Ribosome Decision Graphs for the Representation of Eukaryotic RNA Translation Complexity

Jack A. S. Tierney[1,2,#], Michał Świrski[3,#], Håkon Tjeldnes[1,4], Jonathan M. Mudge[5], Joanna Kufel[3], Nicola Whiffin[6], Eivind Valen[4,7], Pavel V. Baranov[1].

[1] School of Biochemistry and Cell Biology, University College Cork, Cork, Ireland.

[2] SFI Centre for Research Training in Genomics Data Science, University College Cork, Cork, Ireland

[3] Institute of Genetics and Biotechnology, Faculty of Biology, University of Warsaw, Warsaw, Poland.

[4] Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, Norway.

[5] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK.

[6] The Big Data Institute and Wellcome Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, UK.

[7] Department of Biosciences, University of Oslo, Oslo, Norway.

[#] These authors contributed equally to this work.

Corresponding author: p.baranov@ucc.ie

1

Abstract

The application of ribosome profiling has revealed an unexpected abundance of translation in addition to that responsible for the synthesis of previously annotated protein-coding regions. Multiple short sequences have been found to be translated within single RNA molecules, both within annotated protein-coding and non-coding regions. The biological significance of this translation is a matter of intensive investigation. However, current schematic or annotation-based representations of mRNA translation generally do not account for the apparent multitude of translated regions within the same molecules. They also do not take into account the stochasticity of the process that allows alternative translations of the same RNA molecules by different ribosomes. There is a need for formal representations of mRNA complexity that would enable the analysis of quantitative information on translation and more accurate models for predicting the phenotypic effects of genetic variants affecting translation. To address this, we developed a conceptually novel abstraction that we term Ribosome Decision Graphs (RDGs). RDGs represent translation as multiple ribosome paths through untranslated and translated mRNA segments. We termed the later 'translons'. Non-deterministic events, such as initiation, re-initiation, selenocysteine insertion or ribosomal frameshifting are then represented as branching points. This representation allows for an adequate representation of eukaryotic translation complexity and focuses on locations critical for translation regulation. We show how RDGs can be used for depicting translated regions, analysis of genetic variation and quantitative genome-wide data on translation for characterisation of regulatory modulators of translation.

Motivation.

Until relatively recently, available experimental evidence suggested that in eukaryotes each mRNA encoded only a single protein. Since only a single coding region was therefore expected to be translated this region was conventionally termed the CDS (CoDing Sequence). This view has been challenged by the development of the ribosome profiling technique which enables isolation and sequencing of RNA fragments protected by ribosomes and hence detection of regions being translated (1). Numerous ribosome profiling studies carried out in cells from a variety of eukaryotes unexpectedly revealed abundant translation outside of CDS regions. This included the translation of short sequences in the supposedly untranslated regions (UTRs) of mRNAs found upstream and downstream of the CDS, as well as in so-called non-coding RNAs, especially long non-coding RNAs (lncRNAs) (2–11). These studies also demonstrated the translation of N-terminally extended CDS regions due to initiation at upstream non-AUG start codons (12) or C-terminally extended due to stop codon readthrough (13). A certain group of eukaryotic organisms (ciliates *Euplotes*) were found to utilize ribosomal frameshifting in thousands of their genes (14). While most of these phenomena were first described prior to the advent of ribosome profiling (15–19), they were considered rare. Certainly, very few cases have been catalogued by reference gene annotation projects, and no conventional abstraction has been developed to represent this translation complexity in annotations, schematic scientific diagrams or analytical workflows. The lack of a formal framework for the representation of this complexity hampers our ability to generate accurate and biologically realistic annotations of translated sequences and to design mathematical models and computer simulations. In its absence, it is difficult or even impossible to quantitatively characterize multiple translation events and define their interrelationships.

To address this challenge, we developed a conceptually novel framework for abstract representation of translation complexity, which we term Ribosome Decision Graphs (RDGs). RDGs solve many problems, such as the representation of multiple translated regions in the same mRNAs and alternative decoding mechanisms producing multiple proteoforms. We show how RDGs can be used for the accurate depiction of productive and non-productive RNA translation (i.e. translation that does or does not lead to the production of a protein molecule), analysis of quantitative information on translation, and genetic variants affecting mRNA decoding.

Open Reading Frames are inadequate to represent the complexity of mRNA translation

The development of a conventional abstraction is undermined by the ambiguity of the terms used to define translated regions. For example, while translated regions are often described as Open Reading Frames (ORFs) in literature or scientific discourse, gene annotation projects typically utilise only the term CDS, and only for regions considered to be protein-coding. Instead, an ORF would be regarded by implication as a potential translation that can be identified *in silico*. Here, we in effect consider three concepts in an attempt at unification: (1) that ORFs can be identified *in silico* whether or not they have evidence of

translation, (2) that ORFs may undergo translation that does not lead to production of a stable, functional protein, and (3) that ORFs which are known to be translated into proteins should alone be considered as CDS. In other words, most CDS are ORFs, but not all ORFs are CDS. In general, there are two definitions of ORFs, start to stop (Start-Stop) and stop to stop (Stop-Stop)(20), as depicted in Figure 1A. Plotting the locations of potential start codons (usually AUGs) and stop codons in three reading frames is undoubtedly highly instrumental for examining potentially translated sequences. However, the common interpretation of nucleic acid sequences in terms of "translated ORFs" is superficial, frequently inaccurate, and often leads to confusion as illustrated in Figure 1B-D.

Perhaps the most frequent source of alternative translation in many eukaryotes, including humans, is the multiplicity of translation initiation sites. It arises predominantly from two common mechanisms involved in the selection of translation initiation sites: leaky scanning and re-initiation. *Leaky scanning* refers to the inefficient recognition of a start codon by the ribosome resulting in the ribosome scanning complex scanning through the start codon and effectively ignoring it (21). Generally, ribosome scanning complexes assemble at the 5' cap of RNAs and move along the transcript in the 3' direction until they encounter a start site and initiate translation (22–24). However, recognition of start sites is a sequence-dependent stochastic process, in which usually only a proportion of scanning complexes finally initiate. Many factors play a role in determining the efficiency with which a ribosome initiates translation at a given codon. These include the identity of the codon and its surrounding sequence (known as the Kozak context) (25), as well as the dwell time of the scanning ribosome at that codon (26). Unless the combination of these factors is strictly optimal for initiation, at least a small fraction of scanning complexes will bypass the potential start site and continue scanning, allowing translation to be initiated further downstream. When a potential initiation site is a non-AUG codon or an AUG in a weak Kozak context, only a small proportion of scanning complexes will initiate translation. Thus, leaky scanning may result in the translation of different coding sequences using numerous initiation sites, while initiation at start codons in the same reading frame can give rise to proteoforms with alternative N-termini (PANTs) (Figure 1B). A potentially large number of start codons may be used to initiate translate within the same Stop-Stop ORF, as is the case with the well-explored human *PTEN* gene where functionally distinct extended proteoforms are produced from multiple non-AUG starts (27). Annotating all Start-Stop ORFs is problematic due to the large number of potential start codons, and in certain cases, such as Repeat Associated Non-AUG (RAN) translation (28,29), the exact position of the initiation site cannot even be easily identified.

