Shedding Light on the Ghost Proteome

Tristan Cardon, Isabelle Fournier, Michel Salzet

Shedding light on the Ghost proteome

Tristan Cardon[1*], Isabelle Fournier[1,2*] and Michel Salzet[1,2*]

[1]Univ. Lille, Inserm, CHU Lille, U1192 - Protéomique Réponse Inflammatoire Spectrométrie de Masse - PRISM, F-59000 Lille, France
[2]Institut Universitaire de France, Paris, France

Abstract

Conventionally, eukaryotic messenger ribonucleic acid (mRNAs) were thought to be monocistronic, leading to the translation of a single protein. However, large-scale proteomics has led to a massive identification of proteins translated from alternative open reading frame (AltORFs) in mRNAs. AltORFs are find in addition to predicted reference ORF (RefORF) or from non-coding RNA (ncRNAs). AltProts are not represented in the conventional protein databases and this "Ghost proteome" was not considered until recently. Some of these proteins are functional and there is growing evidence that they are involved in central functions in physiological and physiopathological context. Here, we review how this ghost proteome fills the gap in our understanding of signaling pathways, establishes new markers of pathologies and highlights therapeutic targets.

Keywords: Alternative Protein (AltProt), smProt, SEPs, small ORF (smORF), mass spectrometry, hidden proteome

* Corresponding authors: Prof. Michel Salzet (ORCID: 0000-0003-4318-0817), Prof. Isabelle Fournier (ORCID: 0000-0003-1096-5044), PhD Tristan Cardon (ORCID: 0000-0003-1751-0528) Laboratoire Protéomique, Réponse Inflammatoire et Spectrométrie de Masse (PRISM) - Inserm U1192 - Université de Lille, Bât SN3, 1er étage, Cité Scientifique, F-59655 Villeneuve d'Ascq Cedex, France. Phone: +33 (0)3 20 43 41 94; Fax: +33 (0)3 20 43 40 54; email: michel.salzet@univ-lille.fr, isabelle.fournier@univ-lille.fr, tristan.cardon@univ-lille.fr

The hidden side of the eukaryotes protein landscape

Proteins are key players of cell regulation and modern Mass Spectrometry (MS) offers the access to their large-scale identification; though large-scale proteomic strategies rely on the comparison of experimental MS data to theoretical prediction through protein database interrogation. Therefore, since the mid 80's, there has been a tremendous effort from the scientific community to provide reliable protein databases [1–3]. Most proteins included in databases are predicted from gene annotations according to well established rules. By convention, for eukaryotes, unlike prokaryotes, annotations only consider genes or transcripts with ORFs longer than 100 codons, shorter products are disregarded as being noncoding RNAs (ncRNAs). Moreover, only a single ORF per coding gene or coding transcript, referred as the reference ORF (RefORF), is considered, since eukaryotes are known to be **monocistronic (see Glossary)** [4,5]. However, it was demonstrated that a significant number of proteins take their origins from alternative ORF (AltORFs) and were missing from databases, making them invisible to the scientific community for a long time [6,7]. Most importantly, these unknown products from AltORF, have been shown to be functionally active. Because they are from different mechanisms and originate from differently classified RNAs, we will use the term Alternative Proteins (AltProts) regardless of their origin and amino acids length (**see Box 1**) [8]. This missing part of the proteome, called the Ghost proteome, is the hidden part of the iceberg. This review will be devoted to these Ghost proteins; reviewing their origin, the approaches used for their identification and to unveil their functions, and the impact they have on the biochemical landscape of the cell, including the pathophysiological functions.

Origin of the Ghost proteome

The monocistronic mRNA dogma

In eukaryotes, mRNAs carry the genetic information that is translated by ribosomes into protein. Briefly, the main steps of translation are i) the formation of the 43S preinitiation complex, ii) preparation of the mRNA and iii) RNA scanning and the association of the ribosome subunits (**Figure 1A**). These steps are subjected to regulations, like inhibition by protein binding RNA or activation by stimulation of the initiation, resulting in the modification of the protein expression. Regulation can also be affected by modifications of the translation actors (*e.g.* hydrolysis of eIF2, phosphorylation of eIF4E or eIF1, 3 or 5), of the mRNA due to the binding of other proteins at the 5' UTR region causing inaccessibility of the **43S ribosome**. But also, to the micro-RNA binding and blocking the accessibility of mRNA regions to the translation factors [9]. Despite knowledge of these regulation mechanisms, eukaryotic mRNA has always been considered as monocistronic, leading to the translation of a single protein named RefProt. An important amount of work aiming to predict protein sequences was performed by Dr. M. Kozak to understand which sequence is translated into a protein. This work has led Kozak to introduce a scanning model for translation in 1992 [4,5,10], where the ribosome is moving along the mRNA sequence from the 5' down to the 3' UTR. In this model, the first "AUG" start codon read is generally the one to initiate translation. Kozak was describing a strong context favoring the initiation of the translation [11], such as the sequence "gccPccAUGG" in which the capital letters are highly conserved, P being a purine (most frequently an adenosine). The strong and weak context described by Kozak explain the possible translation from other codons than the first "AUG" through alternative mechanisms of translation notably by **reinitiation or leaky scanning** [4]. This model has served as structure for the creation of the first protein database, with the UniProtKB/Swiss-Prot database following the same rules for translation. According to these rules, the coding DNA sequence (CDS) region is the longest nucleotide sequence flanked on

each side by a start and a stop codon. CDS defines the RefORF which is translated into the RefProt.

Existence of mechanisms leading to more than one protein from a single gene is an evolutionary sign that is already established in Bacteria, Archaea, and Viruses. In viruses, multiplicity of AUG initiation sites for translation in the RNA lead to express several proteins in a well-coordinated manner [12]. In eukaryotes, genes which possess multiple initiation site have already been described [13]. They are mostly proteins involved in regulation, such as translation factors, proto-oncogenes, hormone receptors, growth factors, inflammatory mediators, and proteins involved in signal transmission [14–16]. Despite these mechanisms are now considered by prediction algorithms for building databases (e.g. **TrEMBL**) [17] and identify new proteins including their mechanisms; a large part of the proteome has still remained unreferenced.

Non-coding RNAs translates into Ghost proteins

If translation is highly complex, the transcriptional landscape of organisms is as complex. Majority of DNA genomic sequences are transcribed into RNA sequences which are classified between protein-coding RNAs and ncRNAs [18]. Among ncRNAs, the RNA transcripts >200 nucleic acids are called long non-coding RNAs (lncRNAs). They were believed to be molecular relics from the last universal common ancestor of the RNA world and considered as nonfunctional nucleic acid sequences. However, ncRNAs are conserved across most of the cellular life and it was shown that ncRNAs have a vital importance in cells. Although not allowing the translation of proteins, they were demonstrated to be almost uniquely involved in the regulation of the information flow from DNA to protein [19,20]. ncRNAs have been shown to play on epigenetic regulation by modulating the accessibility of chromatin by polymerase, on histone methylation [21], and on transcriptome regulation to control protein expression. But they have also been demonstrated to exert other functions, such as assembling an lncRNA with the RNA-binding protein, TLS [22], as a cofactor modulating transcription factors. Involved in, the protein expression [23] or influencing the choice of RNA polymerase II binding site [24]. Finally, lncRNAs also enable post-transcriptional regulation, as described in ZEB2. The fixation of an LncRNA at the 5' UTR region of RNA causes a modification in the RNA splicing with preservation of the fixation site to the ribosome leading to the production of ZEB2 [25]. Interestingly, ZEB2 expression is subjected to several regulations by lncRNAs, including *i.e.* SPRY4-IT1 [26], LINC00461 [27], HOTAIRM1 [28]. Function of these lncRNAs is also associated with pathophysiological mechanisms. For example, lncRNA imatinib-upregulated (lncRNA-IUR) is described as a tumor suppressor which inhibits STAT5-CD71 signaling pathway [29].

