HHS Public Access

Author manuscript

Exp Cell Res. Author manuscript; available in PMC 2021 June 01.

Published in final edited form as:

Exp Cell Res. 2020 June 01; 391(1): 111973. doi:10.1016/j.yexcr.2020.111973.

Non-AUG start codons: expanding and regulating the small and alternative ORFeome

Xiongwen Cao[†,‡], **Sarah A. Slavoff**[†,‡,⊥,*]

[†]Department of Chemistry, Yale University, New Haven, Connecticut 06520, United States

[‡]Chemical Biology Institute, Yale University, West Haven, Connecticut 06516, United States

[⊥]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06529, United States

Abstract

Recent ribosome profiling and proteomic studies have revealed the presence of thousands of novel coding sequences, referred to as small open reading frames (sORFs), in prokaryotic and eukaryotic genomes. These genes have defied discovery via traditional genomic tools not only because they tend to be shorter than standard gene annotation length cutoffs, but also because they are, as a class, enriched in sequence properties previously assumed to be unusual, including non-AUG start codons. In this review, we summarize what is currently known about the incidence, efficiency, and mechanism of non-AUG start codon usage in prokaryotes and eukaryotes, and provide examples of regulatory and functional sORFs that initiate at non-AUG codons. While only a handful of non-AUG-initiated novel genes have been characterized in detail to date, their participation in important biological processes suggests that an improved understanding of this class of genes is needed.

Keywords

Genomics; proteomics; Non-AUG start codon; small ORF; sORF

Introduction

Over the past decade, advances in genomic and proteomic technologies[1] have accelerated the discovery of thousands of small open reading frames of fewer than 100 codons (sORFs)[2–7] in genomes spanning evolutionary space. These ORFs were previously unannotated due to size cutoffs and other assumptions applied by genome annotation consortia[8,9], and are encoded in RNA regions previously assumed to be non-coding (e.g., long non-coding RNAs, 5′ and 3′ untranslated regions (UTRs)) as well as “alt-ORFs”

* sarah.slavoff@yale.edu.

Xiongwen Cao: Conceptualization; writing – original draft; writing – review and editing. **Sarah Slavoff**: Conceptualization; funding acquisition; writing – original draft; writing – review and editing.

overlapping annotated protein coding sequences or main ORFs[3,6]. We will collectively refer to recently discovered, short genes as sORFs. These newly discovered ORFs can be translated into polypeptide products that we will refer to as microproteins, though nomenclature has varied in prior reports[10–12]. The vast majority of recently reported sORFs remain uncharacterized, and some may represent protogenes[13] or *cis*-translational regulators rather than encoding functional microproteins. However, an increasing number of sORFs have been shown to play important roles in biological processes, including regulation of protein complexes, membrane channels, and transcription factors, among others[14].

Surprisingly, analyses of sORF coding sequences have revealed that these genes escaped annotation not only because they are short, but because, in many cases, they circumvent additional assumptions about the structure of genes. For example, sORFs overlapping main ORFs in alternative reading frames have been reported in both prokaryotic and eukaryotic genomes; examples include human DEDD2[3] and *E. coli gndP*[19]. Even more strikingly, several reports estimate that up to approximately 50% of sORFs detected to date do not initiate with canonical AUG start codons[2,3,7,16,20]. These findings suggest that information may be more densely encoded in non-viral genomes than previously assumed, and that assumptions about the size, sequence and monocistronic nature of eukaryotic genes provide an incomplete picture. In this minireview, we discuss the ubiquity and functions of sORFs that initiate with non-AUG start codons.

Genome-wide frequency of initiation at non-AUG start codons

Translation initiation of main ORFs at non-AUG start codons has recently been extensively reviewed[21,22], so we provide only a brief overview here. Eukaryotic genes were previously assumed to initiate specifically at the first AUG start codon of a transcript, in accordance with the scanning model of translation initiation[22]. Occasional reports of initiation of previously known eukaryotic proteins at upstream or downstream, in-frame non-AUG start codons, such as FGF-2, which initiates at a CUG codon[23,24], and c-myc, which can exist as an N-terminally extended isoform via initiation at an upstream in-frame CUG[25,26], did little to change the perception that these events were rare. In contrast, prokaryotic genes have long been known to initiate at non-AUG start codons; for example, an analysis of 620 bacterial genomes revealed that ~80% of annotated genes initiate at AUG codons, ~12% at GUG and ~8% at UUG, with variable incidences of AUU and AUC across species[27]. The advent of ribosome profiling, or deep sequencing of ribosome-protected RNA footprints as a readout of translation, revealed a much more complex picture of eukaryotic translation initiation: hundreds to thousands of short, unannotated upstream open reading frames in the 5′ leaders of yeast transcripts were found to initiate with near-cognate start codons[2], particularly during meiosis[28]. Application of inhibitors of translation initiation in bacteria, retapamulin and Onc112, have recently been similarly applied to identification of translation initiation sites and dozens of sORFs in bacteria[29], revealing abundant initiation at GUG, UUG, CUG and AUU codons, as well as occasional utilization of AUC[15] (Table 1A). Analogous inhibitors of translation initiation in mammalian cells, such as harringtonine[20] and lactidomycin[16,30], have similarly revealed that approximately 50% of all translation initiation events (including both canonical ORFs and sORFs) occur at non-AUG start codons

