Small proteins encoded by unannotated ORFs are rising stars of the proteome, confirming shortcomings in genome annotations and current vision of an mRNA

Vivian Delcourt[1,2,3], Antanas Staskevicius[1], Michel Salzet[2], Isabelle Fournier[2], Xavier Roucou[1,3*]

[1]Department of Biochemistry, Université de Sherbrooke, Quebec, Canada; [2]Univ. Lille, INSERM U1192, Laboratoire Protéomique, Réponse Inflammatoire & Spectrométrie de Masse (PRISM) F-59000 Lille, France; [3]PROTEO, Quebec Network for Research on Protein Function, Structure, and Engineering, Quebec, Canada

*Correspondance to Xavier Roucou: Department of Biochemistry (Z8-2001), Faculté de Médecine et des Sciences de la Santé, Université de Sherbrooke, 3201 Jean Mignault, Sherbrooke, Quebec J1E 4K8, Canada, Tel. (819) 821-8000x72240; Fax. (819) 820 6831; E-Mail: xavier.roucou@usherbrooke.ca

Abbreviations: CDS, coding sequence; CyPrP, cytosolic prion protein; UTR, untranslated region; ORF, open reading frame

Received: 05 01, 2017; Revised: 05 01, 2017; Accepted: 06 02, 2017

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/pmic.201700058 .

Abstract

Short ORF-encoded peptides and small proteins in eukaryotes have been hiding in the shadow of large proteins for a long time. Recently, improved identifications in MS-based proteomics and ribosome profiling resulted in the detection of large numbers of small proteins. The variety of functions of small proteins is also emerging. It seems to be the right time to reflect on why small proteins remained invisible. In addition to the obvious technical challenge of detecting small proteins, they were mostly forgotten from annotations and they escaped detection because they were not sought. In this review, we identify conventions that need to be revisited, including the assumption that mature mRNAs carry only one coding sequence. The large-scale discovery of small proteins and of their functions will require changing some paradigms and undertaking the annotation of ORFs that are still largely perceived as irrelevant coding information compared to already annotated coding sequences.

Introduction

There is definitively some buzz around the largely unexplored territory of short ORFs, generally defined as ORFs below 100 codons, and their translation products in prokaryotes and in eukaryotes [1–13]. Here, we will focus our discussion on eukaryotes, mainly mammals. Based on the weight of the size criterion in the way annotations are performed, annotated coding ORFs or protein-coding sequences (CDSs) are virtually always the longest ORFs within a coding gene or transcript [14,15]. Thus, unannotated ORFs within coding genes and transcripts are obligatory shorter than CDSs, and although most of these ORFs are shorter than the conventional 100 codons cut-off used to qualify as short ORFs, a fraction of them are longer [16–20]. We previously used the terms alternative ORFs or altORFs for all

unannotated ORFs because they require alternative initiation compared to canonical translation initiation sites mapped to CDSs, a large fraction overlap annotated CDSs in an alternative reading frame, and they may generate alternative translation products different from annotated protein [16]. In this review, we extend our discussion to unannotated ORFs and their translation products in general.

Unannotated ORFs are generally classified in different groups (figure 1). These ORFs, particularly short ORFs, are disturbing. They violate some basic concepts in genome and transcriptome annotations, including the conventional 100 codons cut-off used to identify non-coding ORFs [14,15,21,22] and the one mature mRNA - one protein paradigm [23]. They have been hiding and remained undetected until large scale mapping of ribosome occupancy using ribosome profiling exposed unexpected large numbers of translated ORFs [24–31]. Many small proteins translated from ORFs have now been detected by MS-based proteomics [29,32–34]. Although the physiological function of the majority of these novel small proteins is not known, some have important functions in muscle contraction [35–37], development [38,39] and signaling [40,41]. Finally, the field is recent and it is difficult to predict the implications of the discovery of this new genetic information that has been hiding in genomes.

Here, we mainly discuss some elements of the modern view of protein translation that are definitively not compatible with the recent discoveries of small proteins, and we briefly present some strategies that are used to improve the exploration of this new territory of the proteomic world.

Basic concepts and corollaries of the modern view of the protein coding information

Annotations of protein-coding genes have greatly contributed to the development and the application of omics-based technologies in modern experimental biology and medicine. Each annotated gene, CDS, transcript and protein has a specific entry in databases that are routinely used for research, and these databases are indispensable in the emerging approach of precision medicine. For example, protein databases are central to the success of MS-based protein identification [42].Unfortunately, databases have a dark side; they limit the scope of discoveries that can be made since any unannotated information cannot be detected in a biological sample. Here, we highlight some general concepts in annotations, some of which are clearly hampering the discovery of small proteins.

Coding and non-coding genes and transcripts

Several strategies are used for the annotation of coding sequences (CDSs), including CDS identification based on known protein sequences and ab initio predictions of the most likely CDS [43,44]. And typically, genes are annotated as coding or non-coding. For example, the Ensembl human gene annotation (Genome Reference Consortium Human Build 38) contains 20,310 coding genes and 37,118 non-coding genes (including 14,589 pseudogenes).

In eukaryotes, alternative splicing, alternative transcription and alternative polyadenylation are mechanisms that produce different RNA isoforms with possible variations in the CDS [45]. Therefore, there are more CDSs than protein-coding genes, and most genes encode multiples protein isoforms. CDSs are annotated in databases such as RefSeq [46], Ensembl [47] and the Consensus CDS [48]. As of April 2017, the consensus CDS database contains 18,889 coding genes and 32,524 CDSs for *Homo sapiens*.

If a transcript generated from a protein-coding gene does not contain the annotated CDS or a variation of this CDS, or if it contains a retained intron, it is labeled as a transcript without ORF or non-coding. Thus, according to current annotations, a coding gene can generate both coding and non-coding RNAs. An example is illustrated in figure 2 for the centromere protein A, CENPA. Transcripts 1 and 2 code for the canonical 140 amino acid CENPA protein and a shorter 114 amino acid isoform, respectively. Annotations indicate that the third transcript does not contain any protein-coding ORF. Thus, CENPA is annotated as a coding-gene expressing coding and non-coding transcripts. One could imagine that transcripts 1 and 2 are not expressed in a specific tissue; in that tissue, this gene would functionally be a non-coding gene. One could also imagine that in the same tissue, *CENPA* could be protein-coding in conditions where transcripts 1 and 2 are expressed and non-coding in conditions where transcript 3 only is expressed. This example illustrates how annotations simplify the information and may not reflect the biology.

The presence of unannotated ORFs adds another level of complexity. An overlapping ORF^{CDS} in the +2 reading frame is not annotated (figure 2). Yet, ribosome profiling data detect initiating ribosomes with a number of reads similar to reads on the annotated CDS [49]. Thus, transcript 3 may be a coding transcript after all. Examples of transcripts for which annotation changed from non-coding to coding are shown in table 1.

One mRNA, one protein

Protein synthesis is one of the most complex and energetically demanding cellular processes [50]. The eukaryotic ribosome is a molecular machine with four RNA molecules and at least 80 proteins. Ribosome biogenesis and assembly demands major cellular resources, including 200 assembly protein factors and 80 small nucleolar RNAs [51,52]. Every stage of translation, initiation, elongation, termination and recycling is also regulated by a large

number of specific factors. Overall, if ribosome biosynthesis is not taken into account, more than 100 protein-coding genes and 60 non-protein-coding genes are involved in translation.

