laslop,N. and Mankin,A.S. (2019) Retapamulin-assisted ribosome profiling reveals the alternative bacterial proteome. *Mol. Cell*, **74**, 481–493.
- Hsu,P.Y., Calviello,L., Wu,H.-Y.L., Li,F.-W., Rothfels,C.J., Ohler,U. and Benfey,P.N. (2016) Super-resolution ribosome profiling reveals unannotated translation events in *Arabidopsis*. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, E7126.
- Fritsch,C., Herrmann,A., Nothnagel,M., Szafranski,K., Huse,K., Schumann,F., Schreiber,S., Platzer,M., Krawczak,M., Hampe,J. *et al.* (2012) Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. *Genome Res.*, **22**, 2208–2218.
- Shell,S.S., Wang,J., Lapiere,P., Mir,M., Chase,M.R., Pyle,M.M., Gawande,R., Ahmad,R., Sarracino,D.A. and Ioegeger,T.R. (2015) Leaderless transcripts and small proteins are common features of the mycobacterial translational landscape. *PLoS Genet.*, **11**, e1005641.
- Miravet-Verde,S., Ferrar,T., Espadas-García,G., Mazzolini,R., Gharrab,A., Sabido,E., Serrano,L. and Lluch-Senar,M. (2019) Unraveling the hidden universe of small proteins in bacterial genomes. *Mol. Syst. Biol.*, **15**, e8290.
- Ragan,C., Goodall,G.J., Shirokikh,N.E. and Preiss,T. (2019) Insights into the biogenesis and potential functions of exonic circular RNA. *Sci. Rep.*, **9**, 2048.
- Couso,J.-P. and Patraquim,P. (2017) Classification and function of small open reading frames. *Nat. Rev. Mol. Cell Biol.*, **18**, 575–589.
- Duval,M. and Cossart,P. (2017) Small bacterial and phagic proteins: an updated view on a rapidly moving field. *Curr. Opin. Microbiol.*, **39**, 81–88.
- Meydan,S., Vázquez-Laslop,N. and Mankin,A.S. (2018) Genes within genes in bacterial genomes. *Microbiol. Spectr.*, **6**, doi:10.1128/microbiolspec.RWR-0020-2018.
- Ruiz-Orera,J. and Albà,M.M. (2019) Translation of small open reading frames: roles in regulation and evolutionary innovation. *Trends Genet.*, **35**, 186–198.
- Hinnebusch,A.G. (2014) The scanning mechanism of eukaryotic translation initiation. *Annu. Rev. Biochem.*, **83**, 779–812.
- Laursen,B.S., Sørensen,H.P., Mortensen,K.K. and Sperling-Petersen,H.U. (2005) Initiation of protein synthesis in bacteria. *Microbiol. Mol. Biol. Rev.*, **69**, 101–123.
- Saghatelian,A. and Couso,J.P. (2015) Discovery and characterization of smORF-encoded bioactive polypeptides. *Nat. Chem. Biol.*, **11**, 909–916.
- Pueyo,J.I., Magny,E.G. and Couso,J.P. (2016) New peptides under the s(ORF)ace of the genome. *Trends Biochem. Sci.*, **41**, 665–678.
- Samayoa,J., Yildiz,F.H. and Karplus,K. (2011) Identification of prokaryotic small proteins using a comparative genomic approach. *Bioinformatics*, **27**, 1765–1771.
- Zhou,P., Silverstein,K.A.T., Gao,L., Walton,J.D., Nallu,S., Guhlin,J. and Young,N.D. (2013) Detecting small plant peptides using SPADA (Small Peptide Alignment Discovery Application). *BMC Bioinform.*, **14**, 335.
- Ladoukakis,E., Pereira,V., Magny,E.G., Eyre-Walker,A. and Couso,J.P. (2011) Hundreds of putatively functional small open reading frames in *Drosophila*. *Genome Biol.*, **12**, R118.
- Goli,B. and Nair,A.S. (2012) The elusive short gene – an ensemble method for recognition for prokaryotic genome. *Biochem. Biophys. Res. Commun.*, **422**, 36–41.
- Chen,S., Zhang,C.-y and Song,K. (2013) Recognizing short coding sequences of prokaryotic genome using a novel iteratively adaptive sparse partial least squares algorithm. *Biol. Direct*, **8**, 23.
- Sberro,H., Fremin,B.J., Zlitni,S., Edfors,F., Greenfield,N., Snyder,M.P., Pavlopoulos,G.A., Kyrpides,N.C. and Bhatt,A.S. (2019) Large-scale analyses of human microbiomes reveal thousands of small, novel genes. *Cell*, **30781-0**, S0092-8674.
- Matsui,M., Yachie,N., Okada,Y., Saito,R. and Tomita,M. (2007) Bioinformatic analysis of post-transcriptional regulation by uORF in human and mouse. *FEBS Lett.*, **581**, 4184–4188.
