RNA BIOLOGY
2025, VOL. 22, NO. 1, 1–18
https://doi.org/10.1080/15476286.2025.2498203

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

# The implication of non-AUG-initiated N-terminally extended proteoforms in cancer

Rita Pancsa [a*], Dmitry E. Andreev [b,c*], and Kellie Dean [d]

[a]Institute of Molecular Life Sciences, HUN-REN Research Centre for Natural Sciences, Budapest, Hungary; [b]Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, RAS, Moscow, Russia; [c]Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, Moscow, Russia; [d]School of Biochemistry and Cell Biology, University College Cork, Cork, Ireland

**ABSTRACT**

Dysregulated translation is a hallmark of cancer, and recent genome-wide studies in tumour cells have uncovered widespread translation of non-canonical reading frames that often initiate at non-AUG codons. If an upstream non-canonical start site is located within a frame with an annotated coding sequence (CDS), such translation events can lead to the production of proteoforms with altered N-termini (PANTs). Certain examples of PANTs from oncogenes (e.g. c-MYC) and tumour suppressors (e.g. PTEN) have been previously linked to cancer. We have performed a systematic computational analysis on recently identified non-AUG initiation-derived N-terminal extensions of cancer-associated proteins, and we discuss how these extended proteoforms may acquire new oncogenic properties. We identified a loss of stability for the N-terminally extended proteoforms of oncogenes TCF-4 and SOX2. Furthermore, we discovered likely functional short linear motifs within the N-terminal extensions of oncogenes and tumour suppressors (SOX2, SUFU, SFPQ, TOP1 and SPEN/SHARP) that could provide an explanation for previously described functionalities or interactions of the proteins. In all, we identify novel cases where PANTs likely show different localization, functions, partner binding or turnover rates compared to the annotated proteoforms. Therefore, we propose that alterations in the stringency of translation initiation, often seen under conditions of cellular stress, may result in reprogramming of translation to generate novel PANTs that influence cancer progression.

**ARTICLE HISTORY**
Revised 3 April 2025
Accepted 21 April 2025

**KEYWORDS**
N-terminal extension; translation initiation; start codon; non-AUG initiation; alternative translation start site; proteoforms with altered N-termini; short linear motifs

## Introduction

A crucial step in protein synthesis is the selection of the translation initiation site on the mRNA. In eukaryotic cells, this process is largely accomplished through a scanning mechanism. The pre-initiation complex (PIC), which consists of the 40S ribosomal subunit, certain translation initiation factors, and a methionine-tRNA (met-tRNA), enters the 5'end of the mRNA and scans the mRNA sequence from 5' to 3'. Recognition of a suitable start codon, typically AUG, leads to the dissociation of the initiation factors and the joining of the 60S subunit (reviewed in [1,2]). The ability of the PIC to reach and select a specific start codon on the mRNA depends on several factors, including the availability and activity of the translation initiation factors, the structure of the mRNA, and RNA-binding proteins. These factors can interfere with the scanning process and the selection of the start codon. Differences in mRNA sequences, which are inspected by scanning proteins, dictate translation factor dependency. While some mRNAs require a set of 'canonical' initiation factors for translation, others require additional non-canonical factors (reviewed in [3]). This allows for selective translation control, where alterations in signalling pathways differentially affect the translation of various mRNA species, resulting in different proteomic changes.

When the PIC (pre-initiation complex) scans the mRNA 5' leader region, it encounters a potential start codon that can form a codon–anticodon interaction with the met-tRNA in the P site. The optimality of this interaction is probed by multiple components of the PIC, including rRNA bases, ribosomal proteins, and initiation factors. The efficiency of start codon recognition depends on several factors, including the identity of the codon (some non-AUG codons can also be recognized), its nucleotide context, and the presence of certain initiation factors, such as eIF1, eIF1A, and eIF5. Due to these factors, the scanning ribosome can continue scanning if the start codon is not optimal, a phenomenon known as leaky scanning [4]. Leaky scanning means that translation initiation can occur at more than one start codon on certain mRNAs.

If these start codons are located within the same open reading frame and upstream of each other, translation can produce polypeptides with extended N-termini [5,6] which may yield alternative proteoforms. The term 'proteoform' describes all molecular variants in which a protein can exist and is typically used to refer to differences at the protein level

**CONTACT** Kellie Dean k.dean@ucc.ie School of Biochemistry and Cell Biology, University College Cork, Cork, T12 XF62, Ireland
*These authors contributed equally to the work.
Supplemental data for this article can be accessed online at https://doi.org/10.1080/15476286.2025.2498203

(i.e. translational products). While isoforms are typically produced by alternative splicing of the mRNAs, proteoforms can be produced by translational and/or post-translational mechanisms, such as alternative translation initiation, stop codon readthrough, posttranslational modifications or proteolytic processing. Although we cannot precisely predict all the posttranslational mechanisms, and thus the actual proteoforms that will exist in cells, we refer to the N-terminally extended polypeptides as proteoforms because they differ from the canonical forms on the level of translation.

Deregulated translation is a hallmark of cancer cells [7–14]. Regarding non-AUG initiation events, there are numerous examples of N-terminally extended proteoforms for genes involved in cancer progression, one of the earliest such discoveries was made for the *c-MYC* oncogene in 1988 [15]. By analysing the *in vitro* protein-coding capacity of *c-MYC* cDNAs, Hann et al. found that c-MYC1 (p67) proteoform was initiated by an upstream CUG codon, adding a 14-amino acid extension; whereas the c-MYC2 (p64) proteoform was initiated by the canonical AUG start [15]. While there was lower translation efficiency of *c-MYC1*, with total synthesis about 10–15% of the level of the canonical protein, this changed when cells were grown to high densities, and this effect seemed to be due to amino acid restriction [16,17]. Overall, early studies on *c-MYC* translation pointed to modulation of the scanning ribosomal pre-initiation complex that impacted start codon selection, resulting in the production of different MYC proteoforms. More recently, work by Sato *et al.* (2019), showed that the balance of c-MYC proteoforms could be tipped in favour of the AUG-initiated isoform by the oncogenic protein, 5MP1 (eukaryotic initiation factor 5 (eIF5)-mimic protein) that competes with eIF5 [18] and promotes malignancy in colorectal cancer by translational reprogramming [19].