In the case of stop codon readthrough (13,30) and selenocysteine incorporation (31,32) (Figure 1C), the CDS needs to be described as a fusion of a Start-Stop ORF with a Stop-Stop ORF (gene annotation projects currently resolve these cases by 're-writing' the Stop codon or selenocysteine codon in the protein file, allowing them to code through). For programmed ribosomal frameshifting (Figure 1D), which is common in viruses but also infrequently occurs in cellular genes (33), the description of translation using ORFs would require the introduction of the location of the frameshift site as both start and stop codon. This could enable the designation of the trans-frame protein product as a fusion of the two such "ORFs". In practice, gene annotation may instead introduce an artificial indel

modification of the natural DNA/RNA sequences to yield a single contiguous ORF; for example, the [T] corresponding to human hg38 assembly chr19:2271440 nucleotide is deleted in both RefSeq (e.g. NM_004152.3) and Ensembl (e.g. ENST00000582888.8). To this end, existing gene annotation of in silico trans-frame translation may yield a protein sequence corresponding to the product generated in nature. However, it comes at the expense of producing an incorrect sequence of an mRNA molecule, which does not allow for the regulatory mechanism at play to be accurately represented.
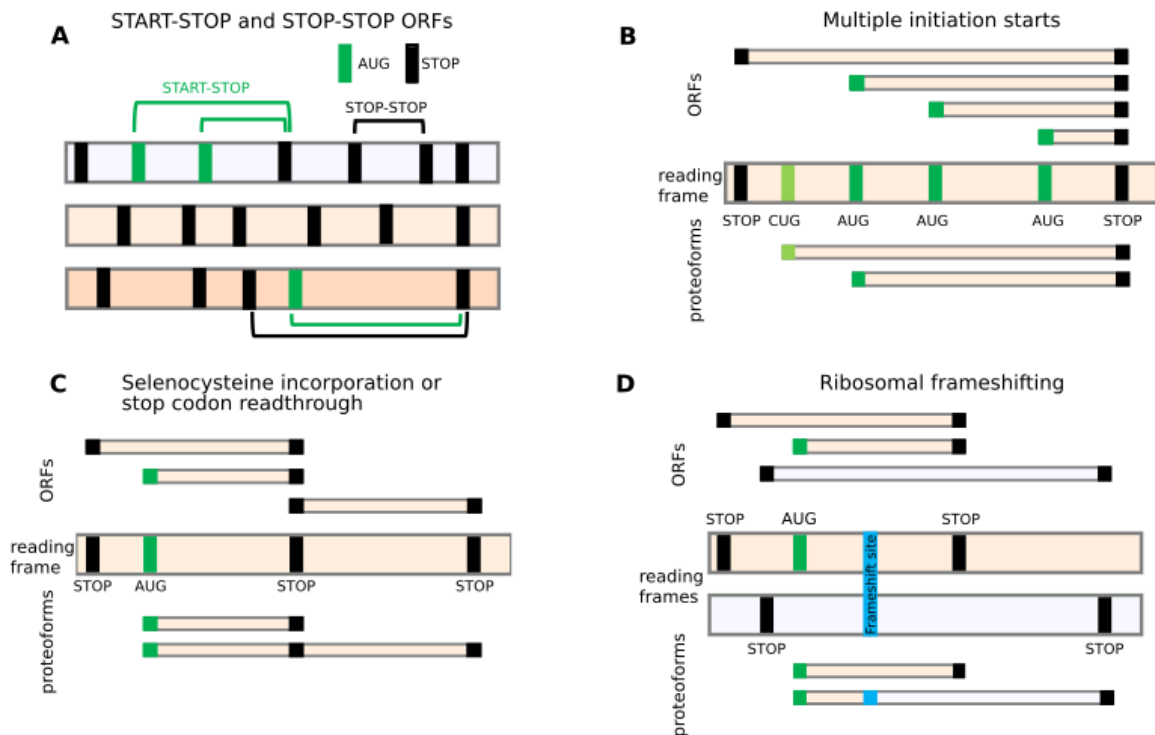

Figure 1. ORFs do not adequately represent translational complexity. A. Two formal definitions of ORFs. Three reading frames are shown as horizontal bars with vertical bars corresponding to AUG (green) or stop codons (black). Several examples of Start-Stop (green arcs) and Stop-Stop (black arcs) ORFs are shown. **B-D.** The relationship between ORFs (top) and expressed proteoforms (bottom) for mRNAs with different locations of starts and stops (middle). Only two relevant reading frames are shown for simplicity. **B.** An RNA encoding two proteoforms with alternative N-termini due to utilization of two start codons. Due to multiple potential AUG codons, there are many Start-Stop ORFs whose conceptual translation does not correspond to encoded proteoforms. Stop-Stop ORF does not reflect the existence of alternative proteoforms. **C.** In the case of stop codon readthrough or selenocysteine insertion, ribosomes may read through specific stop codons by incorporating an amino acid yielding a product (purple) that cannot be described as a product of a single ORF. **D.** Similarly, ribosomal frameshifting generates a trans-frame protein (blue) that does not represent a product of a single ORF.

The examples in Figure 1 are not exhaustive and there are other translation phenomena that cannot be easily described using ORFs, such as translational bypassing (34–36) and StopGo (also known as StopCarryOn or 2A) (37). Regardless of which ORF definition

is used, the concept of a translated ORF is not adequate to represent the complexity of RNA translation.

RNA translation is segmented

Ribosome profiling has revealed the existence of a large number of short translated sequences, currently termed small or short ORFs (smORFs, sORFs) or RiboSeq ORFs, as the term CDS is reserved for sequences encoding classical proteins (38). Many RiboSeq ORFs occur within the same RNA molecules. The lack of appropriate terminology reflecting the complexity of translation becomes even more evident when we consider the relationship between these translation segments. Upstream translation often influences downstream translation and this dependency is known to be utilised to regulate gene expression. For instance, many short translated regions upstream of CDSs (termed upstream ORFs, or uORFs) have been found to regulate translation by blocking ribosomes *via* sensing specific metabolites within the nascent peptide channel (39–43), reviewed in (44). This process is exemplified by translation regulation of the downstream CDS by a short uORF in vertebrate *AMD1* encoding Adenosyl-Methionin Decarboxylase, a key enzyme in polyamine biosynthesis. The uORF encodes a short peptide MAGDIS that stalls ribosomes through its interactions with the ribosome in the presence of polyamines (45). These stalled ribosomes prevent other ribosomes from binding and scanning downstream to initiate at *AMD1* CDS. Thus, the uORF provides a negative feedback control mechanism for *AMD1* expression, inhibiting its synthesis when polyamine concentration is high, but allowing for its synthesis when polyamine levels decrease (Figure 2A).

In addition to leaky scanning, re-initiation is another process impacting start codon selection. Translation re-initiation occurs when small ribosomal subunits remain bound to the mRNA after translation is complete and reinitiate downstream of the terminating stop codon. This is thought to be common after the translation of short ORFs as it takes time for initiation factors to dissociate from the ribosome. In this way, the ribosome may remain capable of initiation after translating a small number of codons, although other factors are known to contribute to this process, allowing for re-initiation in some instances even after the translation of long ORFs. The detailed molecular mechanisms of these processes are described in dedicated reviews (21,22,24,46–49). Re-initiation provides a platform for a rapid switch of gene expression on the translational level. Perhaps the most thoroughly studied is the case of delayed re-initiation (50–52) that protects translation of certain mRNAs (e.g. human A*TF4*, yeast *GCN4*) from downregulation during the Integrated Stress Response (ISR) (53,54). Under this condition, the reduced availability of the ternary complex (tRNAi* eIF2*GTP) increases the time required for post-terminating ribosomes to bind the ternary complex enabling re-initiation. Therefore, the level of stress determines the location of the start codon at which re-initiation occurs. Figure 2B provides a schematic illustrating this mechanism.