The aim of this elaborate translation machinery is to interact with mRNAs and decode CDSs. In the current view of protein synthesis, mRNAs harbour a tripartite structure with a unique CDS and two untranslated regions (5'- and 3'-UTRs). Thus, although the life of an mRNA is complex [53], its function would be to carry a single coding message, the annotated CDS.

The 100 codons cut-off

ORFs are expected to occur by chance in a long nucleotide sequence, and establishing a 100 codon cut-off has been one of the key criteria for the annotation of CDSs, and for the separation of mRNAs from non-coding RNAs [21,22]. Although additional criteria are used to identify likely CDSs, including conservation and known protein domains, the 100 codons cut-off is the main bottleneck that has prevented short ORFs from being included in annotation databases.

Implications

These concepts have an important corollary: the complete set of the coding information is restricted to annotated CDSs, and the proteome is primarily determined according to this assumption. Thus, short ORFs are excluded from the transcriptome coding potential, and the predicted corresponding proteins are omitted from the reference proteome. Consequently, short ORFs and small proteins are not well represented in current databases. The size distribution of the current set of human consensus CDSs indicates that median and average lengths are 434 and 569 codons, respectively (figure 3), confirming previous results obtained with the analysis of a collection of cDNAs [54], GENCODE annotated CDSs [20], and a protein database [7]. The fraction of CDSs with a maximum length of 100 codons is only

2.5%. Interestingly, this observation was used as evidence supporting the 100 codons cut-off to discriminate coding from non-coding RNAs [21,22]. Our discussion below suggests that it is rather because the contribution of small proteins has been largely overlooked that they are few in databases.

Since the fraction of small proteins in databases is so low, there is a generally unrecognized perception that small proteins have less important functions than large proteins. This issue will not be addressed here but it is interesting to note small proteins of 100 amino acids or less are structural subunits or key regulators of two vital molecular machines, the FO-F1 ATP synthase and the sarcoplasmic reticulum Ca^{2+}-ATPase, respectively (table 2).

Translation outside of annotated CDSs: delinquent translation machinery or outdated concepts?

Do ribosomes interact with transcripts annotated as mRNAs but not those annotated as non-coding? Do ribosomes decode annotated CDSs only when they interact with mRNAs? Do only annotated CDSs code for functional proteins? These questions, which might have seemed senseless a few years ago, are now fundamental questions and the answers are no. A large number of studies clearly show that cells can decode both annotated CDSs and unannotated ORFs, and some databases now include information about ORFs with evidence of expression [29,49,55–57]. Two databases are specific for short ORFs and proteins shorter than 100 amino acids in eukaryotic species [56,57].

The role of serendipity in the discovery of overlapping ORFs

Since ORFs are not annotated and the modern view of an mRNA is still a transcript carrying a single annotated CDS, the discovery of small proteins translated from unconventional ORFs is often associated with serendipity [1,2,5,23]. In my laboratory, we stumbled upon a first small protein following a combination of circumstances using a green fluorescent protein (GFP) tagged cytosolic form of the prion protein (CyPrP), CyPrP^{GFP} [58,59]. In this construct, GFP is inserted into a natural restriction site within the unstructured N-terminal domain of CyPrP (figure 4A). Cells transfected with CyPrP^{GFP} display RNA aggregates in the cytoplasm [60], and the expression of several RNA helicases, including DDX3 was tested by western blot using a commercial anti-DDX3 antibody, Ab37160 (figure 4B). DDX3 was detected at the expected size in mock-transfected cells, but a second band labeled X appeared in cells expressing CyPrP^{GFP}. The hypothesis of a novel DDX3 isoform was unlikely based on the observation that three other antibodies directed against different regions of DDX3, Ab50703, NB-200-194 and NB200-195 did not detect protein X (figure 4B). This experiment was later repeated with CyPrP^{GFP} C-terminal deletion mutants (figure 4C). Strikingly, the electrophoretic mobility of protein X increased with the deletions. Eventually, we discovered that PrP CDS contains an ORF^{CDS} in the +3 reading frame [18] (figure 4A, C). In our experiments, the GFP CDS was inserted inside the ORF^{CDS}. We realized that there are no stop codons in the +3 reading frame of GFP CDS, and that a chimeric small protein containing frameshifted GFP might be expressed. In a control experiment, transfected cells expressing frameshifted GFP displayed an epitope that was detected with antibody Ab37160 (figure 4D). Thus, protein X was a chimeric protein containing the Ab37160 epitope translated from an ORF^{CDS} in the prion protein CDS. If a commercial anti- DDX3 antibody different from Ab37160 had been used in the first place, the expression of this novel small

protein would have gone unnoticed, and researchers expressing the prion protein in their experiments would not have known that both a large and a small protein are co-expressed in their experiments. These observations are proof of principle that ribosomes are able to translate two overlapping messages from the same transcript. As discussed below, many studies have confirmed this feature.

Overlapping viral ORFs decoded by the human translation machinery

Some of the earliest evidence that mammalian ribosomes may translate ORFs^{CDS} comes from viruses which use unusual strategies to optimize their coding capacity in small genomes. For space constraints, we provide below a few examples for human viruses only. The hepatitis C virus (HCV) combines the polyprotein and the overlapping strategies. The 10 canonical proteins of the hepatitis C virus are coded in a long CDS that is translated into a large polyprotein. Further proteolytic cleavage results in the production of 7 non-structural and 3 structural proteins. In addition to these proteins, additional smaller proteins encoded by CDSs overlapping the main CDS in a different reading frame are also produced. Such alternative reading frame proteins can be detected in patients developing specific antibodies [61]. One of these proteins, the alternate F protein seems to modulate the immune system [62–64]. West Nile Virus also generates alternative reading frame proteins [65]. Many other RNA viruses express multiple proteins translated from ORFs^{CDS} within their mRNAs [66]. Oncogenic viruses such as Kaposi's sarcoma-associated herpesvirus and Epstein-Barr virus also produce alternative reading frame proteins encoded in alternative CDSs overlapping the latency-associated nuclear antigen and Epstein-Barr nuclear antigen 1 CDSs, respectively [67]. Whether these novel proteins have a function in the oncogenic activity of these viruses remains to be determined. Importantly, these observations indicate that small ORFs^{CDS} hiding in large annotated viral CDSs are translated in infected human cells. Ribosome profiling

confirmed the translation of a large number of viral annotated CDSs and short ORFs^{CDS} in human cells infected with cytomegalovirus and Kaposi's sarcoma-associated herpesvirus [68]. In this study, a fraction of the small proteins were also confirmed by MS-based proteomics.

Overlapping ORFs (ORFs^{CDS}) in mammals

There is accumulating evidence of functional ORFs^{CDS} in mammalian mRNAs [16,18,19,69–72], and computational approaches predict a large number of ORFs^{CDS} in human and murine cDNAs and transcripts [73–75]. Several cDNAs containing the annotated CDS and an ORFs^{CDS} ligated into expression vectors generate a small protein in addition to the larger annotated proteins in cultured cells [16,18,19,69–71] and in vivo [76,77]. These remarkable observations confirm a counterintuitive feature of mammalian ribosomes which are able to decode ORFs^{CDS} in addition to annotated CDSs in the same transcripts. The translation of both the CDS and the ORFs^{CDSs} can be easily tested. Generally, an epitope tag fused in-frame with a CDS facilitates the detection of a protein of interest by western blotting or immunofluorescence (figure 5). However, this approach does not allow the detection of a small protein co-expressed from an ORF^{CDS}. Addition of a second epitope in-frame with the ORF^{CDS} makes the small protein detectable (figure 5). In other words, what you see is what you've tagged; you won't see what you haven't tagged.