However, it has recently been shown that some of these RNAs considered in first place as non-coding, are translated into proteins [30,31]. Ribosome profiling experiments have shown that ribosomes are able to bind and translate ncRNAs [32–34]. These additional functions of ncRNAs have pushed forwards the development of new databases, including additional protein sequences translated from ncRNA and lncRNA, and has contributed to the emergence of a new field of research, the proteogenomics. Moreover, peptides and proteins encoded by ncRNAs (HOXB-AS3, encoded by lncRNA; FBXW7-185aa, PINT-87aa, and SHPRH-146aa, encoded by circular RNA; and miPEP-200a and miPEP-200b, encoded by primary miRNAs) have been shown to be critical players in cancer development and progression, affecting the altered regulation of glucose metabolism, the epithelial-to-mesenchymal transition (EMT), and the ubiquitination pathway [35]. Thus, clearly establishing the role and function of the proteins translated from ncRNAs in the physiological and physiopathological context can be considered as the novel challenge in clinical and research areas, a major effort is necessary to achieve this goal.

AltORFs in mRNA translate into Ghost proteins

The second origin of AltProts is from the translation of AltORFs from mature mRNA which translate a RefProt. Alternative translation mechanisms were already foreseen from the work of Kozak, but these were considered then as minor events by comparison to the translation of RefProt. Nevertheless, translation products from the non-coding 5' and 3' UTR, overlapping between CDS and 5' or 3' UTR or reading frame shifts of one or two nucleotides according to the reference CDS (referred as CDS +2 or +3, CDS +1 being the reference CDS) are observed **(Figure 1B, 1C)**. For example, AltMRV11 protein is located at the 3′ UTR of the mRNA translated from the MRVI1 gene. AltMRVI1^{EGFP} was cloned and transfected into HeLa cells. AltMRVI1^{EGFP} was found to be localized in the nucleus [36] when MRVI1 is cytoplasmic, reflecting that AltMRVI1 is not a proteoform of MRVI1, but represents a completely different sequence to MRVI1. Furthermore, these proteins may represent sequence homologies with other RefProts issued from another mRNA or even ncRNA (**Figure 1D**). AltMRVI1, presents a sequence homology with BRCA1 interacting protein 3 (BRCA1-IP3) [37] and was demonstrated to interact with BRCA1. AltMRVI1 is, therefore, possibly a novel BRCA1 interacting partner that was already identified by yeast two hybrid, but mistakenly rejected as a false positive hit [38].

Ghost proteins: from identification to function

Approaches to AltProt Detection

MS-based proteomic strategies have so far demonstrated to be the most efficient approach to reach massive identification of the Ghost proteome if using a suitable databases, containing predicted sequence of AltProts. Most conventionally, non-targeted large-scale proteomic approaches are performed by **bottom-up**. The usually used **bottom-up shotgun** approaches which involve trypsin digestion of all proteins can be limited for smallest AltProts which do not always present enzymatic cleavage sites and other limitations like the ion charge state of peptides avoiding the identification [39]. This often limits the identification of AltProts because proteins identified on a single peptide match are normally not considered to be of confidence in standard proteomic processing workflow. Using large databases including AltProts have also shown to push the currently used algorithms for protein identification to their limit, in term of FDR calculation and potential false positive identification and redundancy in some peptide sequence based on the AltProt/RefProt homology. Those have been established for a few tens thousands of proteins included in databases but not for hundred thousand ones. It is foreseen that considering AltProts in future studies can push forwards the development of new bio-informatic solutions. Another central aspect of AltProt identification is the sample preparation. Optimized sample preparation protocols, adapted from peptides extraction, using MeOH or boiling water have shown to increase by 30% the number of identified AltProts [40–42]. Top-down proteomics, is another efficient strategy for identifying small proteins such as AltProts. From ovarian cancer tissues, spatially-resolved Top-down proteomics led to identify 15 AltProts which were then back-correlated to the MALDI-MS imaging [43].

Ribosome profiling is an alternative approach to proteomics that was used by different groups to unveil AltProts. Nicholas T.Ingolia and coworkers have used ribosome profiling to demonstrate that noncoding regions of mRNA and also 90% of ncRNA have translation sites capable of producing proteins [33]. Combining Ribosome profiling analysis to the prediction of ribosome binding sites on mRNA, enables to define a score of similarity between prediction and observation, named "**Fragment Length organization similarity score**"

(**FLOSS**) [33]. The measurement of this score gives additional weight to the probability of protein production in noncoding regions as well as in ncRNAs.

Unravelling the functions of Ghost Proteins

Identification of AltProts has revealed a new face of proteomics, challenging the one-gene one-protein dogma. However, although many of previous studies **(see Box2)** have shown the identification of translation sites and AltProts, in mammalian cell lines (HEK, HeLa, NCH82) and tissues (ovarian cancer, glioma, rat spinal cord), and even in extracellular vesicles [44], their functions remain largely under-investigated. As a start, predictions based on AltProt sequences can be achieved. Using homologies found between AltProts and RefProts, domains, families, specific sites, or motif can be identified. This way, more than ~15,000 AltProts were identified with a transmembrane domain or in secreted forms [8] supporting the idea that AltProts are involved in the cellular communication. Domains research via **InterProScan** [45] and **Gene Ontology (GO)**, makes possible to highlight involvement of AltProts in signaling pathways, nucleic acid or protein metabolism, transport or cell organization. Targeted studies, based on screening by CRISPR-Cas9, showed that more than 2,350 AltORFs are involved in modification of cell phenotype. Specifically, co-immunoprecipitation (co-IP) interactome of six lncRNAs has revealed, to name but one, that AltProt derived from lncRNA RP11_469A15.2, is involved in cytochrome C oxidase (COX) complex in the mitochondria [46]. Another way to explore AltProt functions is to search for their RefProt interaction partners [46] and then use this protein-protein interaction (PPI) to place them back into the signaling pathways. This strategy was experimented, first, by re-interrogating publicly available data of cross-linking MS (XL-MS) performed form the nuclei of HeLa cells [47,48]. Among the 1,679 cross-link interactions identified, 292 were shown to involve AltProts. 44 of these AltProts were found to interact with 7 RefProts related to ribonucleoproteins, ribosome subunits and zinc finger protein network. Among these AltProts identified the heterotrimer between the RE/poly(U)-binding/degradation factor 1 (AUF1), the Ribosomal protein 10 (RPL10) and AltATAD2 has been more deeply investigated. Docking models suggest the attachment of AUF1 on the external part of RPL10 and the interaction of AltATAD2 on the RPL10 region interacting with 5S ribosomal RNA as a mechanism of regulation of the ribosome [48]. More recently, new interactome studies by XL-MS, conducted on glioma cell lines in the context of cell transition [49]. have pointed out the involvement of AltProts with the ribosome confirming previous observations [48], where AltTRNAU1AP, AltEPHA5 and AltMAP2 are found in interaction with TPM4 protein [49]. Some of the identified AltProts and their function are listed in **Supplementary Table 1**. This demonstrates that AltProts are largely involved in replication and translation modulation but also in mitochondria, cell signaling and cell motility through cytoskeleton.