In a cancer context, tumour suppressor proteins act as negative regulators in biochemical and cellular processes, essentially serving as brakes within a system. Indeed, loss or disabling mutations of tumour suppressor genes, like *RB1*, *TP53* and *PTEN*, are found in numerous cancers [20]. The phosphatase and tensin homolog on chromosome ten, *PTEN*, encodes a protein that blocks the activation of the phosphatidylinositol 3-kinase (PI3K)/AKT pathway by dephosphorylating lipid substrates [21,22], thereby influencing cell proliferation, survival, growth and metabolism.

The first description of N-terminal extension of PTEN came from a systematic analysis of the extent of non-AUG initiation in human [23]. Ivanov et al. (2011) proposed that PTEN had a 173-aa extension due to translation initiation from a CUG codon located 519 nucleotides upstream of the canonical AUG start [23]. The extended form was supported experimentally and termed PTEN-L (PTEN-Long or PTEN-a) [24–26]. The PTEN family was further expanded by another N-terminal extended proteoform that is initiated by an in-frame AUU codon (PTEN-M or PTEN-b), adding 146 aa [26,27]. Taken together, N-terminally extended proteoforms of PTEN have been shown to have altered localization, substrates and interaction partners [25,27,28].

More recent work suggests that PTEN translational variants can modify gene expression by promoting histone methylation, opposing PTEN's traditional role as a tumour suppressor [29]. Through a direct interaction with WD40 repeat-containing protein 5, WDR5, the extended PTEN proteoforms can modulate the efficiency of histone H3K4 methyltransferases, which in turn, upregulates expression of target genes such as *Notch3* [29]. Binding studies and a crystal structure (PDB:8X3S) showed that the N-terminal extension of PTEN-a contains a WDR5-interacting motif (WIN; aa116–148) [30], wherein mutagenesis of the key residues effectively reduced the protein's pro-proliferative effect in different tumour models [30]. Besides conferring new functionality to PTEN, the N-terminal extension of PTEN is also responsible for altered protein stability due to interactions with a substrate recognition component of the SCF (SKP1-CUL1-F-box protein) E3 ubiquitin-protein ligase complex, mediating ubiquitination [29].

Cancer-associated genes can have conflicting roles as oncogenes and tumour suppressors, depending on the specific context and cancer type. An example is Wilms' tumour, WT1, a transcription factor that has an important role in developmental processes and cell survival [31]. *WT1* was first identified as a tumour suppressor gene in Wilms' tumours [32] but subsequent work showed that overexpression or mutation of WT1 contributes to tumorigenesis of some leukaemias and solid tumours [33,34]. For many years, how WT1 both promotes and suppresses cancers remained elusive [35]. However, recent work indicates that WT1 with a CUG-initiated, 68-aa N-terminal extension is the oncogenic form of the protein, cugWT1 [36,37]. Lee et al. found that cugWT1 extension was phosphorylated at S62 by AKT, leading to increased stability of WT1 and increased expression of its cancer-promoting target genes [36]. Subsequent work supported an oncogenic role of cugWT1 in mouse models of colorectal and lung cancers [37].

All cells rely on growth factors to mediate cell proliferation, differentiation, survival and migration, and dysregulation of growth factor signalling often contributes to cancer [38]. N-terminally extended proteoforms also impact growth factor signalling as illustrated by fibroblast growth factor 2 (FGF2, also known as basic FGF or bFGF) and vascular endothelial growth factor (VEGF). For FGF2, four non-AUG initiated forms of the protein were discovered, resulting in proteins of 22, 22.5, 24 and 34kDa [39–41]. Within the N-terminally extended versions, a conserved glycine-arginine repeat motif with several methylated arginine residues has been found to promote nuclear transport and/or retention [42–44].

Similar to FGF2, VEGF also has several extended proteoforms that can be generated from initiation at upstream, in-frame CUG codons [45,46], with the longest proteoform having a 180-aa N-terminal extension (VEGF-L). VEGF-L is subject to proteolytic processing [47], and in recent work, Katsman *et al.* (2022) provide evidence that the resulting proteolytic fragment, N-VEGF, translocates to the nucleus to participate in transcriptional regulation of angiogenic genes, including *VEGF*, and key genes associated with cell survival under hypoxic conditions [48].

The development of ribosome profiling techniques has allowed researchers to discover novel translated regions in

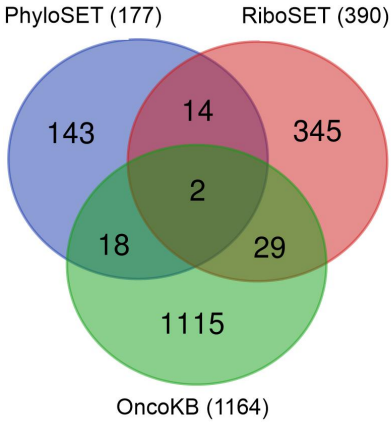
the human genome [49,50]. Recently, Fedorova and colleagues performed global analyses of Ribo-seq data and phylogenetic conservation, uncovering thousands of novel N-terminal protein extensions encoded by human mRNAs [51]. We reasoned that some of these novel N-terminal extensions may be involved in cancer. Our study explores N-terminal extensions of cancer-related proteins and proposes different ways in which they may provide differential intracellular localization, stability, and interactions.

## Results

Fedorova et al. [51] performed two types of analysis. First, they generated a set of genes with translated N-terminal extensions called RiboSET using aggregated Riboseq data. Second, they carried out an analysis of the evolutionary conservation of N-terminal extensions to show evidence of protein coding evolution. Genes with these evolutionary conserved N-terminal extensions were included in PhyloSET. We decided to find which genes from RiboSET and PhyloSET were known in relation to cancer. Analysis of OncoKB [52,53] yielded 49 cancer-related genes with potential N-terminal extensions (Figure 1). Manual annotation of the translated N-terminal extensions of 18 PhyloSET-derived genes using RiboCrypt allowed us to predict translated N-terminal extensions in 9 cases (see Methods). Therefore, our computational analysis was performed on 40 cancer-associated genes with experimentally supported N-terminal extensions (Figure 1; Supplementary Table S1).