- Slavoff,S.A., Mitchell,A.J., Schwaid,A.G., Cabili,M.N., Ma,J., Levin,J.Z., Karger,A.D., Budnik,B.A., Rinn,J.L. and Saghatelian,A. (2013) Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.*, **9**, 59–64.
- Yuan,P., D’Lima,N.G. and Slavoff,S.A. (2018) Comparative membrane proteomics reveals a nonannotated *E. coli* heat shock protein. *Biochemistry*, **57**, 56–60.
- D’Lima,N.G., Khitun,A., Rosenbloom,A.D., Yuan,P., Gassaway,B.M., Barber,K.W., Rinehart,J. and Slavoff,S.A. (2017) Comparative proteomics enables identification of nonannotated cold shock proteins in *E. coli*. *J. Proteome Res.*, **16**, 3722–3731.
- Bienvenut,W.V., Giglione,C. and Meinel,T. (2015) Proteome-wide analysis of the amino terminal status of *Escherichia coli* proteins at

- the steady-state and upon deacetylation inhibition. *Proteomics*, **15**, 2503–2518.
26. Impens, F., Rolhion, N., Radoshevich, L., Bécavin, C., Duval, M., Mellin, J., García del Portillo, F., Pucciarelli, M.G., Williams, A.H. and Cossart, P. (2017) *N*-terminomics identifies Prli42 as a membrane miniprotein conserved in Firmicutes and critical for stressosome activation in *Listeria monocytogenes*. *Nat. Microbiol.*, **2**, 17005.
 27. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S. and Weissman, J.S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, **324**, 218–223.
 28. Johnstone, T.G., Bazzini, A.A. and Giraldez, A.J. (2016) Upstream ORFs are prevalent translational repressors in vertebrates. *EMBO J.*, **35**, 706–723.
 29. Chew, G.L., Pauli, A. and Schier, A.F. (2016) Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish. *Nat. Commun.*, **7**, 11663.
 30. Spealman, P., Naik, A.W., May, G.E., Kuersten, S., Freeberg, L., Murphy, R.F. and McManus, J. (2018) Conserved non-AUG uORFs revealed by a novel regression analysis of ribosome profiling data. *Genome Res.*, **28**, 214–222.
 31. Lei, L., Shi, J., Chen, J., Zhang, M., Sun, S., Xie, S., Li, X., Zeng, B., Peng, L., Hauck, A. *et al.* (2015) Ribosome profiling reveals dynamic translational landscape in maize seedlings under drought stress. *Plant J.*, **84**, 1206–1218.
 32. Lim, C.S., SJ, T.W., Kleffmann, T. and Brown, C.M. (2018) The exon-intron gene structure upstream of the initiation codon predicts translation efficiency. *Nucleic Acids Res.*, **46**, 4575–4591.
 33. Lee, S., Liu, B., Lee, S., Huang, S.X., Shen, B. and Qian, S.B. (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, E2424–E2432.
 34. Gao, X., Wan, J., Liu, B., Ma, M., Shen, B. and Qian, S.-B. (2015) Quantitative profiling of initiating ribosomes *in vivo*. *Nat. Methods*, **12**, 147–153.
 35. Nakahigashi, K., Takai, Y., Kimura, M., Abe, N., Nakayashiki, T., Shiwa, Y., Yoshikawa, H., Wanner, B.L., Ishihama, Y. and Mori, H. (2016) Comprehensive identification of translation start sites by tetracycline-inhibited ribosome profiling. *DNA Res.*, **23**, 193–201.
 36. Weaver, J., Mohammad, F., Buskirk, A.R. and Storz, G. (2019) Identifying small proteins by ribosome profiling with stalled initiation complexes. *mBio.*, **10**, e02819-18.
 37. Santos, D.A., Shi, L., Tu, B.P. and Weissman, J.S. (2019) Cycloheximide can distort measurements of mRNA levels and translation efficiency. *Nucleic Acids Res.*, **47**, 4974–4985.
 38. Khitun, A., Ness, T.J. and Slavoff, S.A. (2019) Small open reading frames and cellular stress responses. *Mol. Omics*, **15**, 108–116.
 39. Occhi, G., Regazzo, D., Trivellin, G., Boaretto, F., Ciato, D., Bobisse, S., Ferasin, S., Cetani, F., Pardi, E. and Korbonits, M. (2013) A novel mutation in the upstream open reading frame of the CDKN1B gene causes a MEN4 phenotype. *PLoS Genet.*, **9**, e1003350.
 40. Gaba, A., Jacobson, A. and Sachs, M.S. (2005) Ribosome occupancy of the yeast CPA1 upstream open reading frame termination codon modulates nonsense-mediated mRNA decay. *Mol. Cell*, **20**, 449–460.
 41. Young, S.K., Palam, L.R., Wu, C., Sachs, M.S. and Wek, R.C. (2016) Ribosome elongation stall directs gene-specific translation in the integrated stress response. *J. Biol. Chem.*, **291**, 6546–6558.