(Table 1B), with CUG appearing as the most common near-cognate initiation codon (15–16% of initiation sites[16]).

Mass spectrometry studies have hinted that sORFs initiate with non-AUG start codons more often than main ORFs in eukaryotic cells. Multiple studies, to date, have aimed to revise the initiation sites of main ORFs, revealing N-terminal extensions initiating from upstream, in-frame non-AUG start codons; however, the numbers of these N-terminal extensions identified in a given study is typically <100 (e.g., 42 by Baranov et al. using a combined computational and experimental approach; 50 by Zhu et al. using a mass spectrometry-based proteogenomics approach; 17 by Van Danne et al. using an N-terminalomics approach combined with ribosome profiling), a small fraction of the ~20,000 annotated human proteins. In contrast, mass spectrometry studies specifically designed for sORF enrichment and detection have revealed that up to ~50% of sORFs initiate at non-AUG start codons[3,6,31,32], though it is important to note that start codons must be indirectly inferred from bottom-up proteomics data. More quantitatively, ribosome profiling has provided direct comparisons of the non-AUG start codon usage between annotated proteins and sORFs in mammalian cells. For example, Ingolia and colleagues[20] identified 5647 start codons for canonical mouse proteins, including N-terminal truncations and extensions; as a class, these canonical proteins were relatively large, exceeding 100 amino acids 93% of the time. Of these, 79% utilized AUG and 21% utilized near-cognate start codons. In the same study, sORFs (which include ORFs mapping to “non-coding” RNAs; ORFs encoded in 5′ UTRs; ORFs internal and out-of-frame with annotated protein coding sequences; and polycistronic/downstream ORFs) were generally (96%) shorter than 100 amino acids, and utilized non-AUG start codons 70% of the time. Similarly, Qian and colleagues reported that global translation initiation sites, including both sORFs and canonical ORFs, occur at non-AUG triplets 50% of the time, while uORFs (upstream ORFs, or sORFs that occur in 5′ UTRs) initiate at near-cognate start codons with a remarkable 74% frequency[16]. While the absolute frequencies of AUG vs. non-AUG initiation vary in each study, the trend toward increased incidence of near-cognate start codons in the sORFeome relative to main ORFs is consistently observed.

Efficiency and mechanism of initiation at non-AUG start codons

Long before the genome-wide ubiquity of non-AUG initiation sites was recognized, the ability of both prokaryotic and eukaryotic ribosomes to initiate at near-cognate start codons was demonstrated, though at fractional efficiencies of initiation at AUG (Table 1C–D). In prokaryotes, a classic study demonstrated that “class I” UUG and GUG codons can initiate translation in *E. coli* with 12–15% the efficiency of an AUG within the same reporter construct; “class IIA” start codons CUG, AUU, AUC, AUA and ACG produced 1–3% protein yields relative to AUG, and the remaining codons tested (AGG, AAG) produced no detectable translation[33]. A more recent study using high-sensitivity reporters remarkably detected some level of translation initiation from almost all 64 codons (Table 1C), though sequences supporting the highest initiation efficiencies followed similar trends to previous reports, and a strong preference for U at the second position was observed[17]. An important early study in eukaryotic cell-free translation systems[34] revealed translation initiation from CUG, GUG, ACG, AUA, UUG, AUU and AUC codons with 36–82% the efficiency of AUG

in rabbit reticulocyte lysates and 8–45% in wheat germ lysate, though the efficiency of initiation at these triplets is likely to be different in living cells. More recent studies using a high-sensitivity reporter revealed translation from these codons with 2.9–18.2% the efficiency of AUG in rabbit reticulocyte lysates[18], and 1.7–19.5% in 293T cells[35]. These reporter assay-based results are consistent with the genome-wide studies, suggesting that non-AUG initiated translation is frequently and efficiently utilized from bacteria to human.