ORFs in regions annotated as untranslated: mRNA UTRs, long non-coding RNAs (lncRNAs), and pseudogenes RNAs

A combination of computational and experimental approaches provide strong evidence for the translation of ORFs present in regions classified as untranslated, from yeast to human [16,26,28,78–85]. These unannotated ORFs include ORFs$^{5'}$ upstream of the CDS, also

termed upstream ORFs [86], ORFs$^{3'}$ downstream of the CDS, and ORFs^{nc} in long non-coding RNA and pseudogenes transcripts (figure 1). A number of studies demonstrate function in development [38,39,87], DNA repair [88], muscle contraction [35–37], cell signaling [40,41,89], mRNA decapping [90] and phagocytosis [91]. As a consequence of these discoveries, an update is performed in databases and the annotation of some genes and their transcripts switches from non-coding to coding (table 1). Once in the databases, the small proteins are officially in the reference proteome and can be detected in routine MS-based proteomics experiments.

Mechanisms for the translation of ORFs

Similar to mRNAs, many lncRNAs are transcribed by RNA polymerase II and are modified with a 5'cap structure and a 3' poly A tail [92]. From the ribosome perspective, there is no specific mechanism requirement for the translation of ORFs^{nc}. Possible mechanisms for the translation of several ORFs within the same transcript have already been reviewed [23] and will not be detailed here. Briefly, these mechanisms can be separated into two classes. In the first class, leaky scanning and translation reinitiation are compatible with the widely accepted ribosome scanning model for translation initiation [93,94]. The second class includes the recruitment of ribosomes on translation initiation sites by tethering, cap-assisted internal initiation and RNA looping [95–97].

The regulation of the translation of ORFs is still unknown but, some insights into the translation of ORFs$^{5'}$ have recently been published. In contrast to annotated CDSs, translation of ORFs$^{5'}$ is resistant to oxidative stress [78], and glucose and oxygen deprivation [98]. Unconventional translation of ORFs$^{5'}$ mediated by the alternative initiation factor eiF2A seems to be crucial in tumor initiation [99].

Taking on the challenges of unifying the nomenclature, annotating, detecting and deciphering the function of currently unannotated ORFs and proteins

Challenge 1: the nomenclature

The heterogeneity in the nomenclature of short ORFs and their translation products in the literature is a good indication that this rapidly expanding research domain is relatively new. ORFs below 100 codons are termed short ORFs (sORFs) [100], small ORFs (smORFs) [5,28,57,80,82,84,101,102], or upstream ORFs (uORFs) when they are located in 5'UTRs. The cutoff may reach 150 codons in some studies [100]. Their translation products are labeled peptides [8,36,38,39,83,87], small proteins [1,2,9,103,104], sORF-encoded peptides or SEP [34,88,100], micropeptides [105], and microproteins [90]. We have termed unannotated ORFs and proteins, alternative ORFs and alternative proteins, respectively to highlight the fact that they represent different ORFs and translation products compared to current annotations [16,18,19,23].

In the short term, it will be useful to introduce a unified nomenclature to avoid confusion and to better organize the field of small proteins.

Challenge 2: the detection

There is no straightforward protocol to detect the short proteome, and it is not as simple as "just mass spec your protein gel dye fronts to death" [11]. Recent computational and biochemical advances, and reduction in the size cutoff for potential coding-ORFs are helping the discovery of small proteins. The different approaches include ribosome profiling [20,25,26,28,30,49,55,57,72,81,105,106] and proteogenomics [16,29,32–34,107–109].

Proteogenomics involve the extraction of ORFs from genomic or transcriptomic data and the generation of custom-made protein databases for MS-based proteomics identification [110].

Challenge 3: the function

There have been recent successes in the discovery of the function of small proteins [35–41,88–91]. Yet, undertaking the functional analyses of small proteins is a multi-level challenge. First, the selection of an unannotated ORF for functional investigation is tricky. ORFs are more likely to occur by chance in the genome [102,111]. Computational tools to find homologs were developed with large proteins but may not perform as well on short ORFs [112]. Protein domains such as those annotated by the InterPro database were determined with annotated and mostly large proteins [113], and the majority of small proteins are unlikely to contain these conventional domains. Second, it is technically challenging to work with small proteins. For example, the use of tags with small proteins is always a concern as it can modify some biochemical features and interfere with normal localization [114]. Third, knockdown experiments to validate function are not possible with standard small interfering or short hairpin RNAs (siRNAs or shRNAs respectively) approaches when the small protein under investigation is encoded in an already annotated mRNA. siRNAs or shRNAs would reduce the expression of both the small and the annotated large proteins at the same time. Here, gene editing methods are required to remove an initiation site or introduce a stop codon to specifically stop the expression of the short ORF. Gene editing may be particularly tricky for ORFs^{CDS} because it would be preferable not to modify the sequence of the annotated protein.

Conclusion and perspective

Only long CDSs and large proteins made it through into current annotation databases routinely used to detect the coding genome and the proteome. Yet, there is strong evidence from ribosome profiling and proteogenomics that there are within our cells new genetic information and novel proteins the vast majority of which remains invisible, similar to a genomic and a proteomic dark matter. The exploration of this dark matter requires questioning the assumptions on which the annotations were based, including (1) that mRNAs carry a single CDS because only one CDS is annotated; and (2) that RNAs annotated as non-coding are necessarily non-coding. Just as annotations are not rigid and change according to new experimental evidence, assumptions should not be interpreted as absolute dogmas either.

The extent of the proteomic dark matter made of small proteins is difficult to assess. Flooding current databases used for MS-based proteomics with all possible ORFs is not an option; the majority of short ORFs may represent expected biological noise [7], and inserting great numbers of irrelevant ORFs would result in large databases that cause challenges for peptide identification [110]. The trend in recent years has been to create customized databases based on a proteogenomic approach. Parallel to the identification of small proteins with this approach, it will be important to facilitate access to and update a database containing the sequences of all detected small proteins. This will enable the MS-based proteomics community to validate and detect the expression of small proteins in routine experiments, an important advance towards the democratization and functional characterization of the small proteome.

Acknowledgements

This research was supported by CIHR grants MOP-137056 and MOP-136962 to X.R, and a Canada Research Chair in Functional Proteomics and Discovery of New Proteins to X.R. X.R is member of the Fonds de Recherche du Québec Santé-supported Centre de Recherche du Centre Hospitalier Universitaire de Sherbrooke. We thank the staff from the Centre for Computational Science at the Université de Sherbrooke, Compute Canada and Compute Québec for access to the Mammouth supercomputer.

References

[1] Ramamurthi, K.S., Storz, G., The small protein floodgates are opening; now the functional analysis begins. *BMC Biol.* 2014, 12, 96.