 42. Ribone, P.A., Capella, M., Arce, A.L. and Chan, R.L. (2017) A uORF represses the transcription factor AtHB1 in aerial tissues to avoid a deleterious phenotype. *Plant Physiol.*, **175**, 1238–1253.
 43. Wethmar, K. (2014) The regulatory potential of upstream open reading frames in eukaryotic gene expression. *Wiley Interdiscip. Rev. RNA*, **5**, 765–778.
 44. Wilson, D.N., Arenz, S. and Beckmann, R. (2016) Translation regulation via nascent polypeptide-mediated ribosome stalling. *Curr. Opin. Struct. Biol.*, **37**, 123–133.
 45. Meijer, H.A. (2003) Ribosomes stalling on uORF1 in the *Xenopus* Cx41 5' UTR inhibit downstream translation initiation. *Nucleic Acids Res.*, **31**, 3174–3184.
 46. Wiese, A., Elzinga, N., Wobbes, B. and Smeekens, S. (2004) A conserved upstream open reading frame mediates sucrose-induced repression of translation. *Plant Cell*, **16**, 1717–1729.
 47. Vattam, 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.
 48. Lykke-Andersen, S. and Jensen, T.H. (2015) Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nat. Rev. Mol. Cell Biol.*, **16**, 665–677.
 49. Jaffrey, S.R. and Wilkinson, M.F. (2018) Nonsense-mediated RNA decay in the brain: emerging modulator of neural development and disease. *Nat. Rev. Neurosci.*, **19**, 715–728.
 50. Colombo, M., Karousis, E.D., Bourquin, J., Bruggmann, R. and Muhlemann, O. (2017) Transcriptome-wide identification of NMD-targeted human mRNAs reveals extensive redundancy between SMG6- and SMG7-mediated degradation pathways. *RNA*, **23**, 189–201.
 51. Celik, A., Baker, R., He, F. and Jacobson, A. (2017) High-resolution profiling of NMD targets in yeast reveals translational fidelity as a basis for substrate selection. *RNA*, **23**, 735–748.
 52. Kurihara, Y., Makita, Y., Kawashima, M., Fujita, T., Iwasaki, S. and Matsui, M. (2018) Transcripts from downstream alternative transcription start sites evade uORF-mediated inhibition of gene expression in *Arabidopsis*. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, 7831–7836.
 53. Saul, H., Elharrar, E., Gaash, R., Eliaz, D., Valenci, M., Akua, T., Avramov, M., Frankel, N., Berezin, I., Gottlieb, D. *et al.* (2009) The upstream open reading frame of the *Arabidopsis* AtMHX gene has a strong impact on transcript accumulation through the nonsense-mediated mRNA decay pathway. *Plant J.*, **60**, 1031–1042.
 54. Nyiko, T., Sonkoly, B., Merai, Z., Benkovics, A.H. and Silhavy, D. (2009) Plant upstream ORFs can trigger nonsense-mediated mRNA decay in a size-dependent manner. *Plant Mol. Biol.*, **71**, 367–378.
 55. 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**, 2830–2849.
 56. Ruiz-Echevarria, M.J. and Peltz, S.W. (2000) The RNA binding protein Pub1 modulates the stability of transcripts containing upstream open reading frames. *Cell*, **101**, 741–751.
 57. Toma, K.G., Rebbapragada, I., Durand, S. and Lykke-Andersen, J. (2015) Identification of elements in human long 3' UTRs that inhibit nonsense-mediated decay. *RNA*, **21**, 887–897.
 58. Park, Y., Reyna-Neyra, A., Philippe, L. and Thoreen, C.C. (2017) mTORC1 balances cellular amino acid supply with demand for protein synthesis through post-transcriptional control of ATF4. *Cell Rep.*, **19**, 1083–1090.
 59. Keeling, K.M., Wang, D., Dai, Y., Murugesan, S., Chenna, B., Clark, J., Belakhov, V., Kandasamy, J., Velu, S.E. and Baasov, T. (2013) Attenuation of nonsense-mediated mRNA decay enhances *in vivo* nonsense suppression. *PLoS One*, **8**, e60478.
 60. Kolter, R. and Yanofsky, C. (1982) Attenuation in amino acid biosynthetic operons. *Annu. Rev. Genet.*, **16**, 113–134.
 61. Gollnick, P. and Babitzke, P. (2002) Transcription attenuation. *Biochim. Biophys. Acta Gene Struct. Expr.*, **1577**, 240–250.
 62. Levin, H.L. and Schachman, H.K. (1985) Regulation of aspartate transcarbamoylase synthesis in *Escherichia coli*: analysis of deletion mutations in the promoter region of the pyrBI operon. *Proc. Natl. Acad. Sci. U.S.A.*, **82**, 4643–4647.
 63. Park, S.-Y., Cromie, M.J., Lee, E.-J. and Groisman, E.A. (2010) A bacterial mRNA leader that employs different mechanisms to sense disparate intracellular signals. *Cell*, **142**, 737–748.