Ghost proteins and their function in physiological and physiopathological mechanisms

AltProt regulation dynamics

As of yet, only very few antibodies have been raised against AltProts for their detection, hence limiting deep functional analysis of specific AltProt candidates. Large-scale analytical pipelines such as above-mentioned XL-MS have shown to be very useful to gather wealth of information at global level on AltProt functions. Considering that AltProt average size is ~57 amino acid [36,50], some assumptions can be made based on their size and the dynamics of their production. For example, AltTRP1 from CDS+3 or AltBING-4 from CDS+2 are described to serve as “cryptic T-cell epitopes” which is of particular importance during thymal selection of T lymphocytes, leading to selection of lymphocytes with antiviral or antitumor activities [51]. Interestingly, it was shown, by studying the abundance of AltProts

and their associated RefProts, that AltProt regulation can be independent from that of the RefProt issued from same mRNA. For example, the main *MIEF1* translational product is not the canonical 463 amino acid MID51 protein but the small 70 amino acid alternative MID51 protein (AltMID51) which is 2-fold more abundant than MID51 [37] though AltMID51 appears to have a shorter lifetime. The re-analysis of data [49] assessing control cells vs. cells under phenotypic transition upon Forskolin treatment (a protein kinase A activator inducing cell differentiation and EMT), indicates major abundance variations between AltProts and RefProts from same mRNA. For AltANKZF1 and AltVPS52 (**Figure 2**), AltANKZF1 is found to be 10-fold more abundant than ANKZF1 in the control; however, after Forskolin stimulation, there is an increase of ANKZF1 which reaches then to the same abundance as AltANKZF1. Similar trend is observed with AltVPS52 and VPS52. In the control, AltVPS52 is 10-fold more abundant than VPS52 and after Forskolin treatment, AltVPS52 level remains higher than that of VPS52. This reflects the dynamic of AltProt expression, suggesting their potential implication in different signaling pathways. However, considering size similarities between AltProts with a mean length of 50 amino acids and polypeptides or neuropeptides typically 3 to 100 amino acids [52] (like the neuropeptide Y of 36 amino acids [53]) ; it can be speculate that AltProts can show similar functions to these peptides. For example, it is well known that snake, spider and Conus venoms contain venom peptides which are known to interact with receptors such as nicotinic acetylcholine receptors and can interfere with cholinergic transmission or have an impact on cardiac myocyte [54]. Neuropeptides also present a comparable size than some of AltProts. Hence, it is conceivable that among other functions, AltProts could be linked to the activation of orphan receptors.

AltProts in physiopathological mechanisms

AltProts have been described in various pathological contexts among which cancer **(Figure 3)**. In ovarian cancer, 15 AltProts have been highlighted, including 4 derived from tumor region in ovarian cancer: AltCMBL, AltGNL1, AltRP11-576E20.1, AltCSNK1A1L [43]. In breast cancer, AltProt CASIMO1, derived from an lncRNA, is described as an oncogene involved in MAPK activation signaling pathway [55]. AltEDARADD which is now annotated by UniProt (accession number L0R849), is shown to be with a high level of expression and associate with a poor prognosis for patients suffering from high grade serous ovarian cancer [40]. In esophageal cancer, lncRNA LINC00278 is described to promote the expression of the YY1BM AltProt [56]. Some AltProts have been as well described in the activation of the EMT [49,57]. AltProt designated as Humanin (HN) polypeptide, was found in Alzheimer disease as a suppressor of genes involved in neuronal death [58]. These evidences demonstrate that in many cases, AltProts are acting in regulation, playing for example on translation of RNAs and therefore regulating gene expression **(Figure 4.A)**. AltProts are involved in mechanisms that were thought to be well understood, though apparently one piece of the puzzle was missing. Cancer development is often driven by gene mutations; among others mutations in P53 or BRCA1/BRCA2 RefProts are well known. It is very likely, though not yet demonstrated, that mutations on AltProts are also driving some pathologies **(Figure 4.B)**. AltGNL1 which results from a shift in the ORF of the mRNA (nucleic acids 1487-1679) coding for RefProt GNL1 is an excellent example. The COSMIC database [59], which references the mutations of various cancers, describes 15 mutations in the region of the mRNA coding both for GNL1 and AltGNL1. Among the 15 mutations expected for AltGNL1; 12 will modify an amino acid of the protein **(Supplementary Table 2)**. The main one, which is a mutation of the nucleic acid 1646 (C>T, COSM8898738) impacts GNL1 by changing A>V. For AltGNL1, this mutation leads to the appearance of a premature stop codon. Thus, upon the effect of this mutation, AltGNL1 (64 amino acids) translates into a truncated proteoform of 53 amino acids, accordingly to the observations of

mRNA expression in endometrium carcinoma. Such modifications can significantly change their response and the cell phenotype. Therefore, tracking for mutations by co-analyses of non-coding regions of mRNA by RNA-seq together with AltProts identification will open the door into new target in cancer induced by mutation.

Concluding Remarks

As a result of the rules defining the translation to predict proteins in eukaryotes, the entire Ghost proteome has until recently, remained unknown. MS-based large-scale analysis and identification of ribosome binding sites on RNA by ribosome profiling has revealed its existence. It is now clear that, as in prokaryotes, eukaryote genes are polycistronic. This idea challenges the paradigm in the view of proteomics, transcriptomics and genetics; though, not radically a new idea since Kozak had already described this principle in 1999 [5]. The polycistronic nature of RNA limits the necessary material for synthesizing proteins, hence the production of AltProt is probably an evolutionary sign already carried out by bacteria [60], plants [61] and even some insects [62]. However, this "Lost World" of proteins remains to be thoroughly explored and might help explaining mechanisms that are still poorly understood. For example, one could hypothesize that AltProts are ligands of orphan receptors. But validating such an assumption will require functional analysis and unveiling the function of AltProts is definitely the next challenge. The ability of AltProts to present post-translational modifications [8] and mutations, establish protein-protein interactions [63], be active player in cellular communication. This can be through by extra cellular vesicle release and bind to membranes or cytoskeleton, demonstrates that they share a similar diversity and complexity in their functions than RefProts. Global large-scale analyses is one way to rapidly predict AltProts function by identifying domains and finding their interaction partners. However, studies have only yet lifted the veil on this new world of proteins and there are still many opened questions to be answered (see "**Outstanding Questions**") and future challenges ahead. Among others the role of AltProts in physiological and pathophysiological mechanisms is one [64]. Another is to grasp knowledge about AltProt mutations in relation to pathologies. Moreover, it is clear that AltProts can potentially interfere with inhibitors of RefProt drug targets which would explain limited efficiency of certain drugs. Further, potentiality of a part of the Ghost proteome to act as a rescue system for cell is a reasonable assumption and something to be investigated. Finally, presence of a subset of mRNA products also rises questions on the strategies developed to target protein expression at DNA level. In particular, CRISPR-Cas9 technology shows a high frequency of off-target activity (≥50%) [65,66] which is a major concern to therapeutic and clinical applications [67]. In light of alternative translation mechanisms, such off-targets can lead to the presence of several AltProts with various specific biological functions. The downstream impact of releasing such sub-products is unknown and might induces aberrant mechanisms and an opposite effect. In a nutshell, by unlocking the Ghost proteome, researchers have opened the Pandora's Box.

ACKNOWLEDGMENT

This research was supported by funding from Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation (MESRI), Institut National de la Santé et de la Recherche Médicale (Inserm) and Université de Lille.

Figures are realized on the Adobe Illustrator software.