[2] Storz, G., Wolf, Y.I., Ramamurthi, K.S., Small Proteins Can No Longer Be Ignored. *Annu. Rev. Biochem.* 2014, 83, 753–777.

[3] Su, M., Ling, Y., Yu, J., Wu, J., Xiao, J., Small proteins: untapped area of potential biological importance. *Front. Genet.* 2013, 4, 286.

[4] Chu, Q., Ma, J., Saghatelian, A., Identification and characterization of sORF-encoded polypeptides. *Crit. Rev. Biochem. Mol. Biol.* 2015, 50, 134–141.

[5] Saghatelian, A., Couso, J.P., Discovery and characterization of smORF-encoded bioactive polypeptides. *Nat. Chem. Biol.* 2015, 11, 909–16.

[6] Andrews, S.J., Rothnagel, J.A., Emerging evidence for functional peptides encoded by short open reading frames. *Nat. Rev. Genet.* 2014, 15, 193–204.

[7] Landry, C.R., Zhong, X., Nielly-Thibault, L., Roucou, X., Found in translation: Functions and evolution of a recently discovered alternative proteome. *Curr. Opin. Struct. Biol.* 2015, 32, 74–80.

[8] Pueyo, J.I., Magny, E.G., Couso, J.P., New Peptides Under the s(ORF)ace of the Genome. *Trends Biochem. Sci.* 2016, 41, 665–678.

[9] Yang, X., Tschaplinski, T.J., Hurst, G.B., Jawdy, S., et al., Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. *Genome Res.* 2011, 21, 634–41.

[10] Staudt, A.-C., Wenkel, S., Regulation of protein function by "microProteins". *EMBO Rep.* 2011, 12, 35–42.

[11] Feller, S., Microproteins (miPs) – the next big thing. *Cell Commun. Signal.* 2012, 10, 42.

[12] Ericson, M., Janes, M.A., Butter, F., Mann, M., et al., On the extent and role of the small proteome in the parasitic eukaryote Trypanosoma brucei. *BMC Biol.* 2014, 12, 14.

[13] Hellens, R.P., Brown, C.M., Chisnall, M.A.W., Waterhouse, P.M., Macknight, R.C., The Emerging World of Small ORFs. *Trends Plant Sci.* 2016, 21, 317–328.

[14] Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., et al., Life with 6000 genes. *Science* 1996, 274, 546, 563–7.

[15] Carninci, P., Kasukawa, T., Katayama, S., Gough, J., et al., The Transcriptional Landscape of the Mammalian Genome. *Science (80-.).* 2005, 309, 1559–1563.

[16] Vanderperre, B., Lucier, J.-F.F., Bissonnette, C., Motard, J., et al., Direct detection of alternative open reading frames translation products in human significantly expands the proteome. *PLoS One* 2013, 8, e70698.

[17] Chalick, M., Jacobi, O., Pichinuk, E., Garbar, C., et al., MUC1-ARF-A Novel MUC1 Protein That Resides in the Nucleus and Is Expressed by Alternate Reading Frame Translation of MUC1 mRNA. *PLoS One* 2016, 11, e0165031.

[18] Vanderperre, B., Staskevicius, A.B., Tremblay, G., McCoy, M., et al., An overlapping reading frame in the PRNP gene encodes a novel polypeptide distinct from the prion protein. *FASEB J.* 2011, 25, 2373–86.

[19] Bergeron, D., Lapointe, C., Bissonnette, C., Tremblay, G., et al., An out-of-frame overlapping reading frame in the ataxin-1 coding sequence encodes a novel ataxin-1 interacting protein. *J. Biol. Chem.* 2013, 288, 21824–35.

[20] Raj, A., Wang, S.H., Shim, H., Harpak, A., et al., Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *Elife* 2016, 5, 1–24.

[21] Dinger, M.E., Pang, K.C., Mercer, T.R., Mattick, J.S., Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput. Biol.* 2008, 4, e1000176.

[22] Ulitsky, I., Bartel, D.P., lincRNAs: Genomics, Evolution, and Mechanisms. *Cell* 2013, 154, 26–46.

[23] Mouilleron, H., Delcourt, V., Roucou, X., Death of a dogma: eukaryotic mRNAs can code for more than one protein. *Nucleic Acids Res.* 2015, 44, 14–23.

[24] Ingolia, N.T., Lareau, L.F., Weissman, J.S., Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 2011, 147, 789–802.

[25] Lee, S.S., Liu, B., Lee, S.S., Huang, S.-X.S.-X., et al., Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U. S. A.* 2012, 109, E2424-2432.

[26] Ingolia, N.T., Brar, G.A., Stern-Ginossar, N., Harris, M.S., et al., Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. *Cell Rep.* 2014, 8, 1365–1379.

[27] Fields, A.P., Rodriguez, E.H., Jovanovic, M., Stern-Ginossar, N., et al., A Regression-

Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. *Mol. Cell* 2015, 60, 816–827.

[28] Bazzini, A.A., Johnstone, T.G., Christiano, R., Mackowiak, S.D., et al., Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.* 2014, 33, 981–993.

[29] Crappé, J., Ndah, E., Koch, A., Steyaert, S., et al., PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Res.* 2015, 43, e29.

[30] Calviello, L., Mukherjee, N., Wyler, E., Zauber, H., et al., Detecting actively translated open reading frames in ribosome profiling data. *Nat. Methods* 2015.

[31] Brar, G.A., Weissman, J.S., Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat. Rev. Mol. Cell Biol.* 2015, 16, 651–64.

[32] Menschaert, G., Van Criekinge, W., Notelaers, T., Koch, A., et al., Deep proteome coverage based on ribosome profiling aids MS-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Mol. Cell. Proteomics* 2013.

[33] Koch, A., Gawron, D., Steyaert, S., Ndah, E., et al., A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites. *Proteomics* 2014, 14, 2688–98.

[34] Ma, J., Ward, C.C., Jungreis, I., Slavoff, S.A., et al., Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. *J. Proteome Res.* 2014, 13, 1757–

65.

[35] Anderson, D.M., Anderson, K.M., Chang, C.-L., Makarewich, C.A., et al., A Micropeptide Encoded by a Putative Long Noncoding RNA Regulates Muscle Performance. *Cell* 2015, 160, 595–606.

[36] Magny, E.G., Pueyo, J.I., Pearl, F.M.G., Cespedes, M. a., et al., Conserved Regulation of Cardiac Calcium Uptake by Peptides Encoded in Small Open Reading Frames. *Science (80-.).* 2013, 341, 1116–1120.

[37] Nelson, B.R., Makarewich, C.A., Anderson, D.M., Winders, B.R., et al., A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science (80-.).* 2016, 351, 271–275.

[38] Kondo, T., Plaza, S., Zanet, J., Benrabah, E., et al., Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis. *Science* 2010, 329, 336–339.

[39] Kondo, T., Hashimoto, Y., Kato, K., Inagaki, S., et al., Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nat. Cell Biol.* 2007, 9, 660–665.

[40] Pauli, A., Norris, M.L., Valen, E., Chew, G.-L., et al., Toddler: An Embryonic Signal That Promotes Cell Movement via Apelin Receptors. *Science* 2014, 343, 1248636–1248636.

[41] Chng, S.C., Ho, L., Tian, J., Reversade, B., ELABELA: A Hormone Essential for Heart Development Signals via the Apelin Receptor. *Dev. Cell* 2013, 27, 672–680.