The mechanism of initiation at non-AUG start codons differs in prokaryotes and eukaryotes, as does translation initiation at standard AUG start codons. In bacteria, assembly of the 30S preinitiation complex requires a Shine-Dalgarno sequence upstream of an AUG or near-cognate start codon. This preinitiation complex positions the start codon in the P-site, base-paired with the anticodon of the initiator fMet-tRNA (N-formylmethionine-tRNA); subsequent steps permit formation of the 70S initiation complex and protein synthesis and have been reviewed elsewhere[36]. A crystal structure of a *T. thermophilus* 70S initiation complex bound to fMet-tRNA base-paired to a near-cognate AUC codon in the P-site reveals that accommodation of the near-cognate-initiator tRNA codon-anticodon mismatch occurs through a wobble mechanism[37]. The relative efficiencies of AUG vs. near-cognate start codon utilization are primarily controlled by bacterial initiation factor 3 (IF3)[33].

In eukaryotes, most translation initiation occurs via a scanning mechanism that begins when the 43S preinitiation complex consisting of the small ribosomal subunit, multiple initiation factors, PABP, and the ternary complex of eIF2 bound to GTP and the initiator tRNA recognize the mRNA 7-methylguanosine cap, then scan 5′-to-3′ until the first start codon is recognized[22]. eIF2 then hydrolyzes GTP and translation initiation occurs. Classic radiolabeling studies demonstrated that methionine is incorporated via scanning initiation at non-AUG start codons[34], suggesting that wobble base pairing between the initiator tRNA and near-cognate codons occurs. The presence of a Kozak consensus sequence (especially an A or G at the −3 position) strongly affects the efficiency of initiation at non-AUG start codons in yeast and human cells[3,38]. More efficient initiation at AUG relative to near-cognate start codons is maintained in eukaryotic cells by translation initiation factors. In particular, mutations in eIF1 that decrease its affinity for eIF2β increase the relative efficiency of binding of the ternary complex to a near-cognate UUG start codon[39]. There are some exceptions to the scanning initiation mechanism at non-AUG start codons; for example, in select cases, initiation can occur at CUG codons using leucyl-tRNA and eIF2A in place of eIF2[40]. A more recently reported class of non-AUG translation events, collectively referred to as RAN (repeat-associated non-AUG) translation, is responsible for translation of toxic dipeptide repeat proteins from microsatellite expansions associated with various diseases including Huntington's disease[41], fragile X-associated tremor ataxia syndrome[42], and amyotrophic lateral sclerosis/frontotemporal dementia[43]. RAN translation may initiate at non-AUG codons via cap-dependent scanning mechanisms in multiple reading frames[43–46]; however, the products of RAN translation are very long repeat polypeptides, and to date no smORFs have been reported to undergo RAN translation.

Functions and regulation of sORFs that initiate with non-AUG start codons

Upstream open reading frames (uORFs) are a class of smORFs encoded in 5′ UTRs of mRNAs that function as *cis*-translational regulators in eukaryotic transcripts. AUG-initiating uORFs have been identified in up to 40% of human mRNAs, and generally exert a repressive effect on translation of the downstream, annotated protein-coding CDS due to inefficient reinitiation[47]. However, non-AUG-initiated uORFs are as common in some genomes as AUG-initiated uORFs[16], and near-cognate initiation in the 5′ UTR has a more complex relationship to translation of the downstream gene. For example, during meiosis in yeast, near-cognate uORF translation was positively correlated with translation of the downstream cistron, whereas upstream AUG start codons were repressive[28]. A more recent study confirmed these findings; using a machine learning approach with a statistical control, 982 statistically significant uORFs were identified in 791 *S. cerevisiae* genes[48]. Examination of seven predictions with dual fluorescence reporter plasmid revealed four non-AUG and two AUG uORFs affected downstream coding sequence translation. Further genome-wide analysis revealed that genes with AUG uORFs had significantly lower translation efficiencies than genes without uORFs, while genes with non-AUG uORFs had higher translation efficiencies than genes without uORFs. These results are consistent with previous reports of upregulation of expression of *EPRS* and *GADD45G* during stress conditions[49]. The mechanism of uORF regulation during cellular stress has been recently reviewed[21,50–52]; briefly, cellular stress in eukaryotes induces phosphorylation of the eIF2α subunit of the eIF2 complex, inhibiting its ability to initiate translation of AUG-initiating uORFs. Under these conditions, relative activity of the alternative initiation factor eIF2A is increased; this complex specifically promotes translation from near-cognate start codons in the 5′ UTR as well as downstream canonical protein coding sequences.