 64. Ramu, H., Mankin, A. and Vazquez-Laslop, N. (2009) Programmed drug-dependent ribosome stalling. *Mol. Microbiol.*, **71**, 811–824.
 65. Ito, K. and Chiba, S. (2013) Arrest peptides: *cis*-acting modulators of translation. *Annu. Rev. Biochem.*, **82**, 171–202.
 66. McCarthy, J.E.G. and Gualerzi, C. (1990) Translational control of prokaryotic gene expression. *Trends Genet.*, **6**, 78–85.
 67. Park, H., McGibbon, L.C., Potts, A.H., Yakhnin, H., Romeo, T. and Babitzke, P. (2017) Translational repression of the RpoS antiadapter IraD by CsrA is mediated via translational coupling to a short upstream open reading frame. *mBio*, **8**, e01355-17.
 68. Zhang, H., Dou, S., He, F., Luo, J., Wei, L. and Lu, J. (2018) Genome-wide maps of ribosomal occupancy provide insights into adaptive evolution and regulatory roles of uORFs during *Drosophila* development. *PLoS Biol.*, **16**, e2003903.

69. Lorenzo-Orts, L., Witthoef, J., Deforges, J., Martinez, J., Loubery, S., Placzek, A., Poirier, Y., Hothorn, L.A., Jaillais, Y. and Hothorn, M. (2019) Concerted expression of a cell cycle regulator and a metabolic enzyme from a bicistronic transcript in plants. *Nat. Plants*, **5**, 184–193.
70. Fields, A.P., Rodriguez, E.H., Jovanovic, M., Stern-Ginossar, N., Haas, B.J., Mertins, P., Raychowdhury, R., Hacohen, N., Carr, S.A., Ingolia, N.T. *et al.* (2015) A regression-based analysis of ribosome-profiling data reveals a conserved complexity to mammalian translation. *Mol. Cell*, **60**, 816–827.
71. Outten, C.E. and Culotta, V.C. (2004) Alternative start sites in the *Saccharomyces cerevisiae* GLR1 gene are responsible for mitochondrial and cytosolic isoforms of glutathione reductase. *J. Biol. Chem.*, **279**, 7785–7791.
72. Liang, H., He, S., Yang, J., Jia, X., Wang, P., Chen, X., Zhang, Z., Zou, X., McNutt, M.A., Shen, W.H. *et al.* (2014) PTEN α , a PTEN isoform translated through alternative initiation, regulates mitochondrial function and energy metabolism. *Cell Metab.*, **19**, 836–848.
73. Iuchi, S., Cameron, D.C. and Lin, E.C. (1989) A second global regulator gene (*arcB*) mediating repression of enzymes in aerobic pathways of *Escherichia coli*. *J. Bacteriol.*, **171**, 868–873.
74. Plaza, S., Menschaert, G. and Payre, F. (2017) In search of lost small peptides. *Annu. Rev. Cell Dev. Biol.*, **33**, 391–416.
75. Storz, G., Wolf, Y.I. and Ramamurthi, K.S. (2014) Small proteins can no longer be ignored. *Annu. Rev. Biochem.*, **83**, 753–777.
76. Safarian, S., Rajendran, C., Müller, H., Preu, J., Langer, J.D., Ovchinnikov, S., Hirose, T., Kusumoto, T., Sakamoto, J. and Michel, H. (2016) Structure of a bd oxidase indicates similar mechanisms for membrane-integrated oxygen reductases. *Science*, **352**, 583–586.
77. Sun, Y.-H., de Jong, M.F., den Hartigh, A.B., Roux, C.M., Rolán, H.G. and Tsolis, R.M. (2012) The small protein CydX is required for function of cytochrome bd oxidase in *Brucella abortus*. *Front. Cell Infect. Microbiol.*, **2**, 47–47.
78. VanOrsdel, C.E., Bhatt, S., Allen, R.J., Brenner, E.P., Hobson, J.J., Jamil, A., Haynes, B.M., Genson, A.M. and Hemm, M.R. (2013) The *Escherichia coli* CydX protein is a member of the CydAB cytochrome bd oxidase complex and is required for cytochrome bd oxidase activity. *J. Bacteriol.*, **195**, 3640–3650.
79. Chen, H., Luo, Q., Yin, J., Gao, T. and Gao, H. (2015) Evidence for the requirement of CydX in function but not assembly of the cytochrome bd oxidase in *Shewanella oneidensis*. *Biochim. Biophys. Acta Gene Subj.*, **1850**, 318–328.
80. Kohlstaedt, M., Buschmann, S., Langer, J.D., Xie, H. and Michel, H. (2017) Subunit CcoQ is involved in the assembly of the Cbb3-type cytochrome c oxidases from *Pseudomonas stutzeri* ZoBell but not required for their activity. *Biochim. Biophys. Acta Bioenerg.*, **1858**, 231–238.