It Is unclear to what extent the translated products of such regulatory translation contribute to the functional cellular proteome beyond their potential contribution to the

antigen pool, as many of them lack conservation at the protein level (38,55,56). Extreme cases of translation regulation without peptide synthesis are represented by minimal ORFs consisting of a start codon immediately followed by a stop codon. While they obviously do not produce any functional peptide, some of them do have regulatory potential as strong ribosome stalling sites (57).
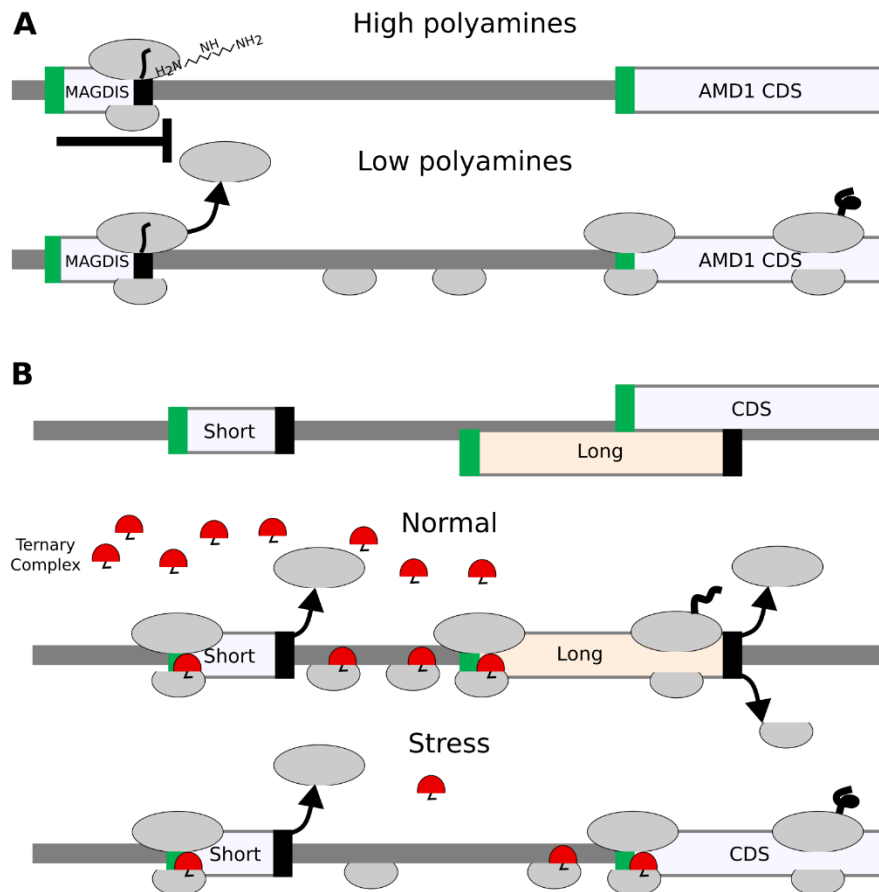

Figure 2. Relationship between translated segments within the same RNA. A. A schematic of metabolite-dependent translation regulation *via* ribosome arrest at uORF, such as in *AMD1* mRNA. In the presence of high concentration of polyamines ribosome with MAGDIS peptide stalls at the end of mRNA. **B.** A schematic of delayed re-initiation mechanism enabling translation of selected mRNAs during global translation suppression caused by eIF2alpha phosphorylation during Integrated Stress Response. When eIF2 is phosphorylated the concentration of the eIF2*tRNAi*GTP ternary complex decreases and it requires a longer time and distance for the scanning ribosome complex to acquire it. As a result the long uORF is bypassed and initiation occurs at CDS.

It is clear that translation complexity requires a unified and comprehensive abstraction that would adequately represent all translated regions – not only those that encode classical proteins – and reflect their mechanistic interrelationships. Such representation should be convenient for use by both, human scientists when examining

individual mRNA sequences and computer agents during programmatic analysis of large datasets.

Ribosome Paths

The complex nature of translational events and regulatory processes reveals the need to consider the entire passage of an individual ribosomal complex containing the same small ribosomal subunit along the mRNA, from the moment of pre-initiation complex assembly at the 5' cap (or IRES element) to the complete dissociation of both ribosomal subunits from the mRNA as a functional unit. We propose to term such a unit a Ribosome Path (RiboPath). It includes both regions that are scanned and those that are translated. As argued above, ORF is an inadequate descriptor of translated regions, and therefore we want to define and assign a new, unambiguous name to an entity denoting *translated region* as encompassing the entire sequence of RNA translated by a fully assembled elongating ribosome from initiation codon through termination and dissociation of the large ribosomal subunit. We term this region *translon*. It has already been suggested as a term specifying a unit of translation (58) but has not yet been adopted. The main advantage of *translon* over ORF is that it is not constrained by the sequence (specific codons as boundaries). It is based on the process of translation and therefore may incorporate a variety of decoding mechanisms such as ribosomal frameshifting, stop codon readthrough, translational bypassing, etc (59,60). The other term commonly used to indicate translated regions is *cistron*, e.g. polycistronic or monocistronic mRNAs. However, this term was originally defined genetically, different cistrons should be responsible for different phenotypes, and it is being used inconsistently in the literature.

To simplify the introduction of the RiboPath concept, for now, we only consider initiation and re-initiation as the mechanisms producing alternative proteoforms. We will exclude other translation mechanisms. Nevertheless, our framework can easily be extended to incorporate other translation mechanisms as we discuss later.

Figure 3 illustrates the RiboPath concept with an example of an mRNA encoding two proteoforms arising from alternative CUG and AUG initiation sites in one reading frame (cream) and a single upstream AUG codon in another frame (light lavender) as depicted in the ORF plot at the top. The corresponding translons are shown beneath. Alternative initiation and re-initiation allow the ribosome to pass through five different RiboPaths. The top RiboPath represents the ribosomes that initiate at the first AUG, but fail to reinitiate further downstream, resulting in a path with a single translon T1. The second path corresponds to the ribosome that successfully re-initiates downstream, thus containing two translons, T1 and T2. In the third RiboPath the ribosomes fail to initiate at the first AUG, but start translation at the CUG allowing for translon T3, which encodes an N-terminally extended proteoform relative to the product of translon T2. The fourth RiboPath corresponds to the ribosomes that fail to initiate at both the first AUG and the CUG but succeed at initiating at the second AUG so that its RiboPath consists of only one translon T2. Finally, the fifth RiboPath is unproductive and represents the ribosomes that have not initiated protein synthesis on this mRNA. The RiboPath presentation makes it clear that certain translons are mutually exclusive as they never occur on the same path, e.g. a single ribosome cannot translate T1 and T3.
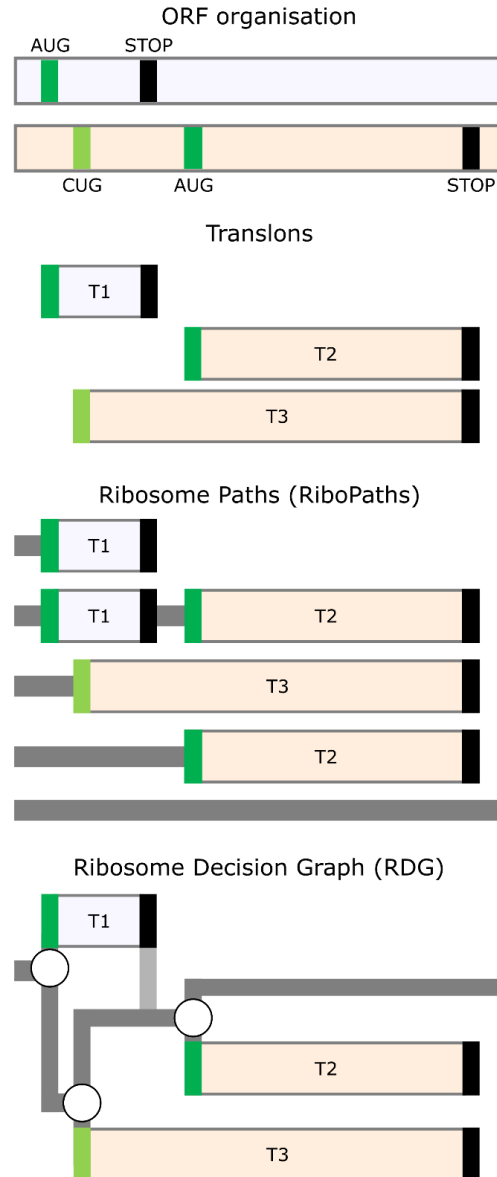
Figure 3. Ribosome paths through mRNA. Top: An example of an RNA with three start codons (green for AUG and light green for CUG) located in two different reading frames depicted as differentially shaded horizontal bars. The translation of this mRNA is represented as a set of translons **below**, Ribosome Paths **further below** and Ribosome Decision Graphs at the **bottom.** RNA regions scanned by the ribosome are shown in dark grey, vertical path in light grey represents the post-terminating small ribosome subunit that continues scanning and remains initiation-competent.