1 Breuza, L. *et al.* (2016) The UniProtKB guide to the human proteome. *Database* 2016, bav120

2 Bateman, A. *et al.* (2017) UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* 45, D158–D169

3 Bateman, A. *et al.* (2015) UniProt: A hub for protein information. *Nucleic Acids Res.* 43, D204–D212

4 Kozak, M. Regulation of translation in eukaryotic systems. , *Annual Review of Cell Biology*, 8. 28-Nov-(1992) , Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA, 197–225

5 Kozak, M. (1999) Initiation of translation in prokaryotes and eukaryotes. *Gene* 234, 187–208

6 Vanderperre, B. *et al.* (2012) HAltORF: Adatabase of predicted out-of-frame alternative open reading frames in human. *Database* 2012, bas025

7 Mouilleron, H. *et al.* Death of a dogma: Eukaryotic mRNAs can code for more than one protein. , *Nucleic Acids Research*, 44. 08-Jan-(2016) , Oxford University Press, 14–23

8 Samandi, S. *et al.* (2017) Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. *Elife* 6, e27860

9 Hershey, J.W.B. (1991) Translational Control in Mammalian Cells. *Annu. Rev. Biochem.* 60, 717–755

10 Kozak, M. (1987) An analysis of 5'-noncoding sequences from 699 vertebrate messenger rNAS. *Nucleic Acids Res.* 15, 8125–8148

11 Kozak, M. (1986) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* 44, 283–292

12 Fouillot, N. *et al.* (1993) Translation of the hepatitis B virus P gene by ribosomal scanning as an alternative to internal initiation. *J. Virol.* 67, 4886–4895

13 Kozak, M. An analysis of vertebrate mRNA sequences: Intimations of translational control. , *Journal of Cell Biology*, 115. (1991) , The Rockefeller University Press, 887–903

14 Prats, H. *et al.* (1989) High molecular mass forms of basic fibroblast growth factor are initiated by alternative CUG codons. *Proc. Natl. Acad. Sci. U. S. A.* 86, 1836–1840

15 Bugler, B. *et al.* (1991) Alternative initiation of translation determines cytoplasmic or nuclear localization of basic fibroblast growth factor. *Mol. Cell. Biol.* 11, 573–577

16 Hann, S.R. *et al.* (1988) A non-AUG translational initiation in c-myc exon 1 generates an N-terminally distinct protein whose synthesis is disrupted in Burkitt's lymphomas. *Cell* 52, 185–195

17 Möller, S. *et al.* (1999) EDITtoTrEMBL: a distributed approach to high-quality automated protein sequence annotation. *Bioinformatics* 15, 219–27

18 Quinn, J.J. and Chang, H.Y. Unique features of long non-coding RNA biogenesis and function. , *Nature Reviews Genetics*, 17. 01-Jan-(2016) , Nature Publishing Group, 47–62

19 Poole, A. *et al.* (1999) Early evolution: Prokaryotes, the new kids on the block. *BioEssays* 21, 880–889

20 Jeffares, D.C. *et al.* (1998) Relics from the RNA world. *J. Mol. Evol.* 46, 18–36

21 Rinn, J.L. *et al.* (2007) Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs. *Cell* 129, 1311–1323

22 Wang, X. *et al.* (2008) Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. *Nature* 454, 126–130

23 Feng, J. *et al.* (2006) The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. *Genes Dev.*

20, 1470–1484
24 Martianov, I. *et al.* (2007) Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature* 445, 666–670
25 Beltran, M. *et al.* (2008) A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition. *Genes Dev.* 22, 756–769
26 Yao, H. *et al.* (2020) LncRNA SPRY4-IT1 promotes progression of osteosarcoma by regulating ZEB1 and ZEB2 expression through sponging of miR-101 activity. *Int. J. Oncol.* 56, 85–100
27 Li, X. *et al.* (2020) The Lnc LINC00461/miR-30a-5p facilitates progression and malignancy in non-small cell lung cancer via regulating ZEB2. *Cell Cycle* 19, 825–836
28 Lin, Y.H. *et al.* (2020) Long non-coding RNA HOTAIRM1 promotes proliferation and inhibits apoptosis of glioma cells by regulating the miR-873-5p/ZEB2 axis. *Chin. Med. J. (Engl).* 133, 174–182
29 Wang, X. *et al.* (2019) Novel lncRNA-IUR suppresses Bcr-Abl-induced tumorigenesis through regulation of STAT5-CD71 pathway. *Mol. Cancer* 18, 84
30 Rion, N. and Rüegg, M.A. LncRNA-encoded peptides: More than translational noise? , *Cell Research*, 27. 01-May-(2017) , Nature Publishing Group, 604–605
31 Matsumoto, A. *et al.* (2017) MTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature* 541, 228–232
32 Ingolia, N.T. *et al.* (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 147, 789–802
33 Ingolia, N.T.T. *et al.* (2014) Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. *Cell Rep.* 8, 1365–1379
34 Bazzini, A.A. *et al.* (2014) Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.* 33, 981–993
35 Wang, J. *et al.* ncRNA-Encoded Peptides or Proteins and Cancer. , *Molecular Therapy*, 27. 02-Oct-(2019) , Cell Press, 1718–1725
36 Vanderperre, B. *et al.* (2013) Direct Detection of Alternative Open Reading Frames Translation Products in Human Significantly Expands the Proteome. *PLoS One* 8, e70698
37 Delcourt, V. *et al.* (2018) The Protein Coded by a Short Open Reading Frame, Not by the Annotated Coding Sequence, Is the Main Gene Product of the Dual-Coding Gene MIEF1. *Mol. Cell. Proteomics* 17, 2402–2411
38 Liu, Y. *et al.* (2011) Yeast two-hybrid junk sequences contain selected linear motifs. *Nucleic Acids Res.* 39, e128
39 Slavoff, S.A. *et al.* (2013) Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.* 9, 59–64
40 Cardon, T. *et al.* (2019) Optimized Sample Preparation Workflow for Improved Identification of Ghost Proteins. DOI: 10.1021/acs.analchem.9b04188
41 Cassidy, L. *et al.* (2019) Depletion of High-Molecular-Mass Proteins for the Identification of Small Proteins and Short Open Reading Frame Encoded Peptides in Cellular Proteomes. *J. Proteome Res.* 18, 1725–1734
42 Cassidy, L. *et al.* (2016) Combination of Bottom-up 2D-LC-MS and Semi-top-down GelFree-LC-MS Enhances Coverage of Proteome and Low Molecular Weight Short Open Reading Frame Encoded Peptides of the Archaeon Methanosarcina mazei. *J. Proteome Res.* 15, 3773–3783
43 Delcourt, V. *et al.* (2017) Combined Mass Spectrometry Imaging and Top-down Microproteomics Reveals Evidence of a Hidden Proteome in Ovarian Cancer. *EBioMedicine* 21, 55–64

44 Murgoci, A.N. *et al.* (2020) Reference and Ghost Proteins Identification in Rat C6 Glioma Extracellular Vesicles. *iScience* 23, 101045

45 Jones, P. *et al.* (2014) InterProScan 5: Genome-scale protein function classification. *Bioinformatics* 30, 1236–1240

46 Chen, J. *et al.* (2020) Pervasive functional translation of noncanonical human open reading frames. *Science (80-.).* 367, 140–146

47 Liu, F. *et al.* (2015) Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. *Nat. Methods* 12, 1179–1184

48 Cardon, T. *et al.* (2019) Nuclei of HeLa cells interactomes unravel a network of ghost proteins involved in proteins translation. *Biochim. Biophys. Acta - Gen. Subj.* DOI: 10.1016/j.bbagen.2019.05.009

49 Cardon, T. *et al.* (2020) Alternative proteins are functional regulators in cell reprogramming by PKA activation. *Nucleic Acids Res.* DOI: 10.1093/nar/gkaa277

50 Delcourt, V. *et al.* (2018) Small Proteins Encoded by Unannotated ORFs are Rising Stars of the Proteome, Confirming Shortcomings in Genome Annotations and Current Vision of an mRNA. *Proteomics* 18, 1700058

51 Ho, O. and Green, W.R. (2006) Unexpected Cryptic T Cell Epitopes: Expecting the Alternative Translational Products and. DOI: 10.4049/jimmunol.177.12.8283

52 Wang, Y. *et al.* (2015) NeuroPep: A comprehensive resource of neuropeptides. *Database* 2015,

53 Reichmann, F. and Holzer, P. Neuropeptide Y: A stressful review. , *Neuropeptides*, 55. 01-Feb-(2016) , Churchill Livingstone, 99–109