81. Kohlstaedt, M., Buschmann, S., Xie, H., Resemann, A., Warkentin, E., Langer, J.D. and Michel, H. (2016) Identification and characterization of the novel subunit ccoM in the cbb3-cytochrome c oxidase from *Pseudomonas stutzeri* ZoBell. *mBio*, **7**, e01921-15.
82. Allen, R.J., Brenner, E.P., VanOrsdel, C.E., Hobson, J.J., Hearn, D.J. and Hemm, M.R. (2014) Conservation analysis of the CydX protein yields insights into small protein identification and evolution. *BMC Genom.*, **15**, 946.
83. Hobson, J.J., Gallegos, A.S., Atha, B.W., III, Kelly, J.P., Lein, C.D., VanOrsdel, C.E., Weldon, J.E. and Hemm, M.R. (2018) Investigation of amino acid specificity in the CydX small protein shows sequence plasticity at the functional level. *PLoS One*, **13**, e0198699.
84. Tada, M., Kirchberger, M.A. and Katz, A.M. (1975) Phosphorylation of a 22,000-dalton component of the cardiac sarcoplasmic reticulum by adenosine 3': 5'-monophosphate-dependent protein kinase. *J. Biol. Chem.*, **250**, 2640–2647.
85. Wawrzynow, A., Theibert, J.L., Murphy, C., Jona, I., Martonosi, A. and Collins, J.H. (1992) Sarcophilin, the “proteolipid” of skeletal muscle sarcoplasmic reticulum, is a unique, amphipathic, 31-residue peptide. *Arch. Biochem. Biophys.*, **298**, 620–623.
86. Magny, E.G., Pueyo, J.I., Pearl, F.M., Cespedes, M.A., Niven, J.E., Bishop, S.A. and Couso, J.P. (2013) Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science*, **341**, 1116–1120.
87. Anderson, D.M., Anderson, K.M., Chang, C.-L., Makarewich, C.A., Nelson, B.R., McAnally, J.R., Kasaragod, P., Shelton, J.M., Liou, J., Bassel-Duby, R. *et al.* (2015) A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell*, **160**, 595–606.
88. Nelson, B.R., Makarewich, C.A., Anderson, D.M., Winders, B.R., Troupes, C.D., Wu, F., Reese, A.L., McAnally, J.R., Chen, X., Kavalali, E.T. *et al.* (2016) A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science*, **351**, 271–275.
89. Payre, F. and Desplan, C. (2016) Small peptides control heart activity. *Science*, **351**, 226–227.
90. Hobbs, E.C., Yin, X., Paul, B.J., Astarita, J.L. and Storz, G. (2012) Conserved small protein associates with the multidrug efflux pump AcrB and differentially affects antibiotic resistance. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 16696–16701.
91. Du, D., Wang, Z., James, N.R., Voss, J.E., Klimont, E., Ohene-Agyei, T., Venter, H., Chiu, W. and Luisi, B.F. (2014) Structure of the AcrAB-TolC multidrug efflux pump. *Nature*, **509**, 512–515.
92. Moon, K. and Gottesman, S. (2009) A PhoQ/P-regulated small RNA regulates sensitivity of *Escherichia coli* to antimicrobial peptides. *Mol. Microbiol.*, **74**, 1314–1330.
93. Wang, H., Yin, X., Wu, M., Dambach, M., Curtis, R. and Storz, G. (2017) Increasing intracellular magnesium levels with the 31-amino acid MgtS protein. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, 5689–5694.
94. Yin, X., Wu, M., Wang, H., Hobbs, E.C., Shabalina, S.A. and Storz, G. (2019) The small protein MgtS and small RNA MgrR modulate the PitA phosphate symporter to boost intracellular magnesium levels. *Mol. Microbiol.*, **111**, 131–144.
95. Alix, E. and Blanc-Potard, A.-B. (2008) Peptide-assisted degradation of the Salmonella MgtC virulence factor. *EMBO J.*, **27**, 546–557.
96. Choi, E., Lee, K.-Y. and Shin, D. (2012) The MgtR regulatory peptide negatively controls expression of the MgtA Mg²⁺ transporter in *Salmonella enterica* serovar Typhimurium. *Biochem. Biophys. Res. Commun.*, **417**, 318–323.
97. Lippa, A.M. and Goulian, M. (2009) Feedback inhibition in the PhoQ/PhoP signaling system by a membrane peptide. *PLoS Genet.*, **5**, e1000788.
98. Salazar, M.E., Podgoraia, A.I. and Laub, M.T. (2016) The small membrane protein MgrB regulates PhoQ bifunctionality to control PhoP target gene expression dynamics. *Mol. Microbiol.*, **102**, 430–445.
99. Zhang, Q., Vashisht, A.A., O'Rourke, J., Corbel, S.Y., Moran, R., Romero, A., Miraglia, L., Zhang, J., Durrant, E., Schmedt, C. *et al.* (2017) The microprotein Minion controls cell fusion and muscle formation. *Nat. Commun.*, **8**, 15664–15664.