[42] Aebersold, R., Mann, M., Mass spectrometry-based proteomics. *Nature* 2003, 422,

198–207.

[43] Mudge, J.M., Harrow, J., The state of play in higher eukaryote gene annotation. *Nat. Rev. Genet.* 2016, 17, 758–772.

[44] Yandell, M., Ence, D., A beginner's guide to eukaryotic genome annotation. *Nat. Rev.* 2012, 13, 329–342.

[45] de Klerk, E., 't Hoen, P.A.C., Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. *Trends Genet.* 2015, 31, 128–139.

[46] O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., et al., Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 2016, 44, D733–D745.

[47] Cunningham, F., Amode, M.R., Barrell, D., Beal, K., et al., Ensembl 2015. *Nucleic Acids Res.* 2014, 43, D662-9.

[48] Farrell, C.M., O'Leary, N.A., Harte, R. a., Loveland, J.E., et al., Current status and new features of the Consensus Coding Sequence database. *Nucleic Acids Res.* 2014, 42, 865–872.

[49] Michel, A.M., Fox, G., M Kiran, A., De Bo, C., et al., GWIPS-viz: development of a ribo-seq genome browser. *Nucleic Acids Res.* 2014, 42, D859-864.

[50] Rolfe, D.F., Brown, G.C., Cellular energy utilization and molecular origin of standard metabolic rate in mammals. *Physiol. Rev.* 1997, 77, 731–58.

[51] Woolford, J.L., Baserga, S.J., Ribosome biogenesis in the yeast Saccharomyces cerevisiae. *Genetics* 2013, 195, 643–81.

[52] Thomson, E., Ferreira-Cerca, S., Hurt, E., Eukaryotic ribosome biogenesis at a glance. *J. Cell Sci.* 2013, 126, 4815–21.

[53] Moore, M.J., From birth to death: the complex lives of eukaryotic mRNAs. *Science* 2005, 309, 1514–8.

[54] Frith, M.C., Forrest, A.R., Nourbakhsh, E., Pang, K.C., et al., The abundance of short proteins in the mammalian proteome. *PLoS Genet.* 2006, 2, 515–528.

[55] Wan, J., Qian, S.B., TISdb: A database for alternative translation initiation in mammalian cells. *Nucleic Acids Res.* 2014, 42, 845–850.

[56] Hao, Y., Zhang, L., Niu, Y., Cai, T., et al., SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. *Brief. Bioinform.* 2017, 420, 563–73.

[57] Olexiouk, V., Crapp??, J., Verbruggen, S., Verhegen, K., et al., SORFs.org: A repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res.* 2016, 44, D324–D329.

[58] Grenier, C., Bissonnette, C., Volkov, L., Roucou, X., Molecular morphology and toxicity of cytoplasmic prion protein aggregates in neuronal and non-neuronal cells. *J. Neurochem.* 2006, 97, 1456–1466.

[59] Beaudoin, S., Vanderperre, B., Grenier, C., Tremblay, I., et al., A large ribonucleoprotein particle induced by cytoplasmic PrP shares striking similarities with the chromatoid body, an RNA granule predicted to function in posttranscriptional gene regulation. *Biochim. Biophys. Acta* 2009, 1793, 335–345.

[60] Goggin, K., Beaudoin, S., Grenier, C., Brown, A.A., Roucou, X., Prion protein

aggresomes are poly(A)+ ribonucleoprotein complexes that induce a PKR-mediated deficient cell stress response. *Biochim. Biophys. Acta* 2008, 1783, 479–491.

[61] Morice, Y., Ratinier, M., Miladi, A., Chevaliez, S., et al., Seroconversion to hepatitis C virus alternate reading frame protein during acute infection. *Hepatology* 2009, 49, 1449–59.

[62] Park, S.B., Seronello, S., Mayer, W., Ojcius, D.M., Hepatitis C Virus Frameshift/Alternate Reading Frame Protein Suppresses Interferon Responses Mediated by Pattern Recognition Receptor Retinoic-Acid-Inducible Gene-I. *PLoS One* 2016, 11, e0158419.

[63] Samrat, S.K., Li, W., Singh, S., Kumar, R., Agrawal, B., Alternate reading frame protein (F protein) of hepatitis C virus: paradoxical effects of activation and apoptosis on human dendritic cells lead to stimulation of T cells. *PLoS One* 2014, 9, e86567.

[64] Xu, X., Yu, X., Deng, X., Yue, M., et al., Hepatitis C virus alternate reading frame protein decreases interferon-α secretion in peripheral blood mononuclear cells. *Mol. Med. Rep.* 2014, 9, 730–6.

[65] Faggioni, G., Pomponi, A., De Santis, R., Masuelli, L., et al., West Nile alternative open reading frame (N-NS4B/WARF4) is produced in infected West Nile Virus (WNV) cells and induces humoral response in WNV infected individuals. *Virol. J.* 2012, 9, 283.

[66] Firth, A.E., Brierley, I., Non-canonical translation in RNA viruses. *J. Gen. Virol.* 2012, 93, 1385–409.

[67] Kwun, H.J., Toptan, T., Ramos da Silva, S., Atkins, J.F., et al., Human DNA tumor

viruses generate alternative reading frame proteins through repeat sequence recoding. *Proc. Natl. Acad. Sci. U. S. A.* 2014, 111, E4342-9.

[68] Stern-Ginossar, N., Ingolia, N.T., Ribosome Profiling as a Tool to Decipher Viral Complexity. *Annu. Rev. Virol.* 2015, 2, 335–49.

[69] Abramowitz, J., Grenet, D., Birnbaumer, M., Torres, H.N., Birnbaumer, L., XLalphas, the extra-long form of the alpha-subunit of the Gs G protein, is significantly longer than suspected, and so is its companion Alex. *Proc. Natl. Acad. Sci. U. S. A.* 2004, 101, 8366–8371.

[70] Klemke, M., Kehlenbach, R.H., Huttner, W.B., Two overlapping reading frames in a single exon encode interacting proteins - A novel way of gene usage. *EMBO J.* 2001, 20, 3849–3860.

[71] Lee, C. -f. C., Lai, H.-L.H.-L., Lee, Y.-C., Chien, C.-L.C.-L., Chern, Y., The A2A Adenosine Receptor Is a Dual Coding Gene: A NOVEL MECHANISM OF GENE USAGE AND SIGNAL TRANSDUCTION. *J. Biol. Chem.* 2014, 289, 1257–1270.

[72] Michel, A.M., Choudhury, K.R., Firth, A.E., Ingolia, N.T., et al., Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Res.* 2012, 22, 2219–2229.

[73] Chung, W.-Y., Wadhawan, S., Szklarczyk, R., Pond, S.K., Nekrutenko, A., A first look at ARFome: dual-coding genes in mammalian genomes. *PLoS Comput. Biol.* 2007, 3, e91.

[74] Ribrioux, S., Brüngger, A., Baumgarten, B., Seuwen, K., John, M.R., Bioinformatics prediction of overlapping frameshifted translation products in mammalian transcripts.

BMC Genomics 2008, 9, 122.

[75] Xu, H., Wang, P., Fu, Y., Zheng, Y., et al., Length of the ORF, position of the first AUG and the Kozak motif are important factors in potential dual-coding transcripts. *Cell Res.* 2010, 20, 445–57.