Non-AUG uORF expression is also highly regulated and contributes to translational control during tumorigenesis. Ribosome profiling of the epidermis of wild-type and SOX2-expressing mice revealed that the translation of ORFs in the 5′ UTR is particularly differential upon oncogene induction, with a median increase of 1.84-fold[53]. Further pathway analysis revealed that these differential uORFs are co-encoded with downstream ORFs involved in mechanisms of cancer, stem-cell pluripotency and Wnt/β-catenin signaling, suggesting that during tumor initiation, the translational initiation apparatus is redirected towards uORFs of cancer-related mRNAs. Interestingly, the majority of these uORFs initiated with a CUG or GUG start codon and enhanced the translational efficiency of oncogenic mRNAs.

Non-AUG sORFs that exhibit regulated expression have also been observed in prokaryotes. Quantitative proteomics of *E. coli* K-12 revealed two sORFs, *ymcF* and *ynfQ*, that are specifically expressed during cold shock[54]. *ymcF* and *ynfQ* map to intergenic sequences downstream of the known cold shock proteins CspG and CspI, respectively, and are initiated from rare AUU start codons. These two sORFs are conserved in related Gram-negative bacteria, and are predicted to be structured. In a parallel study, a small, membrane-associated sORF, *gndA*, was identified during heat shock, which is encoded within the *gndP* gene in an alternative reading frame[19]. Later, an antibiotic-assisted ribosome profiling study revealed that *gndA* is initiated from a UUG start codon[15]. These non-AUG-initiated sORFs exhibit

regulated expression under specific stress conditions, and, considering their conservation, predicted structure and/or subcellular localization, it is possible that the encoded small proteins are functional *in trans*; however, these functions have not yet been demonstrated.

Several sORFs that initiate at non-AUG start codons in mammalian cells have recently been demonstrated to be functional in disease and stress states. In one example, ribosome profiling of lipopolysaccharide (LPS)-treated mouse macrophages revealed 96 non-canonical smORFs are translated from lncRNAs, 55% of which initiated from non-AUG start codons. A smORF translated from lncRNA Aw112010, which initiates from a CUG start codon, was shown to play an important role in the response to bacterial infection. Disruption of translation of Aw112010 with a stop codon knock-in resulted in accelerated weight loss and higher bacterial burden in the liver and spleen of mice during *S. Typhimurium* infection, compared with wild-type litter-mates[55]. In another example, the mitochondrial ribosomal protein L18 (MRPL18) gene initiates translation from a downstream CUG start codon during heat shock. This alternative translation start site generates a truncated protein lacking the mitochondrial targeting signal, promoting localization of this N-terminaly truncated MRPL18 isoform to the cytoplasm instead of the mitochondria during heat shock. Truncated, cytosolic MRPL18 incorporates into the 80S ribosome and promotes synthesis of heat shock proteins during stress.[56]

Perspective

It is now clear that non-AUG-initiated translation is pervasive and conserved from bacteria to human, and is enriched among sORFs relative to main ORFs. Several functions for translation of sORFs from non-AUG start codons have been proposed. First, uORFs, or sORFs in the 5′ UTR, generally switch from negative to positive translational regulatory roles when they initiate with near cognate, rather than AUG, start codons, especially under developmental and/or stress conditions. Secondly, sORFs that function *in trans* can exhibit regulated translation under stress conditions or other stimuli. It is likely that the start codon identity is involved in tuning the expression level and also condition-specific translational regulation of all of these classes of sORFs.

It is also possible that ubiquitous non-AUG initiation of sORFs has an evolutionary origin. It has been proposed that sORFs represent instances of *de novo* gene birth, in which non-genic smORFs are transcribed and translated at low levels, occasionally acquiring adaptive mutations[13]. This model is consistent with the intermediate conservation of sORFs[3,13], and would also be consistent with enrichment of non-AUG start codons among protogenes, which arise randomly from the genomic sequence and have not yet been selected for optimal[57], AUG-driven translation.