54 Utkin, Y.N. (2019) Last decade update for three-finger toxins: Newly emerging structures and biological activities. *World J. Biol. Chem.* 10, 17–27

55 Polycarpou-Schwarz, M. *et al.* (2018) The cancer-associated microprotein CASIMO1 controls cell proliferation and interacts with squalene epoxidase modulating lipid droplet formation. *Oncogene* 37, 4750–4768

56 Wu, S. *et al.* (2020) A Novel Micropeptide Encoded by Y-Linked LINC00278 Links Cigarette Smoking and AR Signaling in Male Esophageal Squamous Cell Carcinoma. *Cancer Res.* DOI: 10.1158/0008-5472.can-19-3440

57 Vergara, D. *et al.* (2020) A Hidden Human Proteome Signature Characterizes the Epithelial Mesenchymal Transition Program. *Curr. Pharm. Des.* 26, 372–375

58 Hashimoto, Y. *et al.* (2001) A rescue factor abolishing neuronal cell death by a wide spectrum of familial Alzheimer's disease genes and Abeta. *Proc. Natl. Acad. Sci. U. S. A.* 98, 6336–41

59 Bamford, S. *et al.* (2004) The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br. J. Cancer* 91, 355–358

60 Güell, M. *et al.* Bacterial transcriptomics: What is beyond the RNA horiz-ome? , *Nature Reviews Microbiology*, 9. Sep-(2011) , 658–669

61 Hanada, K. *et al.* Small open reading frames associated with morphogenesis are hidden in plant genomes. DOI: 10.1073/pnas.1213958110

62 Ladoukakis, E. *et al.* (2011) Hundreds of putatively functional small open reading frames in Drosophila. *Genome Biol.* 12, R118

63 Cardon, T. *et al.* (2020) Optimized Sample Preparation Workflow for Improved Identification of Ghost Proteins. *Anal. Chem.* 92, 1122–1129

64 Le Rhun, E. *et al.* (2017) Evaluation of non-supervised MALDI mass spectrometry imaging combined with microproteomics for glioma grade III classification. *Biochim. Biophys. Acta - Proteins Proteomics* 1865, 875–890

65 Fu, Y. *et al.* (2013) High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.* 31, 822–826

66 Tuladhar, R. *et al.* (2019) CRISPR-Cas9-based mutagenesis frequently provokes on-target mRNA misregulation. *Nat. Commun.* 10, 4056
67 Han, H.A. *et al.* Mitigating off-target effects in CRISPR/Cas9-mediated in vivo gene editing. , *Journal of Molecular Medicine*, 98. 01-May-(2020) , Springer, 615–632
68 Slavoff, S.A. *et al.* (2014) A human short open reading frame (sORF)-Encoded polypeptide that stimulates DNA end joining. *J. Biol. Chem.* 289, 10950–10957
69 Hao, Y. *et al.* (2018) SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. *Brief. Bioinform.* 19, 636–643
70 Peeters, M.K.R. and Menschaert, G. (2020) The hunt for sORFs: A multidisciplinary strategy. *Exp. Cell Res.* DOI: 10.1016/j.yexcr.2020.111923
71 Brunet, M.A. *et al.* (2019) OpenProt: A more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucleic Acids Res.* 47, D403–D410
72 Olexiouk, V. *et al.* (2017) An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res.* 46, 497–502
73 Brunet, M.A. *et al.* (2020) Reconsidering proteomic diversity with functional investigation of small ORFs and alternative ORFs. *Exp. Cell Res.* 393, 112057
74 Fairman, R. *et al.* (1993) Multiple oligomeric states regulate the DNA binding of helix-loop-helix peptides. *Proc. Natl. Acad. Sci. U. S. A.* 90, 10429–10433
75 Tabb-Massey, A. *et al.* Ribosomal proteins Rps0 and Rps21 of Saccharomyces cerevisiae have overlapping functions in the maturation of the 3¢ end of 18S rRNA. DOI: 10.1093/nar/gkg899
76 Vattem, K.M. and Wek, R.C. (2004) Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* 101, 11269–11274
77 Chipuk, J.E. and Green, D.R. How do BCL-2 proteins induce mitochondrial outer membrane permeabilization? , *Trends in Cell Biology*, 18. 01-Apr-(2008) , Elsevier Current Trends, 157–164
78 O'Leary, N.A. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733–D745
79 Pueyo, J.I. *et al.* New Peptides Under the s(ORF)ace of the Genome. , *Trends in Biochemical Sciences*, 41. Aug-(2016) , 665–678
80 Ingolia, N.T. Ribosome profiling: New views of translation, from single codons to genome scale. , *Nature Reviews Genetics*, 15. 28-Jan-(2014) , Nature Research, 205–213
81 Ingolia, N.T. *et al.* (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science (80-.).* 324, 218–223

Glossary

43S ribosome: The preinitiation complex (43S PIC) is a ribonucleoprotein complex that exists during an early step of the translation. The formation of the 43s ribosome complex arises from the recycling of the small ribosome 40S ribosomal subunit bound by the initiation factors (eIF): eIF1, eIF3 and eIF1A. This is followed by the docking of eIF5 and eIF2, itself transporting the tRNA-MET, allowing the recognition of the START codon and the binding of the first amino acid (methionine).

Bottom-up: Bottom-up is the most used large-scale proteomic approach for the identification and relative quantification of proteins in a single experiment. It involves a digestion of the proteins using one (generally trypsin) or several enzymes (*e.g.* LysC/trypsin) and the MS identification of the proteins through measuring their digestion peptides.

Bottom-up shotgun*:* Shotgun is a sub-type of bottom-up approaches where the digestion of all proteins together is performed after the extraction; the tryptic peptides being further analyzed by LC-MS. In shotgun the protein identification is solely based on the molecular weight of the peptides and the structural information collected in tandem MS spectra after gas phase fragmentation of the ions.

Darkome: The Darkome designates transcripts, proteins and metabolites that are missing annotation in the databases. For example, the AltProts and their transcripts showing mutations that could be pathology drivers, such as for cancer, as still missing annotation. Indeed, mutations which are considered as silent mutations because not in the region of the transcript sequence related to the RefProt translation could be non-silencing for the AltProts. Altogether, the Darkome represents still hundred thousand of unknown products.

Fragment Length Organization Similarity Score (FLOSS): The FLOSS is defined as,

$$0.5 \times \sum_{I=26}^{34} f(I) - f_{ref}(I)$$

where f(I) is the fraction of reads of length l in the transcript histogram and $f_{ref}(I)$ is the corresponding fraction in the reference histogram (raw count excluding gene overlap and noncoding).

Gene Ontology (GO): The GO is the largest world knowledgebase source of information, across all species, on the functions of genes that is both human-readable and machine-readable.

InterProScan: InterProScan is a free software that provides functional analysis of proteins by classifying them into families and predicting domains and important sites. It uses predictive models and kown signatures from several different databases from the InterPro consortium.

Monocistronic: refers to the ability of an mRNA to be translated into a single protein chain. Eukaryotes are considered monocistronic.

Reinitiation and leaky scanning: These are 2 mechanisms explaining the translation of AltORFs. The reinitiation mechanism is due to the formation of the 80S ribosome upstream of the most likely AUG site. This small region, formerly called upstream ORF (upORF) is described as not exceeding 30 codons, if its length is greater (more than 120 codons) it no longer allows linking with the main ORF. The upORF region allows the regulation of the transcription of the downstream protein (RefProt), however the role of the small protein is generally attributed to a non-functional peptide. The leaky scanning allows to pass the rule of the first codon "AUG". The second or third Start codon will then be used. The mechanism is due to the absence of an optimal initiation context. Indeed, if the Kozak context around the AUG codon is not strong enough, the 40S subunit does not stop at the expected Start codon and translates the following region.