100. Quinn, M.E., Goh, Q., Kurosaka, M., Gamage, D.G., Petrany, M.J., Prasad, V. and Millay, D.P. (2017) Myomerger induces fusion of non-fusogenic cells and is required for skeletal muscle development. *Nat. Commun.*, **8**, 15665–15665.
101. Bi, P., Ramirez-Martinez, A., Li, H., Cannavino, J., McAnally, J.R., Shelton, J.M., Sánchez-Ortiz, E., Bassel-Duby, R. and Olson, E.N. (2017) Control of muscle formation by the fusogenic micropeptide myomixer. *Science*, **356**, 323–327.
102. Tajbakhsh, S. (2017) lncRNA-encoded polypeptide SPAR(s) with mTORC1 to regulate skeletal muscle regeneration. *Cell Stem Cell*, **20**, 428–430.
103. Pané-Farré, J., Q.M.B., Lewis, R.J. and Marles-Wright, J. (2017) Structure and function of the stressosome signalling hub. *Subcell. Biochem.*, **83**, 1–41.
104. Ramamurthi, K.S., Clapham, K.R. and Losick, R. (2006) Peptide anchoring spore coat assembly to the outer forespore membrane in *Bacillus subtilis*. *Mol. Microbiol.*, **62**, 1547–1557.
105. Ebmeier, S.E., Tan, I.S., Clapham, K.R. and Ramamurthi, K.S. (2012) Small proteins link coat and cortex assembly during sporulation in *Bacillus subtilis*. *Mol. Microbiol.*, **84**, 682–696.
106. Rojas-Duran, M.F. and Gilbert, W.V. (2012) Alternative transcription start site selection leads to large differences in translation activity in yeast. *RNA*, **18**, 2299–2305.
107. Yang, Y.-F., Zhang, X., Ma, X., Zhao, T., Sun, Q., Huan, Q., Wu, S., Du, Z. and Qian, W. (2017) Trans-splicing enhances translational efficiency in *C. elegans*. *Genome Res.*, **27**, 1525–1535.
108. Hemm, M.R., Paul, B.J., Miranda-Ríos, J., Zhang, A., Soltanzad, N. and Storz, G. (2010) Small stress response proteins in *Escherichia*

- coli*: proteins missed by classical proteomic studies. *J. Bacteriol.*, **192**, 46–58.
109. Gerashchenko, M.V., Lobanov, A.V. and Gladyshev, V.N. (2012) Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 17394–17399.
 110. Zhu, X., Thalor, S.K., Takahashi, Y., Berberich, T. and Kusano, T. (2012) An inhibitory effect of the sequence-conserved upstream open-reading frame on the translation of the main open-reading frame of HsfB1 transcripts in *Arabidopsis*. *Plant Cell Environ.*, **35**, 2014–2030.
 111. Wek, R.C. (2018) Role of eIF2 α kinases in translational control and adaptation to cellular stress. *Cold Spring Harb. Perspect. Biol.*, **10**, a032870.
 112. Liu, B. and Qian, S.-B. (2014) Translational reprogramming in cellular stress response. *Wiley Interdiscip. Rev. RNA*, **5**, 301–315.
 113. Kozak, M. (1986) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell*, **44**, 283–292.
 114. Gu, W., Zhou, T. and Wilke, C.O. (2010) A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Comput. Biol.*, **6**, e1000664.
 115. Mortimer, S.A., Kidwell, M.A. and Doudna, J.A. (2014) Insights into RNA structure and function from genome-wide studies. *Nat. Rev. Genet.*, **15**, 469–479.
 116. Fijalkowska, D., Verbruggen, S., Ndah, E., Jonckheere, V., Menschaert, G. and Van Damme, P. (2017) eIF1 modulates the recognition of suboptimal translation initiation sites and steers gene expression via uORFs. *Nucleic Acids Res.*, **45**, 7997–8013.
 117. Kim, T.H., Kim, B.H., Yahalom, A., Chamovitz, D.A. and von Arnim, A.G. (2004) Translational regulation via 5' mRNA leader sequences revealed by mutational analysis of the *Arabidopsis* translation initiation factor subunit eIF3h. *Plant Cell*, **16**, 3341–3356.
 118. Roy, B., Vaughn, J.N., Kim, B.H., Zhou, F., Gilchrist, M.A. and Von Arnim, A.G. (2010) The h subunit of eIF3 promotes reinitiation competence during translation of mRNAs harboring upstream open reading frames. *RNA*, **16**, 748–761.
 119. Schepetilnikov, M., Dimitrova, M., Mancera-Martinez, E., Geldreich, A., Keller, M. and Ryabova, L.A. (2013) TOR and S6K1 promote translation reinitiation of uORF-containing mRNAs via phosphorylation of eIF3h. *EMBO J.*, **32**, 1087–1102.
 120. 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.