[76] Li, C., Goudy, K., Hirsch, M., Asokan, A., et al., Cellular immune response to cryptic epitopes during therapeutic gene transfer. *Proc. Natl. Acad. Sci. U. S. A.* 2009, 106, 10770–10774.

[77] Kracht, M.J.L., van Lummel, M., Nikolic, T., Joosten, A.M., et al., Autoimmunity against a defective ribosomal insulin gene product in type 1 diabetes. *Nat. Med.* 2017, 23, 501–507.

[78] Andreev, D.E., O'Connor, P.B., Fahey, C., Kenny, E.M., et al., Translation of 5' leaders is pervasive in genes resistant to eIF2 repression. *Elife* 2015, 4, e03971.

[79] Prabakaran, S., Hemberg, M., Chauhan, R., Winter, D., et al., Quantitative profiling of peptides from RNAs classified as noncoding. *Nat. Commun.* 2014, 5, 5429.

[80] Smith, J.E., Alvarez-Dominguez, J.R., Kline, N., Huynh, N.J., et al., Translation of Small Open Reading Frames within Unannotated RNA Transcripts in Saccharomyces cerevisiae. *Cell Rep.* 2014, 7, 1858–1866.

[81] 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, 4, e08890.

[82] Mackowiak, S.D., Zauber, H., Bielow, C., Thiel, D., et al., Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol.* 2015, 16, 179.

[83] Ruiz-Orera, J., Messeguer, X., Long non-coding RNAs as a source of new peptides. *arXiv Prepr. arXiv ...* 2014, 1–40.

[84] Aspden, J.L., Eyre-Walker, Y.C., Philips, R.J., Amin, U., et al., Extensive translation of small ORFs revealed by Poly-Ribo-Seq. *Elife* 2014, e03528.

[85] Popa, A., Lebrigand, K., Barbry, P., Waldmann, R., Pateamine A-sensitive ribosome profiling reveals the scope of translation in mouse embryonic stem cells. *BMC Genomics* 2016, 17, 52.

[86] Wethmar, K., Barbosa-Silva, A., Andrade-Navarro, M.A., Leutz, A., uORFdb--a comprehensive literature database on eukaryotic uORF biology. *Nucleic Acids Res.* 2014, 42, D60-7.

[87] Galindo, M.I., Pueyo, J.I., Fouix, S., Bishop, S.A., Couso, J.P., Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol.* 2007, 5, e106.

[88] Slavoff, S. a., Heo, J., Budnik, B. a., Hanakahi, L. a., Saghatelian, A., A human short open reading frame (sORF)-Encoded polypeptide that stimulates DNA end joining. *J. Biol. Chem.* 2014, 289, 10950–10957.

[89] Matsumoto, A., Pasut, A., Matsumoto, M., Yamashita, R., et al., mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature* 2017, 541, 228–232.

[90] D'Lima, N.G., Ma, J., Winkler, L., Chu, Q., et al., A human microprotein that interacts with the mRNA decapping complex. *Nat. Chem. Biol.* 2017, 13, 174–180.

[91] Pueyo, J.I., Magny, E.G., Sampson, C.J., Amin, U., et al., Hemotin, a Regulator of

Phagocytosis Encoded by a Small ORF and Conserved across Metazoans. *PLoS Biol.* 2016, 14, 1–34.

[92] Quinn, J.J., Chang, H.Y., Unique features of long non-coding RNA biogenesis and function. *Nat. Rev. Genet.* 2015, 17, 47–62.

[93] Kozak, M., Pushing the limits of the scanning mechanism for initiation of translation. *Gene* 2002, 299, 1–34.

[94] Schleich, S., Strassburger, K., Janiesch, P.C., Koledachkina, T., et al., DENR-MCT-1 promotes translation re-initiation downstream of uORFs to control tissue growth. *Nature* 2014, 512, 208–12.

[95] Chappell, S.A., Edelman, G.M., Mauro, V.P., Ribosomal tethering and clustering as mechanisms for translation initiation. *Proc. Natl. Acad. Sci. U. S. A.* 2006, 103, 18077–82.

[96] Martin, F., Barends, S., Jaeger, S., Schaeffer, L., et al., Cap-Assisted Internal Initiation of Translation of Histone H4. *Mol. Cell* 2011, 41, 197–209.

[97] Paek, K.Y., Hong, K.Y., Ryu, I., Park, S.M., et al., Translation initiation mediated by RNA looping. *Proc. Natl. Acad. Sci.* 2015, 112, 201416883.

[98] Andreev, D.E., O'Connor, P.B., Zhdanov, A. V, Dmitriev, R.I., et al., Oxygen and glucose deprivation induces widespread alterations in mRNA translation within 20 minutes. *Genome Biol.* 2015, 16, 90.

[99] Sendoel, A., Dunn, J.G., Rodriguez, E.H., Naik, S., et al., Translation from unconventional 5' start sites drives tumour initiation. *Nature* 2017, 541, 494–499.

[100] Slavoff, S. a, Mitchell, A.J., Schwaid, A.G., Cabili, M.N., et al., Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.* 2013, 9, 59–64.

[101] Zanet, J., Benrabah, E., Li, T., Pélissier-Monier, A., et al., Pri sORF peptides induce selective proteasome-mediated protein processing. *Science* 2015, 349, 1356–1358.

[102] Basrai, M. a., Hieter, P., Boeke, J.D., Small open reading frames: Beautiful needles in the haystack. *Genome Res.* 1997, 7, 768–771.

[103] Oyama, M., Itagaki, C., Hata, H., Suzuki, Y., et al., Analysis of small human proteins reveals the translation of upstream open reading frames of mRNAs. *Genome Res.* 2004, 14, 2048–52.

[104] Cabrera-Quio, L.E., Herberg, S., Pauli, A., Decoding sORF translation – from small proteins to gene regulation. *RNA Biol.* 2016, 13, 1051–1059.

[105] Crappé, J., Van Criekinge, W., Trooskens, G., Hayakawa, E., et al., Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. *BMC Genomics* 2013, 14, 648.

[106] Michel, A.M., Baranov, P. V, Ribosome profiling: a Hi-Def monitor for protein synthesis at the genome-wide scale. *Wiley Interdiscip. Rev. RNA* 2013, 4, 473–90.

[107] Menschaert, G., Van Criekinge, W., Notelaers, T., Koch, A., 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. *Mol. Cell. Proteomics* 2013, 12, 1780–90.

[108] Olexiouk, V., Menschaert, G., in:, *Adv. Exp. Med. Biol.*, vol. 926, 2016, pp. 49–64.

[109] Ma, J., Diedrich, J.K., Jungreis, I., Donaldson, C., et al., Improved Identification and Analysis of Small Open Reading Frame Encoded Polypeptides. *Anal. Chem.* 2016, 88, 3967–3975.

[110] Nesvizhskii, A.I., Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* 2014, 11, 1114–1125.

[111] FICKETT, J.W., ORFs and Genes: How Strong a Connection? *J. Comput. Biol.* 1995, 2, 117–123.

[112] Cheng, H., Chan, W.S., Li, Z., Wang, D., et al., Small open reading frames: current prediction techniques and future prospect. *Curr. Protein Pept. Sci.* 2011, 12, 503–7.