TrEMBL: Translation of the EMBL (TrEMBL) is a large protein database in SwissProt format generated by computer translation of all coding sequences in the EMBL Nucleotide Sequence Database database, that are not in SWISS-PROT. Unlike SWISS-PROT, TrEMBL entries are manually annotated.

Text Boxes:

Box1: The Alternative Protein kingdom.

The alternative proteins terminology encompasses a large number of proteins issued from alternative mechanisms (e.g. **reinitiation and leaky scanning**) and various origins (mRNA and ncRNA). These translation products give rise to proteins (as experimentally evidenced) which are in average of small size (<57 amino acids). Currently, there is a clear lack of consensus regarding the terminology of this novel class of proteins in the literature, one reason being the different approaches and angles by which they were initially studied by different groups. Indeed, in the literature, they can be referred to as, short ORFs (sORFs) originating *i.e.* sORF-Encoded Polypeptides (SEPs) [68], Small Proteins (smPROT) [69], polypeptides [70] or AltProts [71]. Indeed, sORFs from the sORFs.org repository [72], are described as ORF in the range of 10 to 100 codons, issued from ribosome profiling analysis, therefore accepting other than AUG start codon, but not including shifts in the CDS. While, AltORF coding AltProt are described as >30 codons, starting with an AUG codon but also found in reading frame shifts [73]. In 2012, 17,096 AltProts sequences were predicted from ~31,422 mRNAs or 8,744 genes from the human genome and referenced as the "Human alternative Open Reading Frame" (HaltORF) database [6]. In this initial database, 83% of predicted AltProts were in the CDS+2 and 17% in the CDS+3 (CDS+1 corresponding to the RefProt). However, this was only taking into account the reading offsets, without referencing the predicted translation products in the 5', 3' UTR regions or ncRNA. In 2019, all of the predictions were included and are now available in the public "OpenProt" database [71]. It is currently predicted ~450,000 proteins in human, without considering post-translational modifications. AltProt prediction from other species such as mice, rat and zebrafish are also now accessible in OpenProt. Because we wish to keep here the big picture on these novel proteins independently of their length in amino acids or nucleotides, their origin (mRNA, ncRNA) or their initiation by an AUG start codon, we prefer to refer to them as proteins issued from alternative mechanisms to the RefProts, or as Alternative Proteins (AltProts).

Box2: History of the Ghost proteome evidence

Despite the general dogma on the monocistronic nature of eukaryotes mRNAs, it was already demonstrated 20 years ago that mRNA as well as certain ncRNA could translate into several proteins. Indeed, in 2000 the assessment of the ribosome profiling has led to prove the capacity of the ribosome to bind on noncoding region of mRNAs and ncRNAs [32,33]. Before the advent of large-scale proteomics, few proteins issued from other ORFs than the RefORF have been punctually discovered and identified such as id peptide [74], RPS21 [75], AltATF4 [76] or BCL-2 [77]. Since protein databases used in large scale proteomic approaches (e.g. UniprotKB [3] or NCBI [78]) are based on genomic annotations, which rule out AltORFs and ncRNA, AltProts have been missing and hence, could not be identified through MS-based proteomics. More recently, it was demonstrated that proteins issued from AltORFs were far more common than initially expected. The first massive identification of AltProts was obtained in 2011 at the protein level by shotgun proteomics using high-resolution (HR) LC-MS. Surprisingly, large scale proteomic bottom-up experiments were always showing more than 10% of tandem MS data remaining unmatched after protein database interrogation. This has led to consider alternative translation mechanisms in the protein predictions and to generate a first database of human AltProts (HaltORF) [6]; though this was far from being complete. A total of 1,259 AltProts were initially identified by shotgun proteomics from human cell lines, FFPE tissues and fluids, both from normal and pathological samples [43]. These results have opened the door to several *in silico* studies to

assess these AltProt databases, leading to the protein interrogation in large-scale proteomic experiments and further evaluate their physiological and pathophysiological roles.

In the following years **(Figure I)**, high-throughput genome and transcriptome sequencing was used to validate an important number of sORFs and AltORF containing less than 100 codons that were previously arbitrary considered as noncoding sequences. The translation of these sORFs into peptides or small proteins (AltProts, microproteins, SEPs) was demonstrated by different strategies such as ribosome profiling [70,79–81] and/or peptidomics coupled to massively parallel RNA-seq [39]. This accumulating evidence both demonstrated by genome annotation and protein translation, has largely questioned the validity of the gene-protein dogma pushing forward the development of new databases including the AltProts as a mandatory step for their identification in large-scale proteomic studies.

Figure Legends:

Figure 1.

Scheme of AltProts expression regions. **A)** The conventional vision of the eukaryotic protein translation: where the RefProt expression comes from the first START codon "AUG", with a CDS framed by a 5'& 3'UTR region. **B)** The 5' & 3' UTR regions can also be the location of the ribosome binding, allowing the translation of AltProt$^{5'UTR}$ or AltProt$^{3'UTR}$. **C)** The ribosome can also shift of the reading frame with a fixation at +2 or +3 of the reference START codon, thus allowing the translation of AltProt^{CDS+2} or AltProt^{CDS+3}. **D)** The last AltProt translation source is ncRNA and lncRNA.

Figure 2.

Variation of the relative abundance (LFQ) of AltProt vs. RefProt associated with a same mRNA. Each condition has been performed in triplicates **A)** AltANKZF1 vs. ANKZF1. The RefProt ANKZF1 shows a lower relative abundance than its AltProt in the control condition, while they show the same abundance under FRSK stimulation, meaning that FRSK stimulation increase the level of ANKZF1 but does not influence AltANKZF1. **B)** AltVPS52 vs. VPS52. AltVPS52 has a higher abundance than the reference VPS52 but both protein show the same regulation trend under FRSK stimulation.

Figure 3.

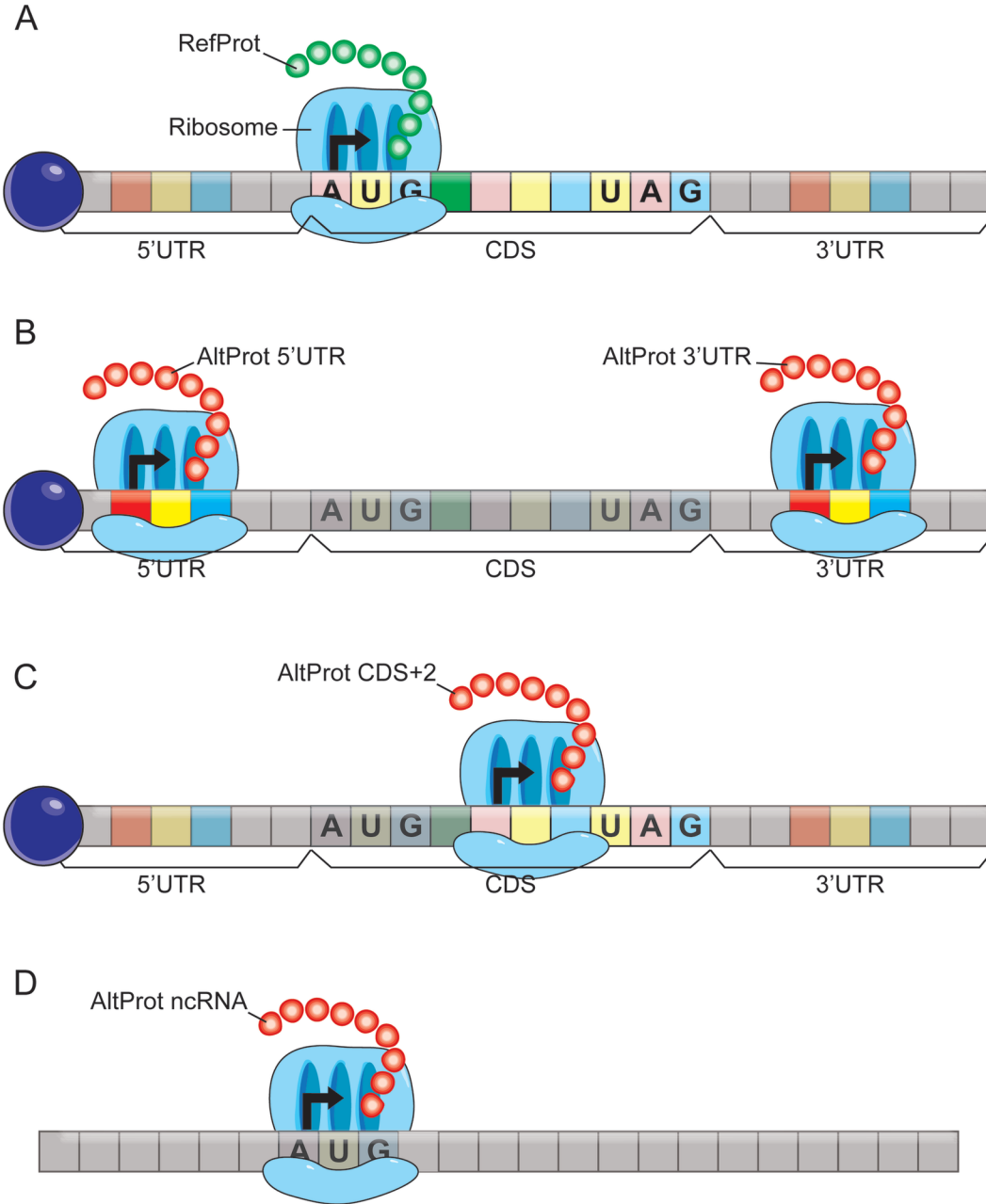
Pathophysiological implication and identification of AltProt. The Ghost proteome drives different proteins involved in the development of cancer. AltProts are plotted back on cancer hallmarks showing that already identified AltProt covers most of hallmarks (in blue circle) and cancers (in red square).