 121. Powley, I.R., Kondrashov, A., Young, L.A., Dobbyn, H.C., Hill, K., Cannell, I.G., Stoneley, M., Kong, Y.-W., Cotes, J.A., Smith, G.C.M. et al. (2009) Translational reprogramming following UVB irradiation is mediated by DNA-PKcs and allows selective recruitment to the polysomes of mRNAs encoding DNA repair enzymes. *Genes Dev.*, **23**, 1207–1220.
 122. Zhang, X., Gao, X., Coots, R.A., Conn, C.S., Liu, B. and Qian, S.-B. (2015) Translational control of the cytosolic stress response by mitochondrial ribosomal protein L18. *Nat. Struct. Mol. Biol.*, **22**, 404–410.
 123. Starck, S.R., Tsai, J.C., Chen, K., Shodiya, M., Wang, L., Yahiro, K., Martins-Green, M., Shastri, N. and Walter, P. (2016) Translation from the 5' untranslated region shapes the integrated stress response. *Science*, **351**, aad3867.
 124. Guenther, U.-P., Weinberg, D.E., Zubradt, M.M., Tedeschi, F.A., Stawicki, B.N., Zagore, L.L., Brar, G.A., Licatalosi, D.D., Bartel, D.P. and Weissman, J.S. (2018) The helicase Ded1p controls use of near-cognate translation initiation codons in 5' UTRs. *Nature*, **559**, 130–134.
 125. Dever, T.E. and Ivanov, I.P. (2018) Roles of polyamines in translation. *J. Biol. Chem.*, **293**, 18719–18729.
 126. Ivanov, I.P., Shin, B.-S., Loughran, G., Tzani, I., Young-Baird, S.K., Cao, C., Atkins, J.F. and Dever, T.E. (2018) Polyamine control of translation elongation regulates start site selection on antizyme inhibitor mRNA via ribosome queuing. *Mol. Cell*, **70**, 254–264.
 127. Hanfrey, C., Franceschetti, M., Mayer, M.J., Illingworth, C. and Michael, A.J. (2002) Abrogation of upstream open reading frame-mediated translational control of a plant S-adenosylmethionine decarboxylase results in polyamine disruption and growth perturbations. *J. Biol. Chem.*, **277**, 44131–44139.
 128. Melamed, S., Peer, A., Faigenbaum-Romm, R., Gatt, Y.E., Reiss, N., Bar, A., Altuvia, Y., Argaman, L. and Margalit, H. (2016) Global mapping of small RNA-target interactions in bacteria. *Mol. Cell*, **63**, 884–897.
 129. Frye, M., Harada, B.T., Behm, M. and He, C. (2018) RNA modifications modulate gene expression during development. *Science*, **361**, 1346–1349.
 130. Roundtree, I.A., Evans, M.E., Pan, T. and He, C. (2017) Dynamic RNA modifications in gene expression regulation. *Cell*, **169**, 1187–1200.
 131. Fu, Y., Dominissini, D., Rechavi, G. and He, C. (2014) Gene expression regulation mediated through reversible m6A RNA methylation. *Nat. Rev. Genet.*, **15**, 293–306.
 132. Helm, M. and Motorin, Y. (2017) Detecting RNA modifications in the epitranscriptome: predict and validate. *Nat. Rev. Genet.*, **18**, 275–291.
 133. Schwartz, S., Agarwala, S.D., Mumbach, M.R., Jovanovic, M., Mertins, P., Shishkin, A., Tabach, Y., Mikkelsen, T.S., Satija, R. and Ruvkun, G. (2013) High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. *Cell*, **155**, 1409–1421.
 134. Anders, M., Chelysheva, I., Goebel, I., Trenkner, T., Zhou, J., Mao, Y., Verzini, S., Qian, S.-B. and Ignatova, Z. (2018) Dynamic m6A methylation facilitates mRNA triaging to stress granules. *Life Sci. Alliance*, **1**, e201800113.
 135. Zhou, J., Wan, J., Gao, X., Zhang, X., Jaffrey, S.R. and Qian, S.-B. (2015) Dynamic m6A mRNA methylation directs translational control of heat shock response. *Nature*, **526**, 591–594.
 136. Zhou, J., Wan, J., Shu, X.E., Mao, Y., Liu, X.-M., Yuan, X., Zhang, X., Hess, M.E., Brüning, J.C. and Qian, S.-B. (2018) N6-methyladenosine guides mRNA alternative translation during integrated stress response. *Mol. Cell*, **69**, 636–647.
 137. Romanelli Tavares, V.L., Kague, E., Musso, C.M., Alegria, T.G.P., Freitas, R.S., Bertola, D.R., Twigg, S.R.F. and Passos-Bueno, M.R. (2019) Craniofrontonasal syndrome caused by introduction of a novel uATG in the 5' UTR of EFN1. *Mol. Syndromol.*, **10**, 40–47.
 138. Wiestner, A., Schlemper, R.J., van der Maas, A.P. and Skoda, R.C. (1998) An activating splice donor mutation in the thrombopoietin gene causes hereditary thrombocythaemia. *Nat. Genet.*, **18**, 49–52.