Ribosome Decision Graphs

Once we represent the behaviour of translating ribosomes in terms of paths, it is only natural to further represent these in terms of graphs (Figure 3). The three initiation events in Figure 3 can be represented as branching points where the ribosome makes a "decision" of whether to initiate or not. We do not imply that ribosomes have free will; the decision is likely determined by the molecular composition and temporal thermodynamics of the local microstate. As in statistical mechanics, for practical purposes, it is appropriate to describe

such decisions probabilistically, even if the underlying molecular processes are deterministic. The mRNA region engaged by ribosomes in Figure 3 can then be represented as a graph with three branching points. Stop codons in this graph are considered as deterministic ends of translons as we exclude the possibility of stop codon readthrough or re-initiation after long translons in our illustrative example. Following this notation, any translated RNA can be represented as a Ribosome Decision Graph (RDG).

As in the representation of translation using ORFs/CDS, RDGs may be either conceptual (representing potential) or real (e.g., experimentally supported). In the case of conceptual RDGs, all potential start codons in mRNAs could be used as branching points, e.g., all AUGs, all CUGs, etc. depending on the specific parameters of the model. Such conceptual RDGs would be very complex graphs with a large number of branching points and possible paths. They are not suitable for evaluation by humans, but provide a straightforward method for generating all theoretically possible products of RNA translation. This can be used for the subsequent mining of mass spectrometry datasets. A set of graphs with branching points sampled from the set of all possible branching points can be used to generate simulated ribosome profiling data. The comparison of simulated and real data would enable the determination of the best RDG fitting the experimental data, thus inferring the real branching points from the data. As exemplified further below, RDGs may also be useful for analysing the impact of genetic variation, because variants that change or introduce new branching points (start and stop codons, frameshifts, etc) would alter the RDG topology.

RDGs could also be used to annotate experimentally validated translations. In this case, only those translation events for which there is experimental evidence will be introduced as branching points. In most cases, these experimentally informed RDGs would be suitable for manual examination by researchers and would overcome the limitations of the data structures that are currently used for protein coding annotation.

Implementations of RDGs

To illustrate how RDGs can be used to represent the impact of variation within 5' leader sequences (a.k.a. 5' UTR) on downstream translation we selected the *NF2* variant responsible for neurofibromatosis type 2 (61). The 5' leader sequence of the *NF2* mRNA contains an AUG start codon followed by an in-frame AUG codon in a strong Kozak context. This suggests that few (if any) ribosomes reach the CDS start *via* leaky scanning. It is far more likely that CDS translation involves re-initiation at the CDS start as depicted in Figure 4A. A single base insertion variant was identified in two unrelated individuals in a cohort of 1,134 individuals diagnosed with neurofibromatosis type 2 (ENST00000338641:−66-65insT; GRCh37:chr22:29999922A >AT) (61). This insertion causes both a shift in the reading frame and the introduction of another AUG. The shift extends translons T1 and T2, abrogating the initiation of translon T3 corresponding to *NF2* CDS (Figure 4B).

To illustrate how RDGs can be used for the representation of real translation data we chose two simple examples, namely human *NRAS* and *NXT1* mRNAs. The criteria for this selection were the existence of only a single transcript per gene according to GENCODE v.42

(43), and the ribosome profiling supporting translation of only a single AUG-initiated translon in addition to the annotated CDS. Of note, translation of most human 5' mRNA leaders is more complex, and therefore the advantages of using RDG representation for these are even greater, but may not be suitable for introducing this concept.
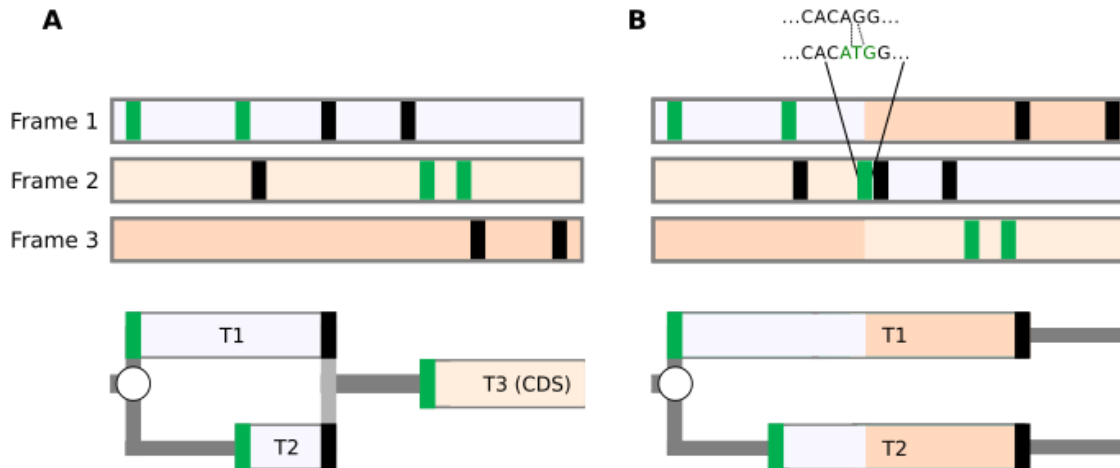

Figure 4: Representation of the effect of a genomic variant on CDS translation. Representations of the *NF2* mRNA for the reference sequence (**A**) and in the presence of the insertion variant (**B**). ORF organisation is shown at the **Top** with reading frames shaded differently according to the reference sequence. AUG and stop codons are represented as green and black vertical bars, respectively. RDGs are shown at the bottom. Given the low probability of leaky scanning through the first two AUGs (see text), it is expected that the translon T3 corresponding to CDS cannot be translated in the pathogenic variant sequence.

Examination of ribosome profiling data in Trips-Viz (62) (Figure 5A) for *NXT1* mRNA reveals translation of an upstream region in the -1 frame (blue translon) relative to the CDS (red translon). Similarly, examination of Trips-Viz data indicates translation upstream and in the +1 reading frame (red translon) relative to the annotated *NRAS* CDS (blue translon). For simplicity, the CDS starts are not depicted as a branching point and are considered to be 100% efficient translation initiation sites. As the translated regions in both graphs are overlapping, it is clear that the simultaneous translation of both translons by the same ribosome cannot occur, at least in the absence of 3' to 5' scanning of post-terminating ribosomes (63).