Moving forward, only a handful of non-AUG translation events have been characterized in molecular detail, and non-AUG-mediated translation has been globally profiled under only a few conditions. It is possible that under additional cellular and disease conditions, other non-AUG-initiated ORFs that exhibit specific expression may remain to be identified. Further investigation of the regulation and functions of these sORFs will aid their identification and characterization. Taken together, the preponderance of non-AUG start codons driving sORF

expression suggests that our model of translation initiation, and our understanding of the mechanism by which information is encoded in genomes, must be revised.

Acknowledgments.

This work was supported in part by the Searle Scholars Program, a Smith Family Foundation Odyssey Award, the NIH (R01GM122984), and Yale University West Campus start-up funds (to S.A.S.). X. C. was supported in part by a Rudolph J. Anderson postdoctoral fellowship from Yale University.

References.

1. Jaffe JD, Berg HC & Church GM Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics 4, 59–77, doi:10.1002/pmic.200300511 (2004). [PubMed: 14730672]
2. Ingolia NT, Ghaemmaghami S, Newman JR & Weissman JS Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324, 218–223, doi:10.1126/science.1168978 (2009). [PubMed: 19213877]
3. Slavoff SA et al. Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat Chem Biol 9, 59–64, doi:10.1038/nchembio.1120 (2013). [PubMed: 23160002]
4. Hemm MR, Paul BJ, Schneider TD, Storz G & Rudd KE Small membrane proteins found by comparative genomics and ribosome binding site models. Mol Microbiol 70, 1487–1501, doi:10.1111/j.1365-2958.2008.06495.x (2008). [PubMed: 19121005]
5. Kastenmayer JP et al. Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae. Genome Res 16, 365–373, doi:10.1101/gr.4355406 (2006). [PubMed: 16510898]
6. Vanderperre B et al. Direct detection of alternative open reading frames translation products in human significantly expands the proteome. PLoS One 8, e70698, doi:10.1371/journal.pone.0070698 (2013). [PubMed: 23950983]
7. Menschaert G et al. Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. Molecular & Cellular Proteomics : MCP 12, 1780–1790, doi:10.1074/mcp.M113.027540 (2013). [PubMed: 23429522]
8. Harrison PM, Kumar A, Lang N, Snyder M & Gerstein M A question of size: the eukaryotic proteome and the problems in defining it. Nucleic Acids Research 30, 1083–1090, doi:10.1093/nar/30.5.1083 (2002). [PubMed: 11861898]
9. Orr MW, Mao Y, Storz G & Qian SB Alternative ORFs and small ORFs: shedding light on the dark proteome. Nucleic Acids Res, doi:10.1093/nar/gkz734 (2019).
10. Makarewich CA et al. MOXI Is a Mitochondrial Micropeptide That Enhances Fatty Acid beta-Oxidation. Cell Rep 23, 3701–3709, doi:10.1016/j.celrep.2018.05.058 (2018). [PubMed: 29949755]
11. Hemm MR et al. Small stress response proteins in Escherichia coli: proteins missed by classical proteomic studies. Journal of Bacteriology 192, 46–58, doi:10.1128/JB.00872-09 (2010). [PubMed: 19734316]
12. D'Lima NG et al. A human microprotein that interacts with the mRNA decapping complex. Nature Chemical Biology 13, 174–180, doi:10.1038/nchembio.2249 (2017). [PubMed: 27918561]
13. Carvunis AR et al. Proto-genes and de novo gene birth. Nature 487, 370–374, doi:10.1038/nature11184 (2012). [PubMed: 22722833]
14. Saghatelian A & Couso JP Discovery and characterization of smORF-encoded bioactive polypeptides. Nat Chem Biol 11, 909–916, doi:10.1038/nchembio.1964 (2015). [PubMed: 26575237]
15. Meydan S et al. Retapamulin-Assisted Ribosome Profiling Reveals the Alternative Bacterial Proteome. Molecular Cell 74, 481–493 e486, doi:10.1016/j.molcel.2019.02.017 (2019). [PubMed: 30904393]