Figure 4.

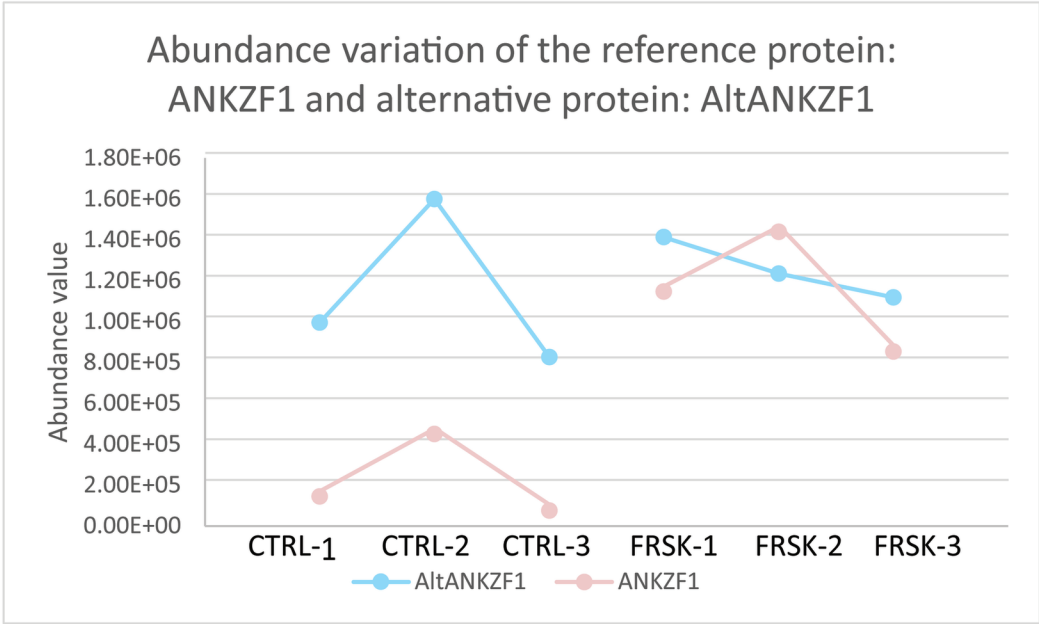
Proteins expression is multi-layered, involving genes, mRNAs, and allowing the translation of RefProts and AltProts. **A)** The conventional representation admits the expression of a single RefProt derived from an mRNA transcript itself derived from a DNA gene, however the presence of AltProt is now proven. **B)** Genetic mutations impact the transcriptome. If some mutations are silencing mutation to the RefProt they will impact region involved in AltProts translation. The study of AltProt mutations in pathologies is a yet unexplored field that could be highly beneficial and provide massive answers to pathophysiological mechanisms.

Figure I.

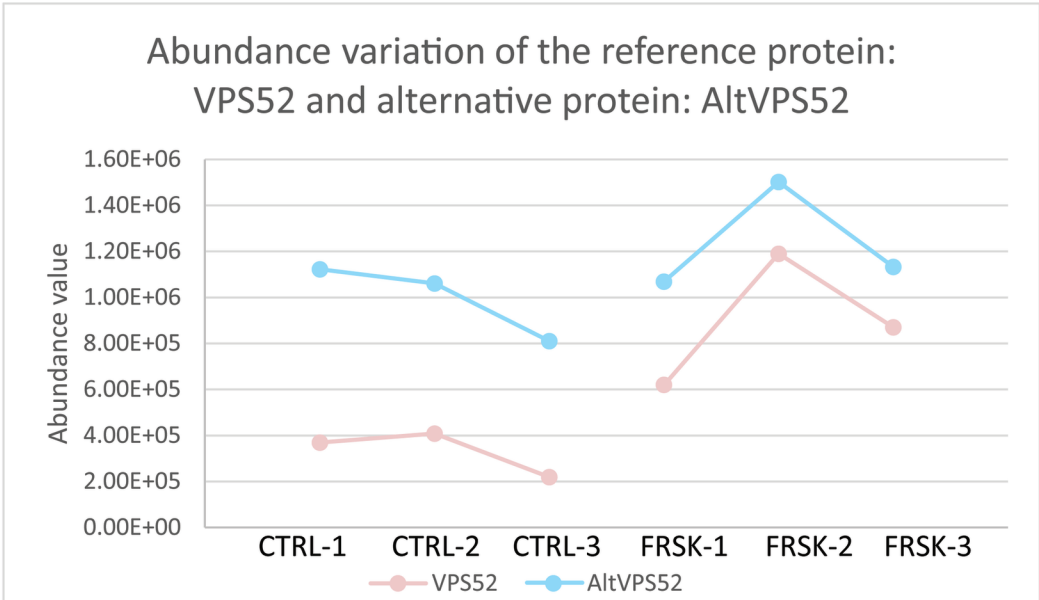
The key events of the Ghost proteome timeline.