In addition to representing qualitative information, RDGs also enable a quantitative representation of translation regulation. Due to the leaky scanning mechanism of translation initiation, the efficiencies of CDS translation in these two examples directly depend on the efficiencies of the upstream starts, e.g., if all ribosomes initiated at the upstream starts, no CDS translation would be observed. The relative translation efficiencies of translons can be used to calculate the probabilities of initiation at the upstream starts (64). These probabilities may vary between different conditions or across different cell types due to a variety of mechanisms, such as global changes in the stringency of start recognition (65,66) or specific

regulation of mRNA via ribosome sensing of particular metabolites through interaction with the nascent peptide (39–43). Using RDG representations in this way makes it easier to characterize the relationship between translation events that are regulated (*via* changes in probabilities at branching points) and the relative rates of translons product synthesis.
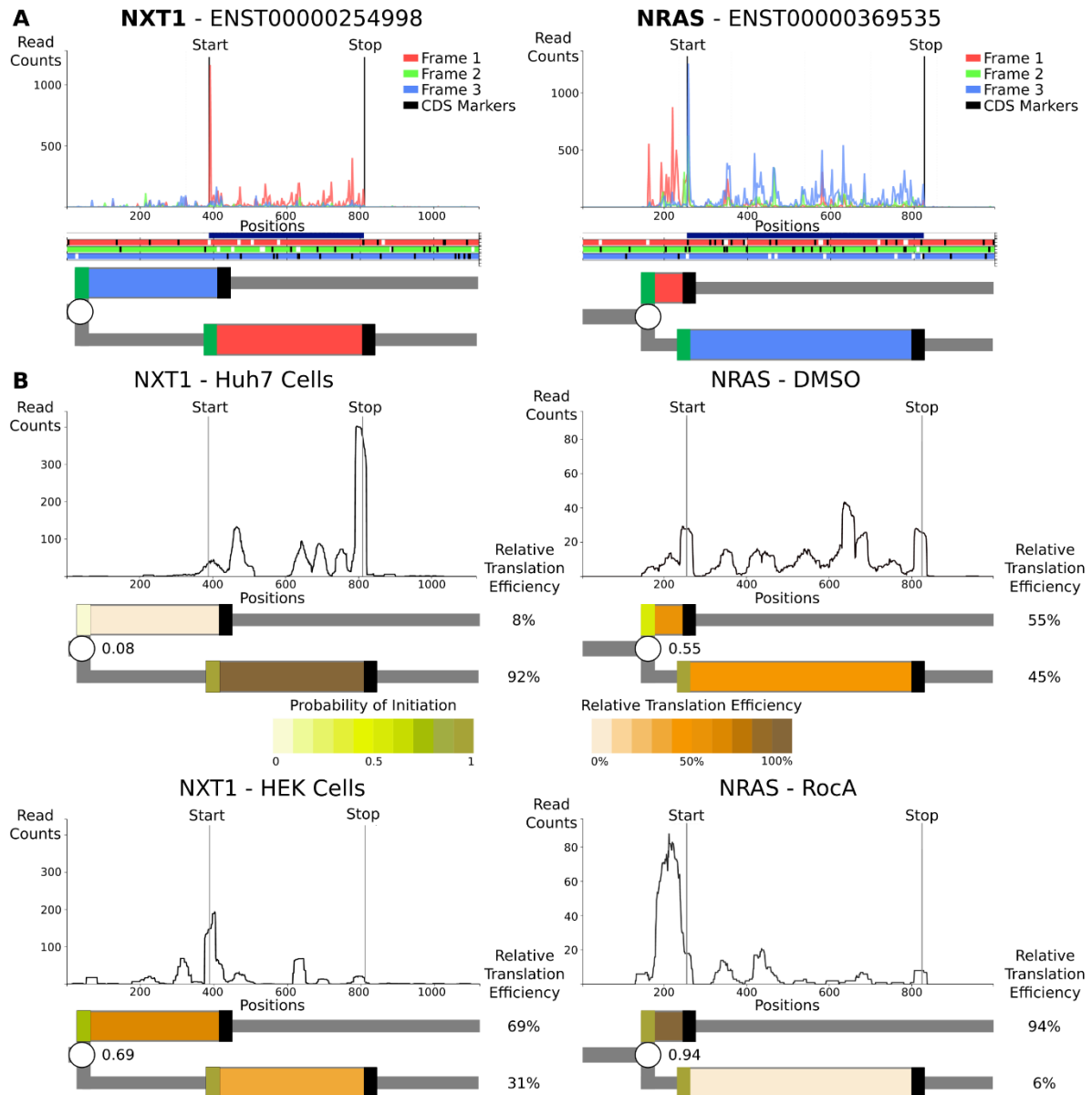

Figure 5: Representation of translation of human mRNAs. A. The top plots are A-site subcodon footprint densities for *NXT1* and *NRAS* mRNAs coloured to match the best-supported reading frame below in the ORF plots, in which AUGs are in white and stops are in black. Further below are RDG representations with translons coloured to match the translated frame. The CDS starts are treated deterministically and unproductive RiboPaths are not shown. **B.** Densities of ribosome A-site footprints obtained from ribosome profiling under different conditions (*NRAS*) or from different cells (*NXT1*). RDGs are shown below each density plot with translons coloured as a heatmap reflecting relative translation efficiencies. Branching points (starts) are also coloured as heatmaps reflecting the inferred probability of their initiation. The probabilities of translation initiation are shown as fractions decimals, while the relative translation efficiencies of each path are shown as percentages (%).

12

To illustrate this with real examples, we examined the translation of the above genes using different ribosome profiling datasets (Figure 5B). For *NRAS,* we used the dataset from cells treated with Rocaglamide A (RocA) and its untreated control (67). As can be seen in Figure 5B the silhouette of ribosome footprint density for the *NRAS* mRNA changes dramatically upon RocA treatment. These footprint densities can be used to calculate the relative translation efficiencies of *NRAS* translons and to derive the probability of translation initiation at the upstream starts (Methods). By showing the relative synthesis rates and initiation probabilities as heatmaps, the relationship between these two translons becomes apparent. RocA treatment greatly increases translation initiation probability at the upstream start, most likely *via* the ability of RocA to clamp eIF4A to mRNAs containing specific sequence motifs (67), which then reduces the downstream CDS translon. In the case of *NXT1,* we examined data obtained in two different cell lines, HeLa (68) and Huh7 (69). The silhouettes of ribosome footprint densities for *NXT1* mRNA are markedly different, as can be seen in Figure 5B. The RDG visualization of these differences in ribosome footprint densities pinpoints the upstream start as the pivotal element of cell-specific regulation of *NXT1* translation. For HeLa samples, the translation initiation at the upstream start is highly efficient, making the upstream start predominant. In contrast, for Huh7 samples efficiency at this start is much lower and, consequently, the CDS translon is predominant. The reasons for these cell-specific differences are beyond the scope of this work, but several mechanisms may be responsible, including different levels of translation factors that recognize translation initiation starts (70).