 139. Zühlke, C., Rless, O., Bockel, B., Lange, H. and Thies, U. (1993) Mitotic stability and meiotic variability of the (CAG) $_n$ repeat in the Huntington disease gene. *Hum. Mol. Genet.*, **2**, 2063–2067.
 140. Nolin, S.L., Brown, W.T., Glicksman, A., Houck, J.G.E., Gargano, A.D., Sullivan, A., Biancalana, V., Bröndum-Nielsen, K., Hjalgrim, H., Holinski-Feder, E. et al. (2003) Expansion of the fragile X CGG repeat in females with premutation or intermediate alleles. *Am. J. Hum. Genet.*, **72**, 454–464.
 141. Schulz, J., Mah, N., Neuenschwander, M., Kischka, T., Ratei, R., Schlag, P.M., Castanos-Velez, E., Fichtner, I., Tunn, P.U., Denkert, C. et al. (2018) Loss-of-function uORF mutations in human malignancies. *Sci. Rep.*, **8**, 2395.
 142. 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.
 143. Barbosa, C., Peixeiro, I. and Romão, L. (2013) Gene expression regulation by upstream open reading frames and human disease. *PLoS Genet.*, **9**, e1003529.
 144. Iqbal, N. and Iqbal, N. (2014) Human epidermal growth factor receptor 2 (HER2) in cancers: overexpression and therapeutic implications. *Biochem. Mol. Biol. Int.*, **2014**, 852748.
 145. Mehta, A., Trotta, C.R. and Peltz, S.W. (2006) Derepression of the Her-2 uORF is mediated by a novel post-transcriptional control mechanism in cancer cells. *Genes Dev.*, **20**, 939–953.
 146. Child, S.J., Miller, M.K. and Geballe, A.P. (1999) Translational control by an upstream open reading frame in the HER-2/neu transcript. *J. Biol. Chem.*, **274**, 24335–24341.
 147. Zhao, R., Choi, B.Y., Lee, M.-H., Bode, A.M. and Dong, Z. (2016) Implications of genetic and epigenetic alterations of CDKN2A (p16^{INK4a}) in cancer. *EBioMedicine*, **8**, 30–39.
 148. Liu, L., Dilworth, D., Gao, L., Monzon, J., Summers, A., Lassam, N. and Hogg, D. (1999) Mutation of the CDKN2A 5' UTR creates an

- aberrant initiation codon and predisposes to melanoma. *Nat. Genet.*, **21**, 128–132.
149. Sandoel,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.
150. Vanderpool,C.K., Balasubramanian,D. and Lloyd,C.R. (2011) Dual-function RNA regulators in bacteria. *Biochimie.*, **93**, 1943–1949.
151. Laouressergues,D., Couzigou,J.-M., Clemente,H.S., Martinez,Y., Dunand,C., Bécard,G. and Combier,J.-P. (2015) Primary transcripts of microRNAs encode regulatory peptides. *Nature*, **520**, 90–93.
152. Lloyd,C.R., Park,S., Fei,J. and Vanderpool,C.K. (2017) The small protein SgrT controls transport activity of the glucose-specific phosphotransferase system. *J. Bacteriol.*, **199**, e00869-16.
153. Gimpel,M. and Brantl,S. (2017) Dual-function small regulatory RNAs in bacteria. *Mol. Microbiol.*, **103**, 387–397.
154. Raina,M., King,A., Bianco,C. and Vanderpool,C.K. (2018) Dual-function RNAs. *Microbiol. Spectr.*, **6**, doi:10.1128/microbiolspec.RWR-0032-2018.
155. Kitano,S., Kurasawa,H. and Aizawa,Y. (2018) Transposable elements shape the human proteome landscape via formation of *cis*-acting upstream open reading frames. *Genes Cells*, **23**, 274–284.
156. Zhang,H., Si,X., Ji,X., Fan,R., Liu,J., Chen,K., Wang,D. and Gao,C. (2018) Genome editing of upstream open reading frames enables translational control in plants. *Nat. Biotechnol.*, **36**, 894–898.
157. Liang,X.H., Shen,W., Sun,H., Migawa,M.T., Vickers,T.A. and Crooke,S.T. (2016) Translation efficiency of mRNAs is increased by antisense oligonucleotides targeting upstream open reading frames. *Nat. Biotechnol.*, **34**, 875–880.
158. Xu,G., Yuan,M., Ai,C., Liu,L., Zhuang,E., Karapetyan,S., Wang,S. and Dong,X. (2017) uORF-mediated translation allows engineered plant disease resistance without fitness costs. *Nature*, **545**, 491–494.
159. Rabouw,H.H., Langereis,M.A., Anand,A.A., Visser,L.J., de Groot,R.J., Walter,P. and van Kuppeveld,F.J.M. (2019) Small molecule ISRIB suppresses the integrated stress response within a defined window of activation. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 2097–2102.