We have considered several different scenarios regarding how N-terminal extension can influence protein availability, activity, or novel function. First, N-terminal extensions can affect proteoform stability. To investigate this possibility, we have compared the predicted stabilities of proteoform pairs using the Degronopedia resource [54]. Second, the N-terminal extension can either add a signal sequence or interfere with an existing signal sequence to alter protein targeting. This was explored using SignalP 6.0 [55]. Third, the N-terminal extension could potentially alter the folding of a proteoform, which can influence or even completely change its activity. For selected cases, we used AF3 [56] to predict the structure of N-terminally extended proteoforms. Finally, the N-terminal extension can obtain short linear motifs (SLiMs/ELMs) that can add novel binding partners, sites for posttranslational modifications, and localization signals. To do this, we explored N-terminally extended proteoforms using the ELM database, which is a resource containing a collection of annotated eukaryotic linear motifs [57].

When investigating the altered stabilities of proteoforms, we consider two possible scenarios. If an N-terminally extended proteoform is more stable than an AUG-initiated proteoform, then increasing initiation at a non-AUG codon could significantly increase the level of the protein. Conversely, if an N-terminal extension makes a proteoform less stable, then increasing non-AUG initiation would rapidly deplete protein levels. Our analysis using Degronopedia shows that for 6 proteins (UBE2A, LPP, SND1, CARM1, TCF7L2 and SOX2) the extended form is considerably less stable (at least 1 unit on the predicted protein stability index (PSI) scale) than the annotated form. Among these, the largest reduction in stability is observed for the TCF7L2/TCF-4 tumour suppressor (more than 2 PSI units) that regulates the WNT signalling pathway and transactivates downstream target genes involved in the progression of colorectal cancer. It is directly involved in regulating the expression of the oncogene MYC and inducing epithelial-mesenchymal transition (EMT) [58–61]. At the same time, there are two proteins, FSTL1 and PRSS8, for which the extended form is considerably more stable (at least 1 unit on the PSI scale) than the annotated form (Table S1). In the case of 12/40 cancer-associated proteins the N-terminally extended form is also predicted to alter the cleavage status of the initiator methionine. For 8 proteins the Met is predicted to be cleaved in the normal form, while not cleaved in the extended form, while for 4 proteins it is the other way around (Table S1).




| Intersection | Genes in intersection |
|---|---|
| PhyloSET & RiboSET & OncoKB | RPTOR, SFPQ |
| PhyloSET & RiboSET | FAM102B, CLK2, PTMS, USP19, CCDC8, HNRNPA0, FAM160B1, CYTH2, MARCKS, FXR2, BRD7, H1FX, HDGF, PIM2 |
| PhyloSET & OncoKB | TRIM33, POU2AF1, ARID3A, DCUN1D1, SPEN, MAF, UBE2A, MIB1, ASXL1, CHD4, ARID3C, SETD1B, FAF1, FGFR2, SOX10, FGF14, FBXW7, MSI1 |
| RiboSET & OncoKB | FSTL1, LPP, YWHAE, NPM1, SND1, EZR, BLM, ZNF384, LASP1, GNA13, PPP4R2, HNRNPA2B1, HSP90AA1, SOX2, SUFU, KLF5, PPP1CB, PMAIP1, ESCO2, KDSR, SMARCA1, TCF7L2, PRSS8, JUN, KAT7, MAD2L2, CARM1, TOP1, TRIP13 |

**Figure 1.** The Venn diagram shows the overlaps between the two datasets of genes encoding N-terminally extended proteins obtained from Fedorova et al., PhyloSET (177 genes) and RiboSET (390 genes), and the list of cancer-associated genes from OncoKB (1164 genes). On the right, genes belonging to the different intersections of the Venn diagram are listed. Cancer-associated genes with experimentally supported N-terminal protein extensions that are in the focus of this study are listed in black, while genes that were excluded due to not being annotated in OncoKB (intersection of PhyloSET & RiboSET) or showing no sign of translation upstream of the annotated start codon by RiboCrypt (intersection of PhyloSET & OncoKB, second half of gene list) are listed in grey.

Next, we investigated whether N-terminal extensions could contribute to signal sequence recognition. Although none of the extensions added novel signal sequences, in two genes, N-terminal extensions may interfere with protein secretion and translocation. For FSTL1, the N-terminal extension significantly decreased the probability of protein targeting (0.5716 vs 0.9998). In contrast, for PRSS8, the N terminus was predicted to preclude targeting. PRSS8 has been described as a potent tumour suppressor in colorectal carcinogenesis and metastasis [62] and is normally secreted to the extracellular space to be part of the seminal fluid. Increased expression of the N-terminally extended, likely mis-targeted proteoform, at the expense of the canonical one, could reduce the availability of PRSS8 in the seminal fluid and thereby contribute to cancer progression.

As a next step, we selected cases for structural modelling by AF3 [56] to identify structural changes introduced by the extensions. We were primarily interested in the cases where the extension could interfere with homo-oligomerization or heterodimerization, so we selected a subset of the 40 proteins with such information in UniProt [63](see Table S1 column U). While there are likely structural changes introduced by the N-terminal extensions, in our tested cases, the AF3 structural predictions did not show consistent changes between the five models produced for WT and five models representing extended forms. Often the five models obtained for the WT form already showed large structural variability (based on visual inspection in ChimeraX), which precluded the detection of structural changes caused by the extensions. Despite considerable effort put into this part of the analysis, we found the data to be too inconclusive to report any results.

Finally, we explored how SLiMs in N-terminal extensions could affect the function of proteins. The extended proteoforms were searched for linear motif candidates using the ELM database as a web server. The resulting hits were filtered for motifs overlapping with the extensions (cleavage motifs and overly redundant motifs with a probability score $> 0.015$ were removed; see candidate extension motifs per protein in Table S1). Although this search resulted in a plethora of candidate motifs, most of which are likely not true, functional linear motifs, we believe that careful analysis of such information on a case-by-case basis including functional profiling, literature mining and consideration of the known interaction partners of the investigated proteins could result in the discovery of new protein functional modules. Therefore, we decided to conduct targeted analyses of selected candidates and present our results in the next section of the manuscript.

## SFPQ

Splicing factor, proline- and glutamine-rich, SFPQ (Uniprot P23246, also known as PSF; 707 aa) is a predominantly nuclear matrix-associated, nucleic acid-binding protein that is one member of the *Drosophila behavior/human splicing* (DBHS) family of proteins [64]. Although the protein was first identified through its participation in messenger RNA (mRNA) splicing as a component of the spliceosome and U4/U6.U5 small nuclear ribonucleoprotein (tri-snRNP) complexes [65–67], it is clear that SFPQ mediates many nuclear events in addition to splicing [64].