[113] Mitchell, A., Chang, H.-Y., Daugherty, L., Fraser, M., et al., The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.* 2014, 43, D213-221.

[114] Viallet, P.M., Vo-Dinh, T., Monitoring intracellular proteins using fluorescence techniques: from protein synthesis and localization to activity. *Curr. Protein Pept. Sci.* 2003, 4, 375–88.

Table 1: the discovery of small proteins leads to changes in annotations

Annotation[a]	Gene	Gene type	Ensembl Transcript	Transcript type	Protein (aa)
GRCh37	APELA/ELABELA (ENSG00000248329)	Non coding	ENST00000507152	lincRNA[b]	-
GRCH38		Coding		Coding	54
GRCh37	MRLN (ENSG00000227877)	Non-coding	ENST00000414264	lincRNA[b]	-
GRCH38		Coding		Coding	46
GRCh37	SLC35A4 (ENSG00000176087)	Coding	ENST00000323146[c]	Coding	324
GRCH38		Coding	ENST00000623481	Novel coding	103[d]

[a]Ensembl human genome assembly.

[b]lincRNA refers to long intergenic non-coding RNAs in Ensembl annotations

[c]Although this canonical transcript contains both the CDS and a short ORF encoding a novel 103 amino acids protein, and was proposed to be a bicistronic mRNA [78], it is annotated as coding the 324 aa SLC35A4 protein only.

[d]This small protein is not an isoform, and SLC35A4 is a dual-coding gene.

Table 2: Small proteins and molecular machines

Protein	Size (aa)	Description
ATP synthase subunit f	94	ATP synthase, H+ transporting, mitochondrial Fo complex subunit F2
ATP synthase subunit epsilon	51	ATP synthase, H+ transporting, mitochondrial F1 complex, epsilon subunit
ATP synthase subunit e	69	ATP synthase, H+ transporting, mitochondrial Fo complex subunit E
ATP synthase subunit g 2	100	ATP synthase, H+ transporting, mitochondrial Fo complex subunit G2
Sarcolipin	31	Sarcoplasmic reticulum Ca^{2+}-ATPase (SERCA), negative regulator
Phospholamban	52	Sarcoplasmic reticulum Ca^{2+}-ATPase (SERCA), negative regulator

Figure 1: Unannotated ORFs may be found in mRNAs or non-coding RNAs.

Molecules shown here represent mature or processed RNAs. Unannotated ORFs are found in untranslated regions of the transcriptome. Within mRNAs, these regions include 5'UTRs, 3'UTRs, and alternative reading frames overlapping annotated CDSs. An $ORF^{5'}$ (also termed upstream ORF or uORF) may be found outside the CDS in the three reading frames, or partially overlapping a CDS in one of the two alternative reading frames. ORF^{CDS} may be found completely nested inside the CDS, or partially overlapping the 3'UTR in one of the two alternative reading frames1. An $ORF^{3'}$ may be in one of the 3 reading frames. Non-coding RNAs do not have a tripartite structure similar to mRNAs, and an ORF^{nc} may be found anywhere.

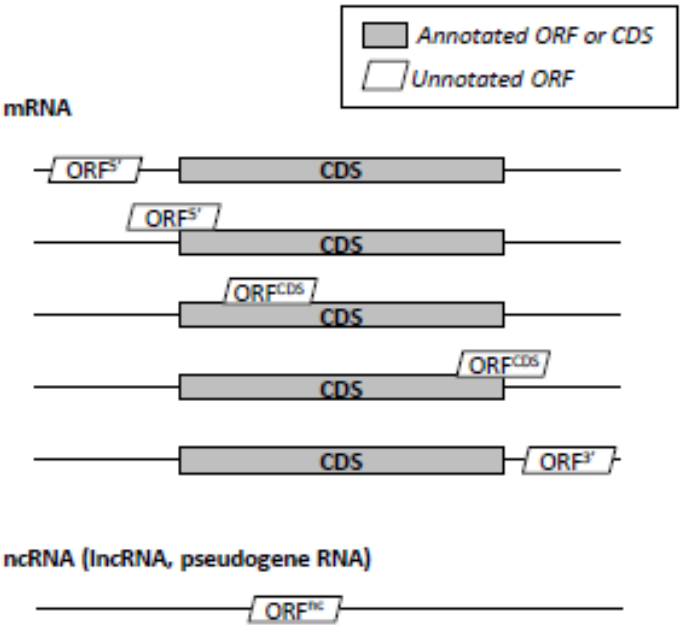

Figure 2. Typical Ensembl transcripts annotation for a human gene.

Three transcripts from the CENPA gene are illustrated. Sequences annotated as coding sequences are shown as gray boxes, and sequences annotated as untranslated regions are shown as open boxes. Shaded boxes represent AUG translation initiation sites. Black boxes represent stop codons. ENST00000335756 carries the canonical 420 bps CDS with 4 coding exons. The end of exon 4 after the stop codon is non-coding. ENST00000233505 is an isoform lacking exon 3 and carries a shorter version of the canonical CDS. ENST00000472719 is an isoform lacking exon 1. In the absence of the translation initiation site, the Ensembl annotation indicates the absence of a canonical CDS, and this transcript is believed to be non-coding. An unannotated overlapping long ORF of 53 codons in the +2 reading frame starts with an AUG codon within exon 4 and ends with a stop codon in exon 5. Translation initiation ribosome profiling data clearly show initiating ribosomes on both the annotated CDS and the unannotated ORF with similar number of reads (graph below). Thus, transcripts 1 and 2 might be bicistronic, and transcript 3 is likely a coding transcript for a novel small protein.