16. Lee S et al. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proc Natl Acad Sci U S A 109, E2424–2432, doi:10.1073/pnas.1207846109 (2012). [PubMed: 22927429]
17. Hecht A et al. Measurements of translation initiation from all 64 codons in E. coli. Nucleic Acids Res 45, 3615–3626, doi:10.1093/nar/gkx070 (2017). [PubMed: 28334756]
18. Wei J, Zhang Y, Ivanov IP & Sachs MS The stringency of start codon selection in the filamentous fungus Neurospora crassa. J Biol Chem 288, 9549–9562, doi:10.1074/jbc.M112.447177 (2013). [PubMed: 23396971]
19. Yuan P, D'Lima NG & Slavoff SA Comparative Membrane Proteomics Reveals a Nonannotated E. coli Heat Shock Protein. Biochemistry 57, 56–60, doi:10.1021/acs.biochem.7b00864 (2018). [PubMed: 29039649]
20. Ingolia NT, Lareau LF & Weissman JS Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell 147, 789–802, doi:10.1016/j.cell.2011.10.002 (2011). [PubMed: 22056041]
21. Kearse MG & Wilusz JE Non-AUG translation: a new start for protein synthesis in eukaryotes. Genes Dev 31, 1717–1731, doi:10.1101/gad.305250.117 (2017). [PubMed: 28982758]
22. Hinnebusch AG Molecular mechanism of scanning and start codon selection in eukaryotes. Microbiology and Molecular Biology Reviews : MMBR 75, 434–467, first page of table of contents, doi:10.1128/MMBR.00008-11 (2011). [PubMed: 21885680]
23. Florkiewicz RZ & Sommer A Human basic fibroblast growth factor gene encodes four polypeptides: three initiate translation from non-AUG codons. Proceedings of the National Academy of Sciences of the United States of America 86, 3978–3981, doi:10.1073/pnas.86.11.3978 (1989). [PubMed: 2726761]
24. Arnaud E et al. A new 34-kilodalton isoform of human fibroblast growth factor 2 is cap dependently synthesized by using a non-AUG start codon and behaves as a survival factor. Molecular and Cellular Biology 19, 505–514, doi:10.1128/mcb.19.1.505 (1999). [PubMed: 9858574]
25. Hann SR, King MW, Bentley DL, Anderson CW & Eisenman RN A non-AUG translational initiation in c-myc exon 1 generates an N-terminally distinct protein whose synthesis is disrupted in Burkitt's lymphomas. Cell 52, 185–195, doi:10.1016/0092-8674(88)90507-7 (1988). [PubMed: 3277717]
26. Hann SR, Sloan-Brown K & Spotts GD Translational activation of the non-AUG-initiated c-myc 1 protein at high cell densities due to methionine deprivation. Genes & Development 6, 1229–1240, doi:10.1101/gad.6.7.1229 (1992). [PubMed: 1628829]
27. Villegas A & Kropinski AM An analysis of initiation codon utilization in the Domain Bacteria - concerns about the quality of bacterial genome annotation. Microbiology 154, 2559–2661, doi:10.1099/mic.0.2008/021360-0 (2008). [PubMed: 18757789]
28. Brar GA et al. High-resolution view of the yeast meiotic program revealed by ribosome profiling. Science 335, 552–557, doi:10.1126/science.1215110 (2012). [PubMed: 22194413]
29. Weaver J, Mohammad F, Buskirk AR & Storz G Identifying Small Proteins by Ribosome Profiling with Stalled Initiation Complexes. MBio 10, doi:10.1128/mBio.02819-18 (2019).
30. Gao X et al. Quantitative profiling of initiating ribosomes in vivo. Nature Methods 12, 147–153, doi:10.1038/nmeth.3208 (2015). [PubMed: 25486063]
31. Ma J et al. Improved Identification and Analysis of Small Open Reading Frame Encoded Polypeptides. Anal Chem 88, 3967–3975, doi:10.1021/acs.analchem.6b00191 (2016). [PubMed: 27010111]
32. Ma J et al. Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. J Proteome Res 13, 1757–1765, doi:10.1021/pr401280w (2014). [PubMed: 24490786]
33. Sussman JK, Simons EL & Simons RW Escherichia coli translation initiation factor 3 discriminates the initiation codon in vivo. Molecular Microbiology 21, 347–360, doi:10.1046/j.1365-2958.1996.6371354.x (1996). [PubMed: 8858589]
34. Peabody DS Translation initiation at non-AUG triplets in mammalian cells. J Biol Chem 264, 5031–5035 (1989). [PubMed: 2538469]