SFPQ has an N-terminal disordered region that is enriched with proline and glutamine residues, a DNA-binding domain (DBD), two RNA-recognition motifs (RRM1 and RRM2), and a protein interaction domain (NonA/paraspeckle (NOPS)) that includes an extended coiled-coil subdomain that is important for dimerization, as shown in the resolved crystal structure (PDB: 4WII [68]). SFPQ has a nuclear localization sequence (NLS) at its C-terminus (aa701–707); however, there are reports that the protein has non-nuclear roles in the cytoplasm [69,70].

One of SFPQ's main protein-binding partners is non-POU domain-containing octamer-binding protein, NONO, a multifunctional, RNA-binding protein that is also a DBHS family member [71,72]. SFPQ and NONO can be arranged as a heterodimer and a crystal structure has been resolved (PDB: 7PU5 [73]). SFPQ and NONO can also form a ternary complex with topoisomerase I, TOP1 (see below) that stimulates the enzyme's activity [74,75].

Given a wide range of cellular activities, it is perhaps unsurprising that SFPQ has been implicated in several human diseases, including neurological disorders and cancer [76]. Knockdown of SFPQ in colorectal cancer cells was shown to enhance apoptosis [77], and multiple reports link SFPQ to altered transcriptional, transport and splicing profiles of mRNAs involved in cancer progression and drug resistance [78–80]. Still the most direct involvement in cancer comes from Xp11.2 translocation renal cell carcinoma (XP11.2 tRCC) in which *SFPQ* is the fusion partner of MiT (microphthalmia transcription factor) family member gene, transcription factor binding to IGHM enhancer 3 (*TFE3*) [81]. The *SFPQ/TFE3* gene fusion is also present in perivascular epithelioid cell tumours (PEComas) and melanotic Xp11 translocation renal cancers [82].

The N-terminal extension of SFPQ is supported by strong ribosome footprint densities within the region upstream of the annotated start codon (Figure 2(A)). To search for proteomic evidence of the N-terminal extension of SFPQ, we implemented targeted peptide search engine PepQuery2 [83]. The search in 48 MS/MS datasets available in the web version (https://pepquery2.pepquery.org/) yielded 71 confident peptide spectrum matchings (PSMs) corresponding to 3 peptides: FCLDRPLTTDMSR (64×), MASTFPER (3×), MASTFPERLLR (4×) (Supplementary Table S2). In the 21-residue N-terminal extension of SFPQ we could identify two unique motifs that likely confer new protein–protein interactions (Figure 2(B)). The first is DOC_CYCLIN_RxL_1 (aa5--15, FPERLLRFCLD), a cyclin N-terminal domain docking motif. Work by Rayner et al. (2021), using a proximity ligation method (*Bio*tin *Id*entification (BioID)), immunoprecipitations and mass spectrometry, identified novel interaction partners of cyclin F, including SFPQ [84]. This interaction was further supported by high-throughput data [85].

Although not examined in a cancer context, an interaction between cyclin F and SFPQ is of clinical relevance to familial and sporadic amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD). In 2016, mutations of the cyclin F gene, *CCNF*, were discovered in ALS/FTD, and *CCNF*