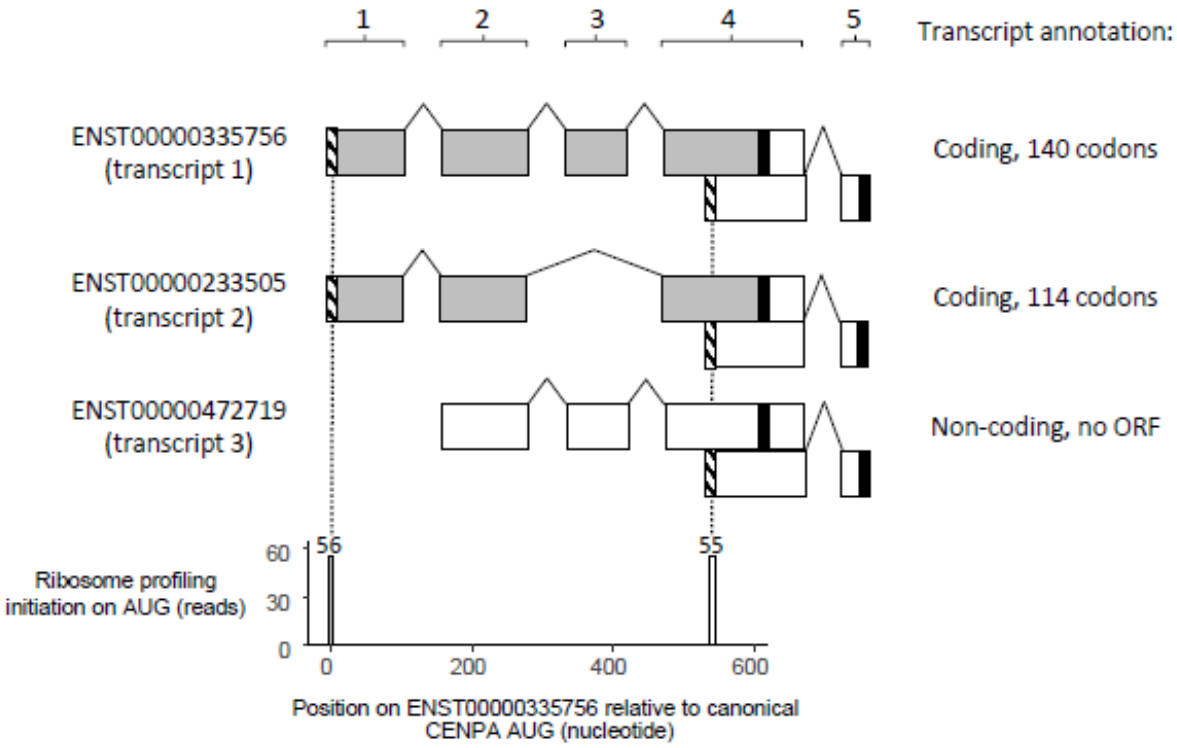

Figure 3: Distribution of the number and the size of human consensus CDSs.

Calculations were performed with consensus CDS release 20. Median, 434 codons; average, 570 codons.

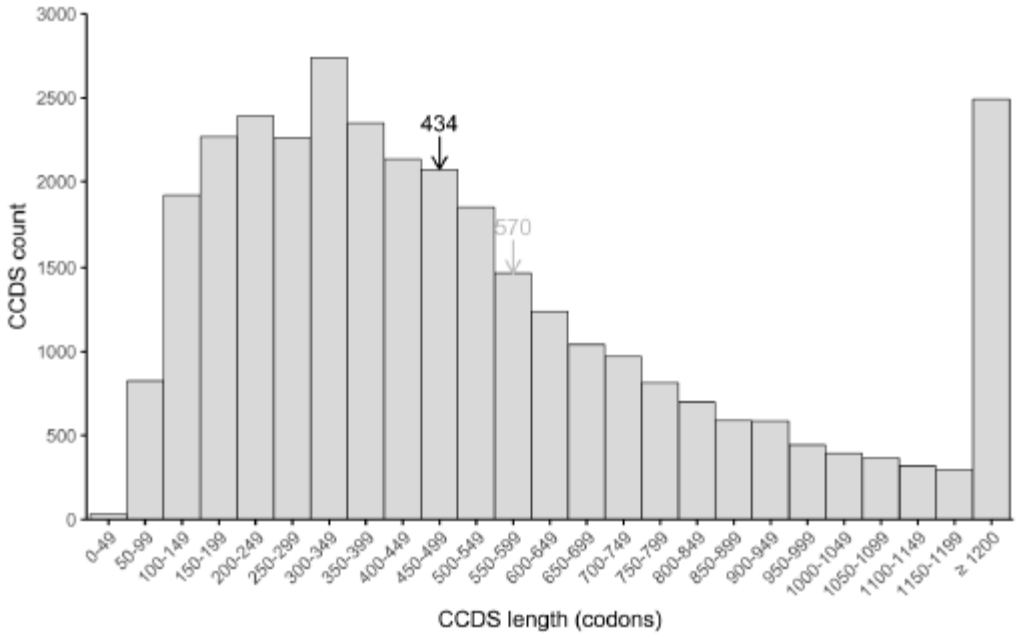

Figure 4: Translation of an overlapping ORF in the prion protein (PrP) CDS.

A. Diagram of CyPrP^{GFP} CDS (+1 reading frame) and frameshifted alternative CyPrP^{GFP} (+3 reading frame). GFP CDS is shown as an empty box, CyPrP CDS is shown in grey. **B.** Western blot analysis of mock-transfected HEK293 cells (lane 1) and CyPrP^{GFP}-transfected cells (lane 2) with four antibodies directed against DDX3, as indicated. DDX3 is indicated by an arrow above the 72 kDa marker. An unknown protein labeled X is detected in CyPrP^{GFP}-expressing cells with antibodies Ab37160. **C.** Western blot analysis of different CyPrP^{GFP} C-terminal deletion mutants with anti-PrP (left blot) or anti-DDX3 (right blot, Ab37160) antibodies. Lane 1: Mock-transfected cells; lane 2: cells transfected with wild-type CyPrP^{GFP}; lanes 3-5: deletion mutants. A diagram of each construct is indicated on the right side of each blot. C-terminal deletions within CyPrP^{GFP} introduce C-terminal deletions within overlapping alternative CyPrP^{GFP}. **D.** Western blot analysis of cells transfected with CyPrP^{GFP} (lane 1) or frameshifted GFP (lane 2) with anti-DDX3 antibodies (Ab37160).

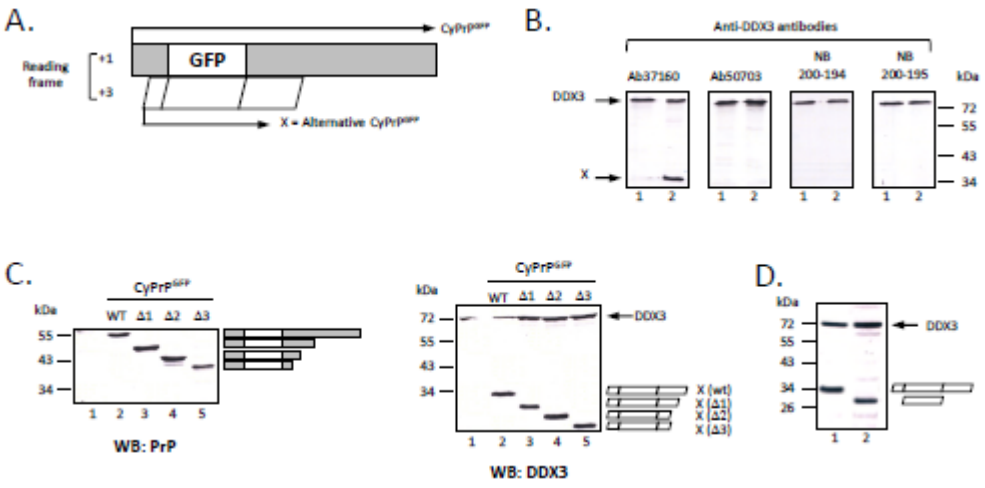

Figure 5. A simple double epitope tagging strategy for the detection of ORFs^{CDS}.

Upper panel: experimental design. (**A**) Empty vector. (**B**) Expression plasmid containing an epitope tagged CDS (tag1). (**C**) Expression plasmid containing both an epitope tagged CDS (tag 1) and an epitope tagged ORF^{CDS} (tag 2). Tag 2 is in-frame with the ORF^{CDS}, but out-of-frame with the CDS. Tag 2 should not introduce a stop codon in the CDS frame.

Bottom panel: after transfection and expression, cell lysates are analyzed by western blot with anti-tag 1 and anti-tag 2 antibodies. (**A**) No signals are detected in mock-transfected cells. (**B**) Anti-tag 1 antibodies detect the expression of the protein of interest. A second protein expressed from the ORF^{CDS} remain invisible. (**C**) Anti-tag 2 antibodies detect the expression of the unannotated small protein.