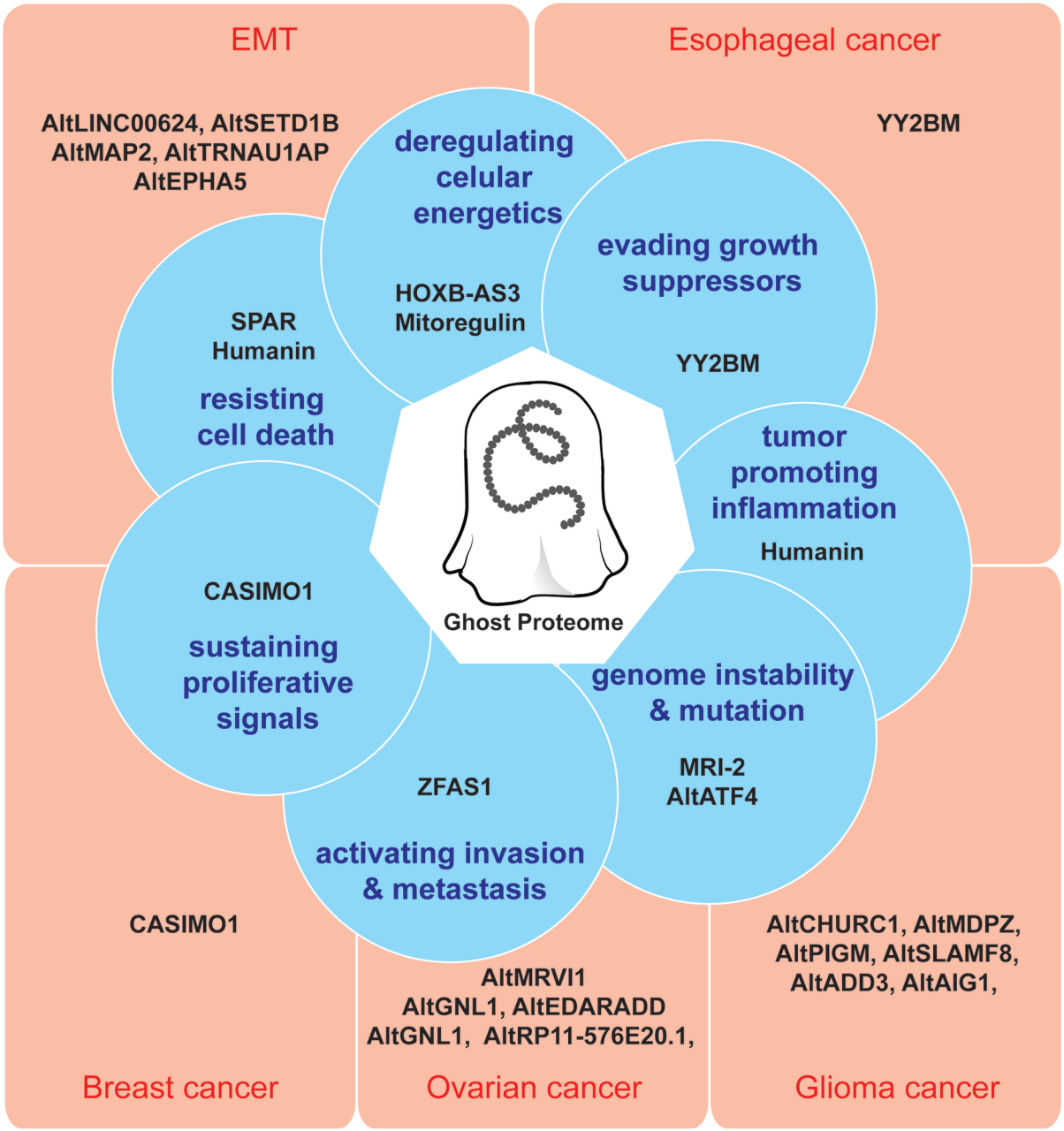

A
RefProt
Ribosome
A U G
U A G
5'UTR
CDS
3'UTR
B
AltProt 5'UTR
AltProt 3'UTR
A U G
U A G
5'UTR
CDS
3'UTR
C
AltProt CDS+2
A U G
U A G
5'UTR
CDS
3'UTR
D
AltProt ncRNA
A U G

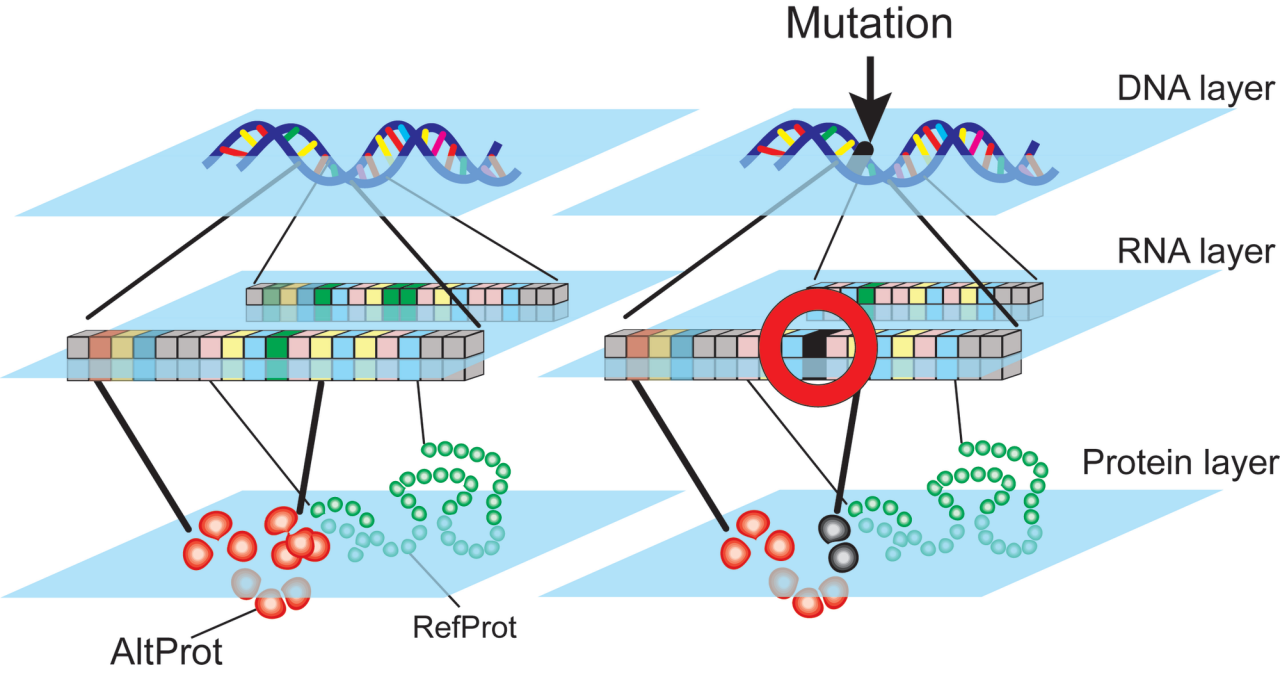
A


Abundance variation of the reference protein:
ANKZF1 and alternative protein: AltANKZF1
Abundance value
1.80E+06
1.60E+06
1.40E+06
1.20E+06
1.00E+06
8.00E+05
6.00E+05
4.00E+05
2.00E+05
0.00E+00
CTRL-1
CTRL-2
CTRL-3
FRSK-1
FRSK-2
FRSK-3
AltANKZF1
ANKZF1

B


Abundance variation of the reference protein:
VPS52 and alternative protein: AltVPS52
Abundance value
1.60E+06
1.40E+06
1.20E+06
1.00E+06
8.00E+05
6.00E+05
4.00E+05
2.00E+05
0.00E+00
CTRL-1
CTRL-2
CTRL-3
FRSK-1
FRSK-2
FRSK-3
VPS52
AltVPS52


EMT
AltLINC00624, AltSETD1B
AltMAP2, AltTRNAU1AP
AltEPHA5
Esophageal cancer
YY2BM
deregulating celular energetics
HOXB-AS3
Mitoregulin
evading growth suppressors
YY2BM
SPAR
Humanin
resisting cell death
tumor promoting inflammation
Humanin
Ghost Proteome
CASIMO1
sustaining proliferative signals
genome instability & mutation
MRI-2
AltATF4
ZFAS1
activating invasion & metastasis
CASIMO1
AltMRVI1
AltGNL1, AltEDARADD
AltGNL1, AltRP11-576E20.1,
AltCHURC1, AltMDPZ,
AltPIGM, AltSLAMF8,
AltADD3, AltAIG1,
Breast cancer
Ovarian cancer
Glioma cancer


Mutation
DNA layer
RNA layer
Protein layer
AltProt
RefProt