35. Ivanov IP, Loughran G, Sachs MS & Atkins JF Initiation context modulates autoregulation of eukaryotic translation initiation factor 1 (eIF1). Proceedings of the National Academy of Sciences of the United States of America 107, 18056–18060, doi:10.1073/pnas.1009269107 (2010). [PubMed: 20921384]
36. Gualerzi CO & Pon CL Initiation of mRNA translation in bacteria: structural and dynamic aspects. Cell Mol Life Sci 72, 4341–4367, doi:10.1007/s00018-015-2010-3 (2015). [PubMed: 26259514]
37. Svidritskiy E & Korostelev AA Ribosome Structure Reveals Preservation of Active Sites in the Presence of a P-Site Wobble Mismatch. Structure 23, 2155–2161, doi:10.1016/j.str.2015.08.011 (2015). [PubMed: 26412335]
38. Zitomer RS, Walthall DA, Rymond BC & Hollenberg CP Saccharomyces cerevisiae ribosomes recognize non-AUG initiation codons. Molecular and Cellular Biology 4, 1191–1197, doi:10.1128/mcb.4.7.1191 (1984). [PubMed: 6390186]
39. Thakur A, Marler L & Hinnebusch AG A network of eIF2beta interactions with eIF1 and Met-tRNAi promotes accurate start codon selection by the translation preinitiation complex. Nucleic Acids Research 47, 2574–2593, doi:10.1093/nar/gky1274 (2019). [PubMed: 30576497]
40. Starck SR et al. Leucine-tRNA initiates at CUG start codons for protein synthesis and presentation by MHC class I. Science 336, 1719–1723, doi:10.1126/science.1220270 (2012). [PubMed: 22745432]
41. Banez-Coronel M et al. RAN Translation in Huntington Disease. Neuron 88, 667–677, doi:10.1016/j.neuron.2015.10.038 (2015). [PubMed: 26590344]
42. Todd PK et al. CGG repeat-associated translation mediates neurodegeneration in fragile X tremor ataxia syndrome. Neuron 78, 440–455, doi:10.1016/j.neuron.2013.03.026 (2013). [PubMed: 23602499]
43. Cleary JD, Pattamatta A & Ranum LPW Repeat-associated non-ATG (RAN) translation. The Journal of Biological Chemistry 293, 16127–16141, doi:10.1074/jbc.R118.003237 (2018). [PubMed: 30213863]
44. Mori K et al. The C9orf72 GGGGCC repeat is translated into aggregating dipeptide-repeat proteins in FTLD/ALS. Science 339, 1335–1338, doi:10.1126/science.1232927 (2013). [PubMed: 23393093]
45. Ash PE et al. Unconventional translation of C9ORF72 GGGGCC expansion generates insoluble polypeptides specific to c9FTD/ALS. Neuron 77, 639–646, doi:10.1016/j.neuron.2013.02.004 (2013). [PubMed: 23415312]
46. Tabet R et al. CUG initiation and frameshifting enable production of dipeptide repeat proteins from ALS/FTD C9ORF72 transcripts. Nature Communications 9, 152, doi:10.1038/s41467-017-02643-5 (2018).
47. Calvo SE, Pagliarini DJ & Mootha VK Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc Natl Acad Sci U S A 106, 7507–7512, doi:10.1073/pnas.0810916106 (2009). [PubMed: 19372376]
48. Spealman P et al. Conserved non-AUG uORFs revealed by a novel regression analysis of ribosome profiling data. Genome Research 28, 214–222, doi:10.1101/gr.221507.117 (2018). [PubMed: 29254944]
49. Young SK & Wek RC Upstream Open Reading Frames Differentially Regulate Gene-specific Translation in the Integrated Stress Response. The Journal of Biological Chemistry 291, 16927–16935, doi:10.1074/jbc.R116.733899 (2016). [PubMed: 27358398]
50. Wek RC, Jiang HY & Anthony TG Coping with stress: eIF2 kinases and translational control. Biochem Soc Trans 34, 7–11, doi:10.1042/BST20060007 (2006). [PubMed: 16246168]
51. Jackson RJ, Hellen CU & Pestova TV The mechanism of eukaryotic translation initiation and principles of its regulation. Nat Rev Mol Cell Biol 11, 113–127, doi:10.1038/nrm2838 (2010). [PubMed: 20094052]
52. Khitun A, Ness TJ & Slavoff SA Small open reading frames and cellular stress responses. Mol Omics 15, 108–116, doi:10.1039/c8mo00283e (2019). [PubMed: 30810554]
53. Sendoel A et al. Translation from unconventional 5′ start sites drives tumour initiation. Nature 541, 494–499, doi:10.1038/nature21036 (2017). [PubMed: 28077873]

54. D'Lima NG et al. Comparative Proteomics Enables Identification of Nonannotated Cold Shock Proteins in E. coli. J Proteome Res 16, 3722–3731, doi:10.1021/acs.jproteome.7b00419 (2017). [PubMed: 28861998]

55. Jackson R et al. The translation of non-canonical open reading frames controls mucosal immunity. Nature 564, 434–438, doi:10.1038/s41586-018-0794-7 (2018). [PubMed: 30542152]

56. Zhang X et al. Translational control of the cytosolic stress response by mitochondrial ribosomal protein L18. Nature Structural & Molecular Biology 22, 404–410, doi:10.1038/nsmb.3010 (2015).