Future prospects

One of the attractive features of the RDG concept is its expandability. In the RDG examples above, we limited branching points only to starts where initiation and re-initiation events can occur. The most basic information for generating RDGs requires only locations of starts in a transcript because in-frame stop codons are identifiable from the sequence and are treated deterministically as the ends of translons. Therefore, the example shown in Figure 1B can be represented as a short array:

starts(x_1;x_2;x_3;x_4)

while the example in Figure 3 as

starts(x_1;x_2;x_3)

However, the concept can be extended to incorporate annotations for any non-deterministic translation events. For example, to capture such phenomena as stop codon readthrough or selenocysteine insertion (Figures 1C and 6A), an addition of a new type of branching point would be necessary: a stop codon, at which the ribosome could either terminate or incorporate an amino acid and continue translation.

starts(x_1); secs(y_1)

A frameshifting site would be yet another type of branching point where the ribosome could either continue translation in the same frame or shift to one of the alternative reading

frames. For example, the case of ribosomal frameshifting (Figures 1D and 6B) can be represented as

starts(x_1); shifts(i_1,s_1)

where shifts need to be assigned two coordinates, the coordinate of the last codon in the initial frame i_1 and the coordinate of the first codon in the shifted frame s. Accordingly, $s_1 - i_1$ = 1 would correspond to +1 frameshifting, $s_1 - y_1$ = -1 to -1 frameshifting, and so on. This notation is compatible not only with common types of frameshifting but also with extreme events such as bypassing in bacteriophage T4 gene *60* (34), where 50 nucleotides are skipped by the ribosome without translation, i.e. expressed as **shifts(i,s)** where $s - i$ = 50.

Algorithms interpreting such annotations will be backward and forward-compatible since older algorithms could simply ignore new types of branching points.
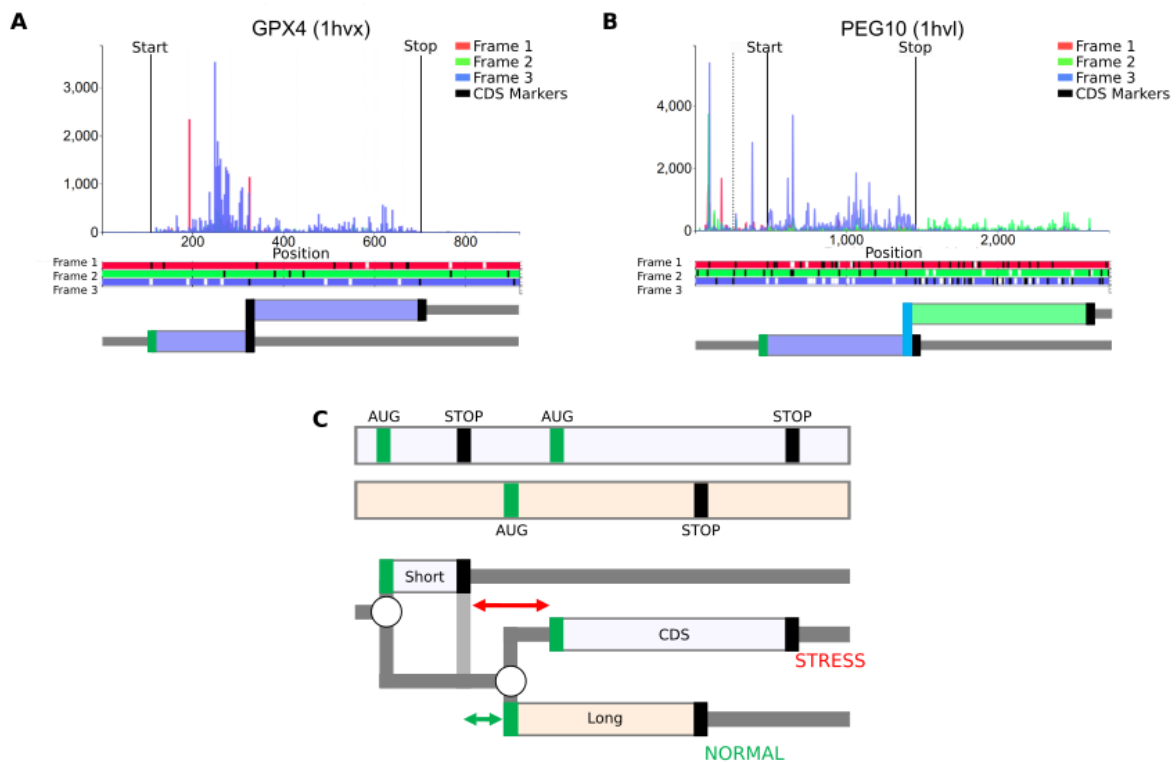

Figure 6: RDG representations of special cases. (**A**) mRNA encoding selenoprotein GPX4 and (**B**) PEG10 mRNA requiring ribosomal frameshifting for its expression. Trips-Viz screenshots of ribosomal profiling density for these mRNAs are at the top with ORF plots beneath and RDG representations further below. **C. Top**: An ORF plot containing three translons (protein-coding CDS and two uORFs, short and long), which is a minimal requirement for the mechanism of delayed re-initiation. **Below**: The corresponding RDG with green and red representing the predominant ribosome paths for normal and stress conditions, respectively. Arrows indicate the distance sufficient for scanning ribosomes to bind ternary complex under normal (green) or stress (red) conditions.

14

Despite the apparent simplicity of RDGs notations, it would be naive to expect that it can represent a full range of translational mechanisms. For example, in the case of a delayed re-initiation mechanism (50–52) that makes translation of certain mRNAs resistant to global downregulation during Integrated Stress Response (ISR), it is not sufficient to simply add a stop codon as a branching point, allowing either ribosome dissociation from mRNA or re-initiation downstream. This is because the reduced availability of the ternary complex (tRNAi*eIF2*GTP) increases the time required for the post-terminating ribosomes to bind the ternary complex, thereby enabling re-initiation (Figure 2B). Thus, it is not the probability of re-initiation, but the location of the start at which re-initiation will occur that changes during ISR. However, even in this case, the RDG concept can be useful to illustrate the mechanism, as shown in Figure 6C. It is also conceivable to extend the concepts of RDGs with parameters linking scanning distance to re-initiation probability.

An important shortcoming of the presented solution is the difficulty of its application to genomic loci encoding multiple transcript isoforms. The purpose of RDGs is to represent molecular events that take place during the translation of a single mRNA molecule. Therefore, a single RDG can only be applied to a single mRNA sequence. However, the concept of representing biological sequences as graphs is gaining momentum with splice graphs for representing alternative splicing (71) and variation graphs for representing pangenomes (72). Therefore, we envision that the RDG concept will fit into the emerging bioinformatics infrastructure of hierarchical representation of biological sequences as graphs, from genome to transcriptome to translatome.

Methods

We used GENCODE v42 (73) as a source of transcriptome annotation when searching for suitable loci and GENCODE v25 when producing visualizations in Trips-Viz (62). Pre-processed alignments from datasets GSE70211, GSE79664, and GSE94454 obtained in relevant studies (67–69) were obtained from Trips-Viz. The translation efficiencies of examined regions were calculated as the number of ribosome footprints uniquely aligning to individual translons normalised over the length of translons used for the mapping. Five first codons of translons were excluded from mapping to avoid the distortions introduced by the high peaks at the starts. The probability of translation initiation was then calculated as $p=R_u/(R_u+R_d)$, where R_u and R_d are translation efficiencies of the upstream and downstream translons.

Acknowledgements

Conflict of Interests

P.V.B. is a co-founder and a shareholder of Eirna Bio.

References

1. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009 Apr 10;324(5924):218–23.

2. Ingolia NT, Lareau LF, Weissman JS. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell. 2011 Nov 11;147(4):789–802.

3. Andreev DE, O'Connor PBF, Fahey C, Kenny EM, Terenin IM, Dmitriev SE, et al. Translation of 5' leaders is pervasive in genes resistant to eIF2 repression. Elife. 2015 Jan 26;4:e03971.

4. Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM. Long non-coding RNAs as a source of new peptides. Elife. 2014 Sep 16;3:e03523.

5. Calviello L, Mukherjee N, Wyler E, Zauber H, Hirsekorn A, Selbach M, et al. Detecting actively translated open reading frames in ribosome profiling data. Nat Methods. 2016 Feb;13(2):165–70.

6. Chong C, Müller M, Pak H, Harnett D, Huber F, Grun D, et al. Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. Nat Commun. 2020 Mar 10;11(1):1293.

7. Ji Z, Song R, Regev A, Struhl K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. Elife. 2015 Dec 19;4:e08890.

8. Johnstone TG, Bazzini AA, Giraldez AJ. Upstream ORFs are prevalent translational repressors in vertebrates. EMBO J. 2016 Apr 1;35(7):706–23.

9. Michel AM, Choudhury KR, Firth AE, Ingolia NT, Atkins JF, Baranov PV. Observation of dually decoded regions of the human genome using ribosome profiling data. Genome Res. 2012 Nov;22(11):2219–29.

10. Chothani SP, Adami E, Widjaja AA, Langley SR, Viswanathan S, Pua CJ, et al. A high-resolution map of human RNA translation. Molecular Cell. 2022 Aug;82(15):2885-2899.e8.

11.   Wright BW, Molloy MP, Jaschke PR. Overlapping genes in natural and engineered genomes. Nat Rev Genet. 2022 Mar;23(3):154–68.

12.   Fedorova AD, Kiniry SJ, Andreev DE, Mudge JM, Baranov PV. Thousands of human non-AUG extended proteoforms lack evidence of evolutionary selection among mammals. Nat Commun. 2022 Dec 23;13(1):7910.

13.   Dunn JG, Foo CK, Belletier NG, Gavis ER, Weissman JS. Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. Elife. 2013 Dec 3;2:e01179.

14.   Lobanov AV, Heaphy SM, Turanov AA, Gerashchenko MV, Pucciarelli S, Devaraj RR, et al. Position-dependent termination and widespread obligatory frameshifting in Euplotes translation. Nat Struct Mol Biol. 2017 Jan;24(1):61–8.

15.   Ivanov IP, Firth AE, Michel AM, Atkins JF, Baranov PV. Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences. Nucleic Acids Research. 2011 May;39(10):4220–34.

16.   Wethmar K, Smink JJ, Leutz A. Upstream open reading frames: Molecular switches in (patho)physiology. Bioessays. 2010 Oct;32(10):885–93.

17.   Namy O, Rousset JP, Napthine S, Brierley I. Reprogrammed Genetic Decoding in Cellular Gene Expression. Molecular Cell. 2004 Jan;13(2):157–68.

18.   Baranov PV, Gesteland RF, Atkins JF. Recoding: translational bifurcations in gene expression. Gene. 2002 Mar;286(2):187–201.

19.   Ivanov IP, Atkins JF, Michael AJ. A profusion of upstream open reading frame mechanisms in polyamine-responsive translational regulation. Nucleic Acids Research. 2010 Jan;38(2):353–9.

20.   Sieber P, Platzer M, Schuster S. The Definition of Open Reading Frame Revisited. Trends Genet. 2018 Mar;34(3):167–70.

21.   Kozak M. Pushing the limits of the scanning mechanism for initiation of translation. Gene. 2002 Oct;299(1–2):1–34.

22.   Jackson RJ, Hellen CUT, Pestova TV. The mechanism of eukaryotic translation initiation and principles of its regulation. Nat Rev Mol Cell Biol. 2010 Feb;11(2):113–27.

23.   Hinnebusch AG. The scanning mechanism of eukaryotic translation initiation. Annu Rev Biochem. 2014;83:779–812.

24.   Sonenberg N, Hinnebusch AG. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. Cell. 2009 Feb 20;136(4):731–45.

25.   Kozak M. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Res. 1987 Oct 26;15(20):8125–48.

26.   Kozak M. Context effects and inefficient initiation at non-AUG codons in eucaryotic cell-free translation systems. Mol Cell Biol. 1989 Nov;9(11):5073–80.

27.   Tzani I, Ivanov IP, Andreev DE, Dmitriev RI, Dean KA, Baranov PV, et al. Systematic analysis of the PTEN 5' leader identifies a major AUU initiated proteoform. Open Biol. 2016 May;6(5):150203.

28.   Cleary JD, Ranum LPW. Repeat associated non-ATG (RAN) translation: new starts in microsatellite expansion disorders. Curr Opin Genet Dev. 2014 Jun;26:6–15.

29.   Nguyen L, Cleary JD, Ranum LPW. Repeat-Associated Non-ATG Translation: Molecular Mechanisms and Contribution to Neurological Disease. Annu Rev Neurosci. 2019 Jul 8;42:227–47.

30.   Loughran G, Chou MY, Ivanov IP, Jungreis I, Kellis M, Kiran AM, et al. Evidence of efficient stop codon readthrough in four mammalian genes. Nucleic Acids Res. 2014 Aug 18;42(14):8928–38.

31.   Driscoll DM, Copeland PR. Mechanism and regulation of selenoprotein synthesis. Annu Rev Nutr. 2003;23:17–40.

32.   Labunskyy VM, Hatfield DL, Gladyshev VN. Selenoproteins: molecular pathways and physiological roles. Physiol Rev. 2014 Jul;94(3):739–77.

33.   Atkins JF, Loughran G, Bhatt PR, Firth AE, Baranov PV. Ribosomal frameshifting and transcriptional slippage: From genetic steganography and cryptography to adventitious use. Nucleic Acids Res. 2016 Sep 6;44(15):7007–78.

34.   Herr AJ, Atkins JF, Gesteland RF. Coupling of open reading frames by translational bypassing. Annu Rev Biochem. 2000;69:343–72.

35.   Nosek J, Tomaska L, Burger G, Lang BF. Programmed translational bypassing elements in mitochondria: structure, mobility, and evolutionary origin. Trends Genet. 2015 Apr;31(4):187–94.

36.   Klimova M, Senyushkina T, Samatova E, Peng BZ, Pearson M, Peske F, et al. EF-G–induced ribosome sliding along the noncoding mRNA. Sci Adv. 2019 Jun 7;5(6):eaaw9049.

37.   Atkins JF, Wills NM, Loughran G, Wu CY, Parsawar K, Ryan MD, et al. A case for "StopGo": reprogramming translation to augment codon meaning of GGN by promoting unconventional termination (Stop) after addition of glycine and then allowing continued translation (Go). RNA. 2007 Jun;13(6):803–10.

38.   Mudge JM, Ruiz-Orera J, Prensner JR, Brunet MA, Calvet F, Jungreis I, et al. Standardized annotation of translated open reading frames. Nat Biotechnol. 2022 Jul;40(7):994–9.

39.   Law GL, Raney A, Heusner C, Morris DR. Polyamine regulation of ribosome pausing at the upstream open reading frame of S-adenosylmethionine decarboxylase. J Biol Chem. 2001 Oct 12;276(41):38036–43.

40.   Hardy S, Kostantin E, Wang SJ, Hristova T, Galicia-Vázquez G, Baranov PV, et al. Magnesium-sensitive upstream ORF controls PRL phosphatase expression to mediate energy metabolism. Proc Natl Acad Sci U S A. 2019 Feb 19;116(8):2925–34.