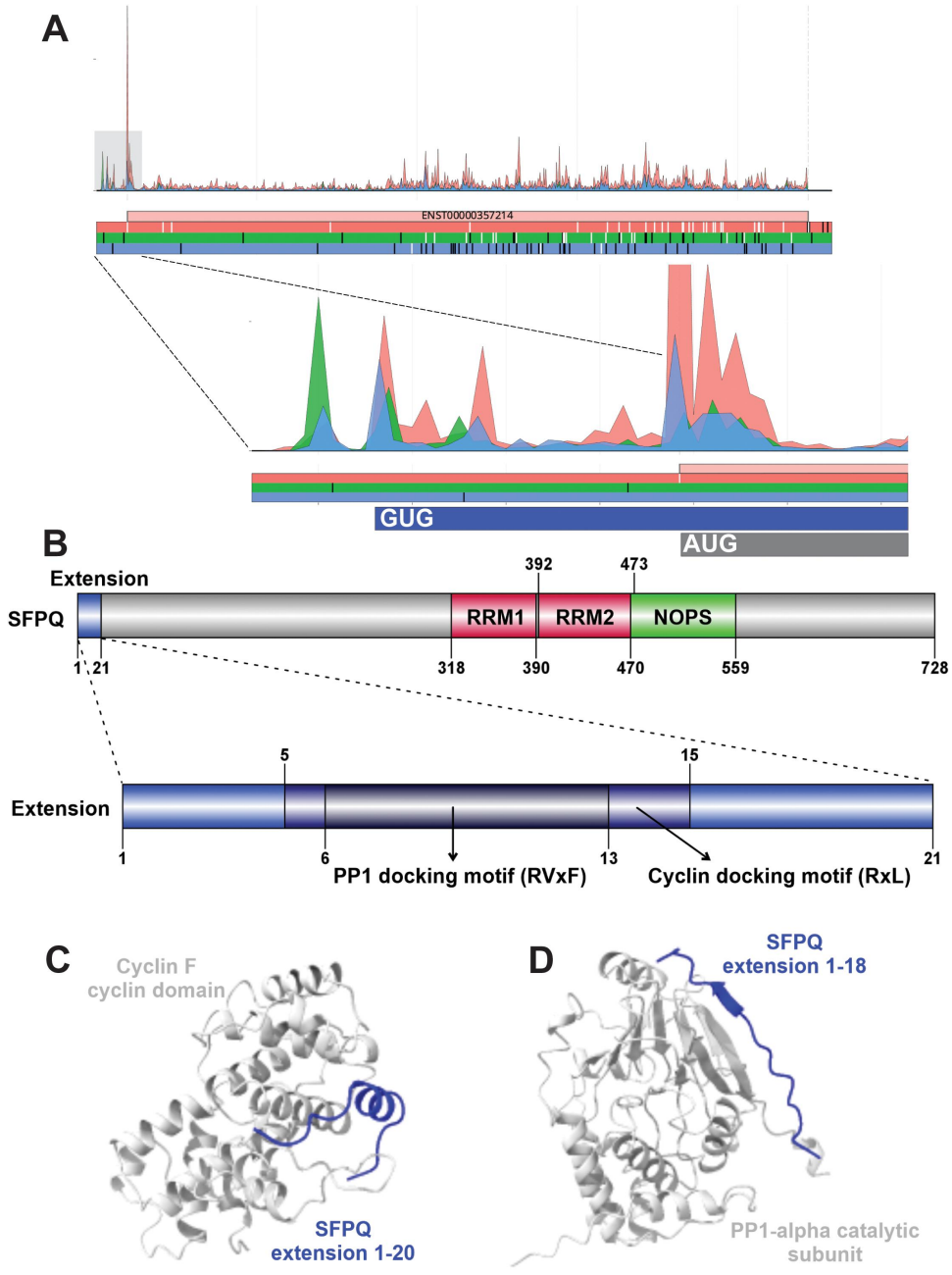



**Figure 2.** (A) Aggregated Riboseq data for *SFPQ* (ENST00000357214) from the Ribocrypt transcriptome browser (ribocrypt.Org, in preparation). Ribosome footprints are colour-coded according to the translated reading frames shown below the diagram. (B) Domain map of the N-terminally extended SFPQ protein is depicted. The 21 aa extension (blue) and the 707 aa canonical protein are shown, with known domains of the protein depicted (residue boundaries provided with respect to the extended proteoform). Below, the extension is depicted separately with the two identified, likely functional short linear motifs marked by dark blue and black, respectively. (C) AlphaFold 3 (AF3) models of the proposed domain-motif complexes are provided below the domain maps in cartoon representation. Structural model of human cyclin F cyclin domain (P41002, residues 290–531) modelled with SFPQ extension peptide residues 1–20, containing the predicted cyclin-binding RxL motif. (D) Structural model of human Serine/threonine-protein phosphatase PP1-alpha catalytic subunit (P62136, residues 1–305) modelled with SFPQ extension peptide residues 1–18, containing the predicted PP1 docking RVxF motif.

variant (cyclin F: S621G) resulted in an accumulation of ubiquitinated proteins, likely due to abnormal ubiquitination or transport to the proteasome [86]. SFPQ was already linked to pathological features of neurological disorders, including ALS [87]. Following on from this, it was shown that overexpression of the cyclin F mutant S621G led to an increase in the insoluble fraction of SFPQ and disrupted its subcellular distribution [84]; thus, cyclin F mutations could lead to dysregulation of SFPQ's influence on RNA metabolism, which contributes to the pathomechanisms of ALS/FTD.

It is possible that the RxL motif in the N-terminal extension might mediate this interaction, as there is no other copy of this motif within SFPQ. Although cyclin F is more of an E3 ligase adaptor than an ordinary cyclin [88], it has cell cycle regulatory activity under certain circumstances and a cyclin domain that was shown to be functional and binding to RxL motifs (also called Cy motif) in several proteins [89–91]. Modeling of the interaction by AF3 led to a complex wherein the RxL motif of the SFPQ extension binds to the known RxL-binding groove of the Cyclin F cyclin domain in helical

conformation, similarly to RxL motif complexes with known structure (e.g. PDB: 1H24, 1H25, 1H26, 1H27, 1H28 [92]) (Figure 2(C)).

The second motif identified in the N-terminal extension is DOC_PP1_RVXF_1 (aa6–13, PERLLRFC), which overlaps with the RxL motif. The RVxF motif mediates the interaction with serine/threonine-protein phosphatase 1, PP1 [93]. Previous data provide evidence for an interaction between SFPQ and PP1. SFPQ was found to interact with PP1CA (the catalytic subunit of the PP1 phosphatase) in a low- and high-throughput study [94]. While for the close-relative NONO, the RVxF motif could be identified and was determined to mediate an interaction with PP1, it could not be found in SFPQ [94].

We suggest that the RVxF motif within the N-terminal extension of SFPQ could be functional and mediating the interaction with PP1, which could be especially important when SFPQ forms homodimers and does not form a complex with NONO. In the AF3-predicted complex the RVxF motif of the SFPQ extension binds to PP1-alpha catalytic subunit in beta-augmentation, similarly to other bound RVxF motifs with known structure (e.g. PDB: 3N5U [95]) (Figure 2(D)). This interaction could lead to dephosphorylation of SFPQ and loss of its transcriptional co-repressor activity [94].

## SPEN

The SHARP protein encoded by gene SPEN is a huge transcriptional repressor of 3664 residues that acts as a scaffold for different proteins and complexes important for expression regulation. It was identified as a component of transcriptional repression complexes in both nuclear receptor and Notch/RBP-Jkappa signalling pathways. The N-terminal ~ 600 residues contain 4 RRM domains that were reported to mediate DNA as well as RNA binding, and the protein has extended disordered regions that exceed thousand residues in length. A central RID domain was reported to bind nuclear receptors, while the C-terminal SPOC domain is known to recruit transcriptional corepressors SMRT/NCoR and histone deacetylases HDAC1/2 through binding their LSD motifs [96–98], as well as the C-terminal disordered domain of RNA polymerase II (Pol II [99]). The RBP-Jkappa/SHARP complex was shown to recruit the CtIP and CtBP corepressors to silence Notch target genes in a manner that CtIP/CtBP functionally complement SHARP in repression [100]. While CtIP binding could be mapped to the repressor domain of SHARP, CtBP binding could not be precisely mapped.

However, translation from an in-frame CUG codon in 5'leader leads to a 22 aa-long N-terminal extension (Figure 3(A)), which is supported by PepQuery2 proteomics analysis uncovering 6 confident peptide spectrum matchings (Table S2). Interestingly, this short extension of SHARP contains a CtBP-binding motif, while the annotated fraction of the giant protein does not, which makes it highly likely that the extended form plays a key role at least in the silencing of Notch genes, where CtBP binding is required (Figure 3(A,B)). Modeling of the interaction by AF3 resulted in a complex structure wherein the SHARP CtBP-binding motif binds to the same groove of the CtBP dimer with beta-augmentation as seen for known CtBP-motif complexes (e.g. PDB:1HL3 [101]) (Figure 3(C)). Thus, the extension might contribute to CtBP binding simultaneously to CtIP binding by the annotated part of the protein, and/or could enable direct binding of SHARP to CtBP without the contribution of CtIP in certain processes.