57. Belinky F, Rogozin IB & Koonin EV Selection on start codons in prokaryotes and potential compensatory nucleotide substitutions. Scientific Reports 7, 12422, doi:10.1038/s41598-017-12619-6 (2017). [PubMed: 28963504]

Table 1.

Frequency and efficiency of non-AUG start codon utilization in bacterial and eukaryotic cells from selected sources.

Observed frequency of non-AUG start codon occupancy by ribosome profiling in (A) retapamulin-treated bacterial cells[15] and (B) lactidomycin-treated human and mouse cells[16]. Efficiency of non-AUG start codon use, expressed as percentage AUG efficiency, in (C) a GFP reporter gene assay inside living bacterial cells[17] and (D) a firefly luciferase reporter gene assay in *Neurospora crassa*[18].

A		U	C	A	G		B		U	C	A	G	
	U	0.04%	NR	0.04%	NR	U		U	0.22%	0.08%	NR	NR	U
		0.09%	NR	0.02%	NR	C			0.29%	0.13%	0.23%	NR	C
		0.06%	NR	0.02%	NR	A			0.09%	0.09%	0.01%	NR	A
		5.62%	NR	NR	NR	G			4.17%	0.13%	<0.01%	NR	G
	C	NR	NR	NR	NR	U		C	0.05%	0.11%	NR	0.07%	U
		NR	NR	NR	0.02%	C			0.25%	0.16%	0.25%	0.42%	C
		0.02%	NR	0.06%	NR	A			0.16%	0.28%	0.07%	0.12%	A
		3.12%	0.02%	NR	NR	G			15.44%	0.29%	0.51%	0.11%	G
	A	6.18%	NR	0.02%	0.02%	U		A	3.05%	0.13%	NR	0.04%	U
		4.12%	0.06%	0.04%	0.04%	C			2.71%	0.42%	0.32%	0.18%	C
		0.02%	0.06%	0.13%	0.06%	A			1.29%	0.17%	1.34%	0.15%	A
		69.47%	0.02%	0.02%	0.02%	G			49.76%	3.89%	1.00%	0.91%	G
	G	NR	0.02%	NR	0.02%	U		G	NR	0.22%	NR	0.10%	U
		NR	NR	NR	NR	C			0.25%	0.49%	0.31%	0.39%	C
		0.07%	0.02%	0.02%	NR	A			0.12%	0.32%	0.31%	0.31%	A
		10.40%	0.06%	0.02%	0.02%	G			7.17%	0.38%	0.44%	0.10%	G

C		U	C	A	G		D		U	C	A	G	
	U	0.10%	0.11%	0.58%	0.07%	U		U	NR	NR	NR	NR	U
		0.05%	0.07%	0.10%	0.07%	C			NR	NR	NR	NR	C
		0.18%	0.08%	0.09%	0.08%	A			NR	NR	NR	NR	A
		63.99%	0.11%	0.10%	0.17%	G			8.30%	NR	NR	NR	G
	C	0.04%	0.05%	0.39%	0.07%	U		C	NR	NR	NR	NR	U
		0.08%	0.10%	0.11%	0.08%	C			NR	NR	NR	NR	C
		0.11%	0.06%	0.06%	0.06%	A			NR	NR	NR	NR	A
		2.72%	0.05%	0.05%	0.06%	G			18.20%	NR	NR	NR	G
	A	1.92%	0.12%	0.35%	0.08%	U		A	5.90%	NR	NR	NR	U
		1.03%	0.11%	0.08%	0.07%	C			2.90%	NR	NR	NR	C
		1.84%	0.10%	0.10%	0.09%	A			5.20%	NR	NR	0.01%	A
		100.00%	0.42%	0.05%	0.10%	G			100.00%	4.50%	0.01%	0.02%	G
	G	0.55%	0.13%	0.34%	0.17%	U		G	NR	NR	NR	NR	U
		0.15%	0.07%	0.15%	0.11%	C			NR	NR	NR	NR	C

		0.63%	0.15%	0.43%	0.37%	A			NR	NR	NR	NR	A
		121.84%	0.26%	0.23%	0.23%	G			11.40%	NR	NR	NR	G