41.   Rahmani F, Hummel M, Schuurmans J, Wiese-Klinkenberg A, Smeekens S, Hanson J. Sucrose control of translation mediated by an upstream open reading frame-encoded peptide. Plant Physiol. 2009 Jul;150(3):1356–67.

42.   Laing WA, Martínez-Sánchez M, Wright MA, Bulley SM, Brewster D, Dare AP, et al. An upstream open reading frame is essential for feedback regulation of ascorbate biosynthesis in Arabidopsis. Plant Cell. 2015 Mar;27(3):772–86.

43.   Ivanov IP, Shin BS, Loughran G, Tzani I, Young-Baird SK, Cao C, et al. Polyamine Control of Translation Elongation Regulates Start Site Selection on Antizyme Inhibitor mRNA via Ribosome Queuing. Mol Cell. 2018 Apr 19;70(2):254-264.e6.

44.   Dever TE, Ivanov IP, Hinnebusch AG. Translational regulation by uORFs and start codon selection stringency. Genes Dev. 2023 Jun 1;37(11–12):474–89.

45.   Ruan H, Shantz LM, Pegg AE, Morris DR. The upstream open reading frame of the mRNA encoding S-adenosylmethionine decarboxylase is a polyamine-responsive translational control element. J Biol Chem. 1996 Nov 22;271(47):29576–82.

46.   Andreev DE, Loughran G, Fedorova AD, Mikhaylova MS, Shatsky IN, Baranov PV. Non-AUG translation initiation in mammals. Genome Biol. 2022 May 9;23(1):111.

47.   Hinnebusch AG, Ivanov IP, Sonenberg N. Translational control by 5′-untranslated regions of eukaryotic mRNAs. Science. 2016 Jun 17;352(6292):1413–6.

48.   Kearse MG, Wilusz JE. Non-AUG translation: a new start for protein synthesis in eukaryotes. Genes Dev. 2017 Sep 1;31(17):1717–31.

49.   Pestova TV, Kolupaeva VG, Lomakin IB, Pilipenko EV, Shatsky IN, Agol VI, et al. Molecular mechanisms of translation initiation in eukaryotes. Proc Natl Acad Sci U S A. 2001 Jun 19;98(13):7029–36.

50.   Baird TD, Wek RC. Eukaryotic Initiation Factor 2 Phosphorylation and Translational Control in Metabolism12. Adv Nutr. 2012 May 4;3(3):307–21.

51.   Andreev DE, Arnold M, Kiniry SJ, Loughran G, Michel AM, Rachinskii D, et al. TASEP modelling provides a parsimonious explanation for the ability of a single uORF to derepress translation during the integrated stress response. eLife. 7:e32563.

52.   Hinnebusch AG. Translational regulation of yeast GCN4. A window on factors that control initiator-trna binding to the ribosome. J Biol Chem. 1997 Aug 29;272(35):21661–4.

53.   Costa-Mattioli M, Walter P. The integrated stress response: From mechanism to disease. Science. 2020 Apr 24;368(6489):eaat5314.

54.   Pakos-Zebrucka K, Koryga I, Mnich K, Ljujic M, Samali A, Gorman AM. The integrated stress response. EMBO Reports. 2016 Oct;17(10):1374–95.

55.   Prensner JR, Abelin JG, Kok LW, Clauser KR, Mudge JM, Ruiz-Orera J, et al. What Can Ribo-Seq, Immunopeptidomics, and Proteomics Tell Us About the Noncanonical Proteome? Molecular & Cellular Proteomics. 2023 Sep;22(9):100631.

56.   Wacholder A, Parikh SB, Coelho NC, Acar O, Houghton C, Chou L, et al. A vast evolutionarily transient translatome contributes to phenotype and fitness. Cell Systems. 2023 May;14(5):363-381.e8.

57.   Tanaka M, Sotta N, Yamazumi Y, Yamashita Y, Miwa K, Murota K, et al. The Minimum Open Reading Frame, AUG-Stop, Induces Boron-Dependent Ribosome Stalling and mRNA Degradation. Plant Cell. 2016 Nov;28(11):2830–49.

58.   Goel SC. Transcription Unit. Nature. 1973 Oct;245(5425):397–397.

59.   Baranov PV, Atkins JF, Yordanova MM. Augmented genetic decoding: global, local and temporal alterations of decoding processes and codon meaning. Nat Rev Genet. 2015 Sep;16(9):517–29.

60.   Rodnina MV, Korniy N, Klimova M, Karki P, Peng BZ, Senyushkina T, et al. Translational recoding: canonical translation mechanisms reinterpreted. Nucleic Acids Research. 2020 Feb 20;48(3):1056–67.

61.   Whiffin N, Karczewski KJ, Zhang X, Chothani S, Smith MJ, Evans DG, et al. Characterising the loss-of-function impact of 5' untranslated region variants in 15,708 individuals. Nat Commun. 2020 May 27;11(1):2523.

62.   Kiniry SJ, Judge CE, Michel AM, Baranov PV. Trips-Viz: an environment for the analysis of public and user-generated ribosome profiling data. Nucleic Acids Res. 2021 Jul 2;49(W1):W662–70.

63.   Gould PS, Dyer NP, Croft W, Ott S, Easton AJ. Cellular mRNAs access second ORFs using a novel amino acid sequence-dependent coupled translation termination-reinitiation mechanism. RNA. 2014 Mar;20(3):373–81.

64.   Michel AM, Andreev DE, Baranov PV. Computational approach for calculating the probability of eukaryotic translation initiation from ribo-seq data that takes into account leaky scanning. BMC Bioinformatics. 2014 Nov 21;15(1):380.

65.   Loughran G, Sachs MS, Atkins JF, Ivanov IP. Stringency of start codon selection modulates autoregulation of translation initiation factor eIF5. Nucleic Acids Res. 2012 Apr;40(7):2898–906.

66.   Fijalkowska D, Verbruggen S, Ndah E, Jonckheere V, Menschaert G, Van Damme P. eIF1 modulates the recognition of suboptimal translation initiation sites and steers gene expression via uORFs. Nucleic Acids Res. 2017 Jul 27;45(13):7997–8013.

67.   Iwasaki S, Floor SN, Ingolia NT. Rocaglates convert DEAD-box protein eIF4A into a sequence-selective translational repressor. Nature. 2016 Jun 23;534(7608):558–61.

68.   Park JE, Yi H, Kim Y, Chang H, Kim VN. Regulation of Poly(A) Tail and Translation during the Somatic Cell Cycle. Mol Cell. 2016 May 5;62(3):462–71.

69.   Lintner NG, McClure KF, Petersen D, Londregan AT, Piotrowski DW, Wei L, et al. Selective stalling of human translation through small-molecule engagement of the ribosome nascent chain. PLoS Biol. 2017 Mar;15(3):e2001882.

70.   Anisimova AS, Kolyupanova NM, Makarova NE, Egorov AA, Kulakovskiy IV, Dmitriev SE. Human Tissues Exhibit Diverse Composition of Translation Machinery. IJMS. 2023 May 6;24(9):8361.

71.   Ryan MC, Cleland J, Kim R, Wong WC, Weinstein JN. SpliceSeq: a resource for analysis and visualization of RNA-Seq data on alternative splicing and its functional impacts. Bioinformatics. 2012 Sep 15;28(18):2385–7.

72.   Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, et al. A draft human pangenome reference. Nature. 2023 May 11;617(7960):312–24.

73.   Frankish A, Carbonell-Sala S, Diekhans M, Jungreis I, Loveland JE, Mudge JM, et al. GENCODE: reference annotation for the human and mouse genomes in 2023. Nucleic Acids Res. 2023 Jan 6;51(D1):D942–9.