## SUFU

Suppressor of fused homolog, SUFU (Uniprot Q9UMX1; 484 aa), is a negative regulator within the evolutionarily conserved Hedgehog (HH) signalling pathway that is fundamental in embryogenesis and adult tissue homoeostasis [102]. SUFU interacts with zinc finger GLI transcription factors (GLI1, GLI2 and GLI3) in the cytoplasm [103,104], preventing their translocation to the nucleus to activate transcriptional programs. However, it was also shown that SUFU can interact with GLI1 while bound to DNA [104]. SUFU N-terminal domain (aa 64–240) binds to the C-terminal region of GLI; while SUFU_C (aa 253–473) binds to the N terminal of GLI, and interactions between SUFU and GLI1 at both sites are required for cytoplasmic tethering and repression of GLI1 [105]. The repressive effect of SUFU on GLI transcription factors can be overcome by serine-threonine kinase, Fused (serine-threonine kinase protein-36, STK36 in humans) [106].

Disruption of the HH pathway is linked to cancer development and components of the pathway are targets for anticancer therapeutics [107]. In humans, germline and somatic mutations in *SUFU*, accompanied by loss of heterozygosity (LOH), results in a predisposition to medulloblastoma [108], a type of brain cancer most often affecting children. Subsequent work in mouse models showed that *Sufu* $^{\pm}$ mice crossed within a *p53* null background led to a stark increase in medulloblastoma in the animals [109]. SUFU has also been implicated in other cancers, including basal cell carcinoma [110]. In somatic cells, missense mutations across the protein are most common, with 27.94% of 1124 unique samples containing *SUFU* missense mutations recorded in COSMIC (cancer.sanger.ac.uk) [111].

SUFU is normally targeted for degradation by the Skp1-Cul1-F-box protein complex (SCF) through its polyubiquitination by E3 ubiquitin ligase, FBXL17 (F-box and leucine-rich repeat protein 17), while in complex with GLI1 [112]. Although FBXL17 directs downregulation of SUFU in the nucleus [112], it does not seem to do this in a cell cycle-dependent fashion, due to the lack of cyclin-dependent kinase (CDK) phosphorylation sites in SUFU. More recent work has shown that SUFU can operate outside of its role in HH signalling by negatively regulating initiation of centrosome duplication and DNA replication at the G1-S transition of the cell cycle [113].

Analysis of Riboseq datasets allowed the detection of translation of CUG-initiated N-terminal extension (Figure 4(A)), which is also supported by PeptideQuery2-based identification of 5 confident PSMs (Table S2). Based on our subsequent motif analysis, the N-terminal extended form of SUFU could be under the control of the cell cycle (Figure 4(B)), as suggested by the joint presence of several motifs implicated in granting cell cycle phase-dependent availability. Efficient

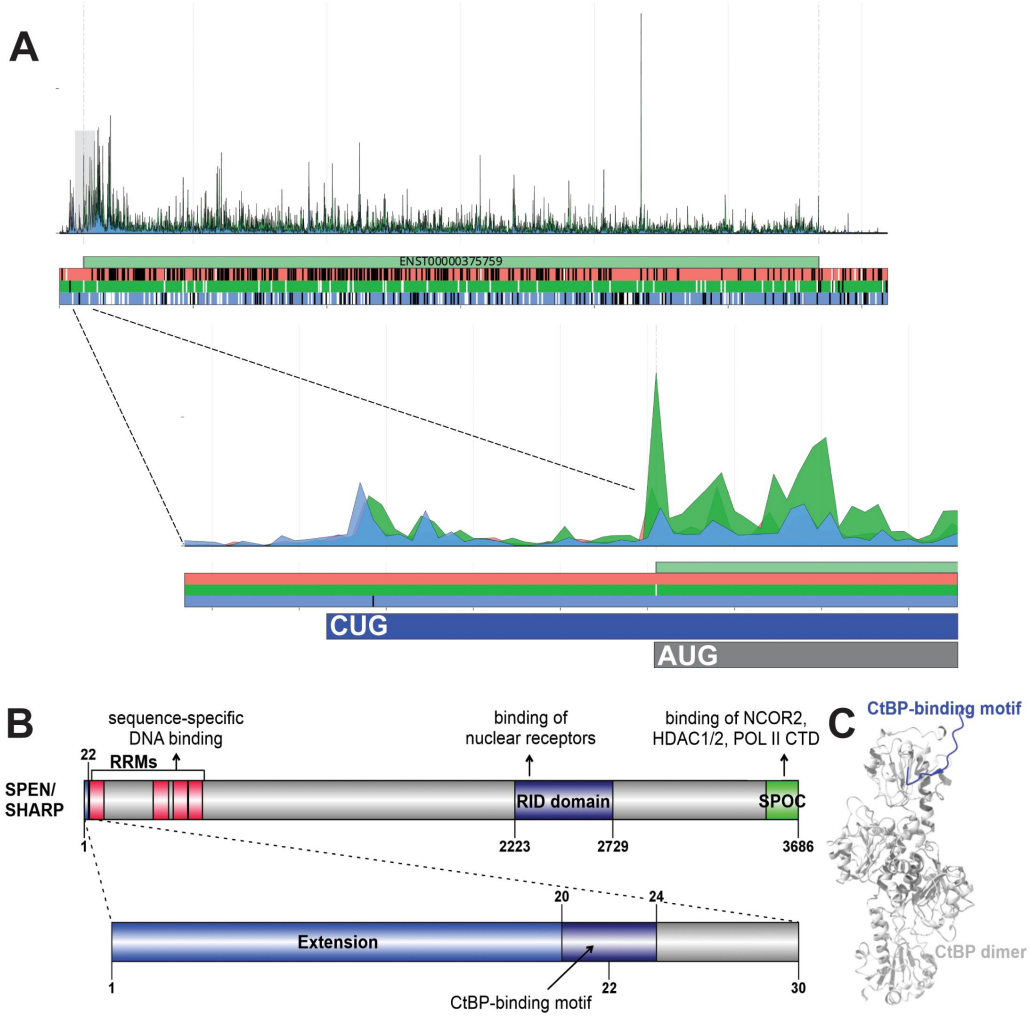



**Figure 3.** (A) Aggregated Riboseq data for *SPEN* (ENST00000375759) from the Ribocrypt transcriptome browser (ribocrypt.Org, in preparation). Ribosome footprints are colour-coded according to the translated reading frames shown below the diagram. (B) Domain map of the N-terminally extended SPEN/SHARP protein is depicted. The 22 aa extension (blue) and the 3664 aa canonical protein are shown, with known domains of the protein depicted (residue boundaries provided with respect to the extended proteoform). Below, the extension is depicted separately with the identified, likely functional CtBP-binding motif marked by dark blue. (C) the AF3 model of the proposed domain-motif complex between human CtBP1 dimer (2× P41002, residues 28–370) and the SHARP extension peptide residues 15–30 (containing the predicted CtBP-binding motif) is provided on the right of the domain maps in cartoon representation.

phosphorylation by CDKs at motifs (MOD_CDK_SPK_2, MOD_CDK_SPxK_1) could be largely fostered by the observed cyclin-dependent kinases regulatory subunit 1 (CKS1) docking motif (DOC_CKS1_1). The sequence pattern of the CDK phosphorylation site also matches the target sequence pattern required for dephosphorylation by the CDK-antagonistic phosphatase, CDC14 (MOD_CDC14_SPxK_1). Furthermore, on the border of the extension and the canonical protein sequence, a DEG_SCF_FBW7_2 motif is also formed. The SCF ubiquitin ligase substrate recognition subunit F-box/WD repeat-containing protein 7, FBXW7, is a well-known tumour suppressor, initiating the ubiquitination and subsequent degradation of many cell cycle-regulated proteins, including oncogenes such as *CCNE1*(cyclin E), *MYC* and *RICTOR* [114–116]. In the AF3-predicted model of the SUFU-FBXW7 interaction, the SUFU motif is placed onto the top of the doughnut-shaped WD40 domain of FBXW7 in a similar conformation as seen for other FBXW7 substrates (e.g. PDB:2OVQ [117]) (Figure 4(C)). Modeling of the potential SUFU-CKS1 interaction resulted in a complex, where the clam-shaped CKS1 binds the SUFU extension peptide (phosphorylated at T22) in a similar conformation as seen for other CKS1-phosphopeptide complexes (e.g. PDB:2AST [118]) ((Figure 4(D)). Presence of all these motifs conferring cell-cycle dependent control within the relatively short, 23-residue disordered extension of SUFU can be considered as a solid support for the existence of a cell cycle-regulated SUFU proteoform.

## SOX2

One example of N-terminal extension is found in *SOX2*. The high-mobility group (HMG) box transcription factor, SOX2, is a master regulator of stemness and pluripotency. By promoting oncogenic signalling and maintaining cancer stem cells, SOX2 is implicated in the development of several different cancer types, including breast [119], prostate [120,121], pancreatic [122], gastric [123], lung [124] and cervical cancers

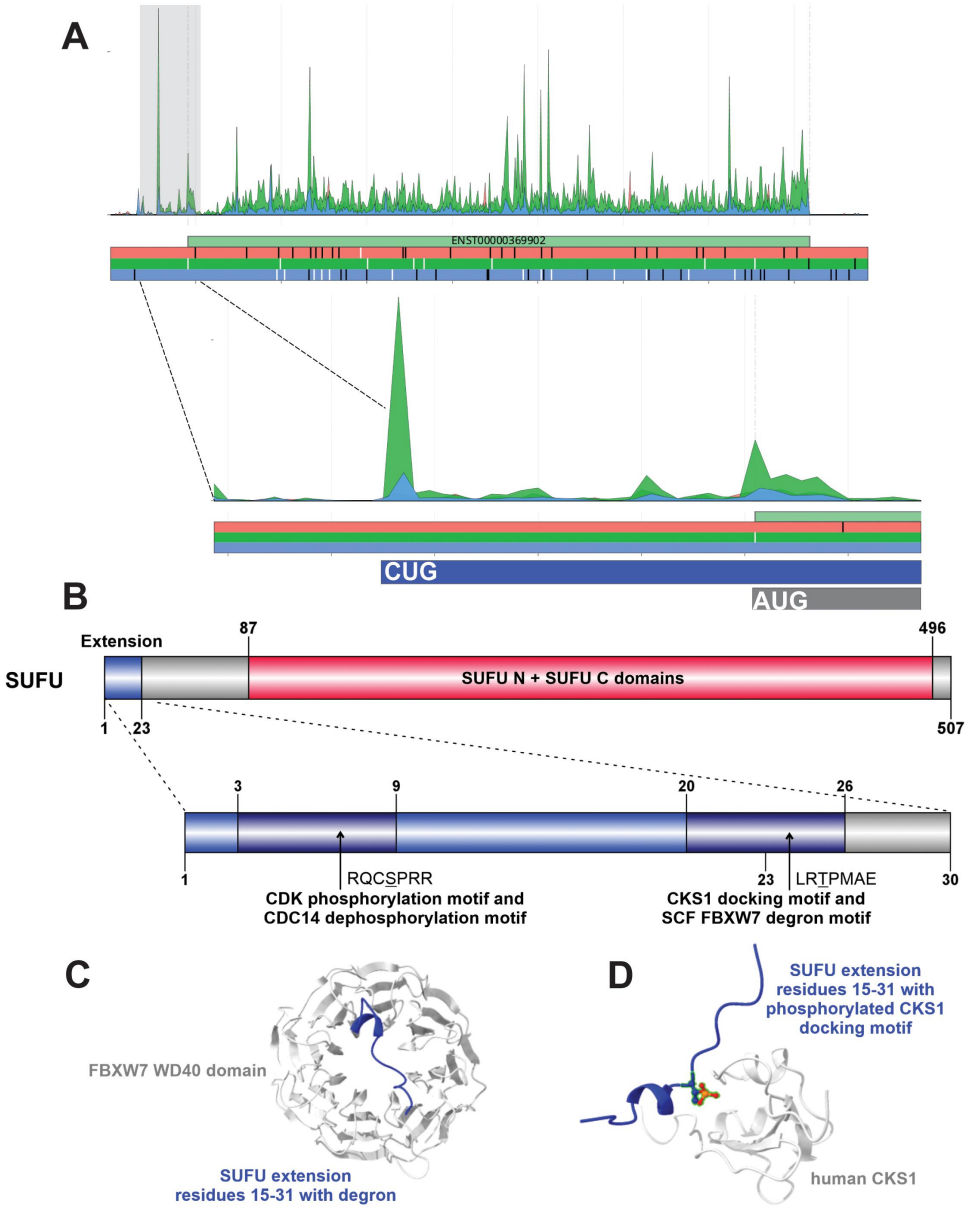



**Figure 4.** (A) Aggregated Riboseq data for *SUFU* (ENST00000369902) from the Ribocrypt transcriptome browser (ribocrypt.Org, in preparation). Ribosome footprints are colour-coded according to the translated reading frames shown below the diagram. (B) Domain map of the N-terminally extended SUFU protein is depicted. The 23 aa extension (blue) and the 484 aa canonical protein are shown, with known domains of the protein depicted (residue boundaries provided with respect to the extended proteoform). Below, the extension is depicted separately with the four identified, likely functional short linear motifs mapped onto two regions (in dark blue), respectively. (C) AF3 models of two of the proposed domain-motif complexes are provided below the domain maps in cartoon representation. Structural model of human FBXW7 WD40 domain (Q969H0, residues 370–707) bound to SUFU extension peptide residues 15–31, containing the predicted degron in a phosphorylated form (p22T). (D) Structural model of human CKS1 (P61024, residues 1–79) modelled with SUFU extension peptide residues 15–31, containing the predicted CKS1 docking motif in a phosphorylated form (p22T). The phosphorylation is highlighted in stick representation. For the identified modification-type motifs, complexes could not be modelled due to the transient nature of the interactions and lack of available experimentally determined complex structures.

[123]. SOX2 was proposed to promote metastasis [125–128], and its amplification was observed in several cancer types [125,129–135].

Aggregated Riboseq data for *SOX2* is consistent with translation of an N-terminal extension from the UUG codon located 288nts upstream of the annotated AUG start codon (Figure 5(A)). PepQuery2 analysis [83] allowed identification of as many as 90 peptide spectrum matches, corresponding to several unique peptides including AGPAHSAR (23X), MITIIGGGR (18X), MITIIGGGRIGQR (22X) and LPSSSPPAR (11X) (Table S2). Identification of semi-tryptic peptides started with methionine (MITI …) supports the assumption that the UUG codon of SOX2 N-terminal extension is decoded with initiator met-tRNAi.

The sequence of the predicted N-terminal extension was analysed with the eukaryotic linear motif (ELM) resource [57] and yielded 36 motifs that met our criteria (see Methods and Table S1). Interestingly, one of such motifs is a putative reverse nuclear export signal (NES; ELM class: TRG_NESrev_CRM1_2) located in the middle of the N-terminal extension (Figure 5(B)). While as a transcription factor, SOX2 is expected to operate only in nuclei, a cytosolic

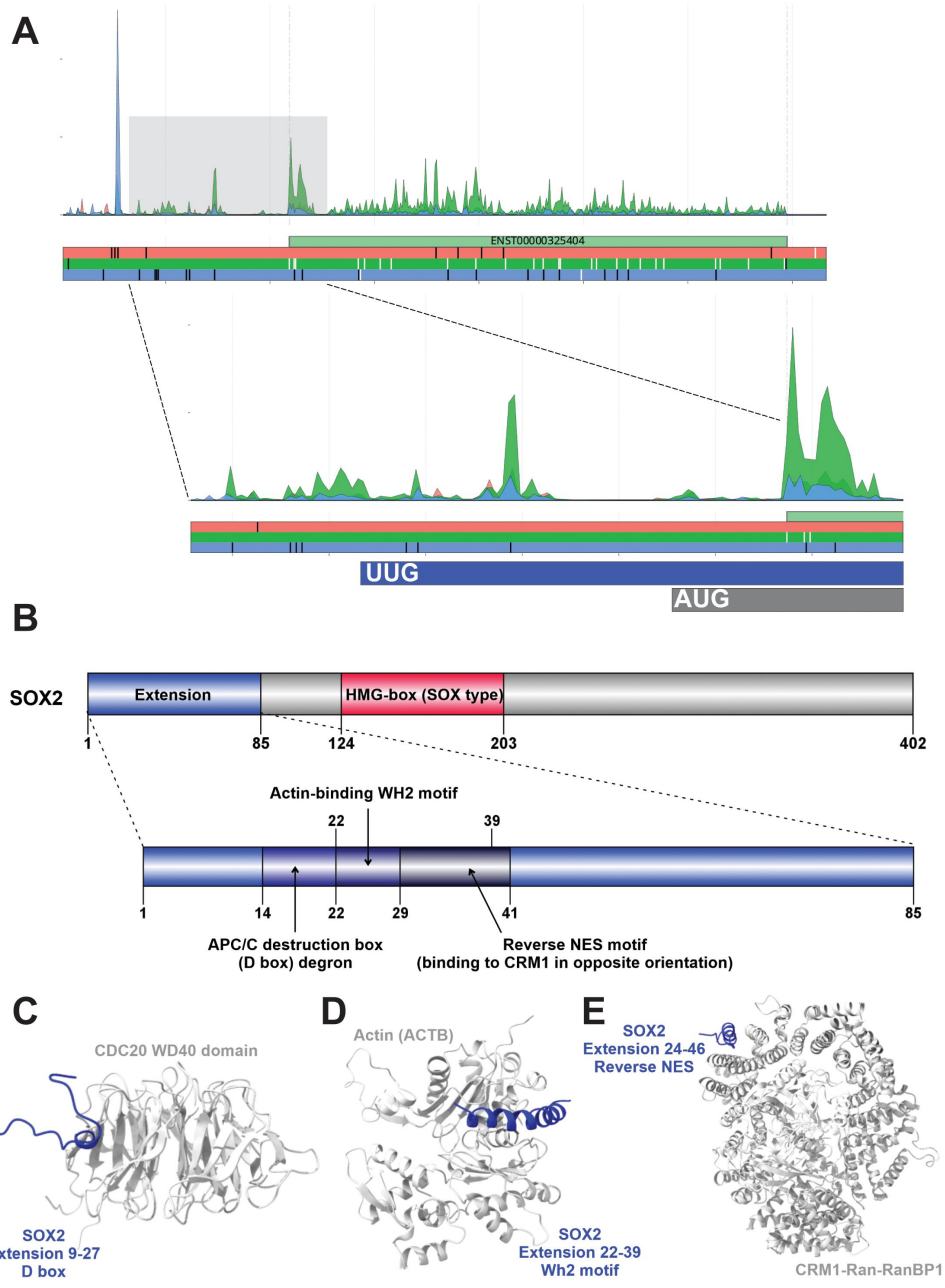



**Figure 5.** (A) Aggregated Riboseq data for *SOX2* (ENST00000325404) from the Ribocrypt transcriptome browser (ribocrypt.Org, in preparation). Ribosome footprints are colour-coded according to the translated reading frames shown below the diagram. (B) Domain map of the N-terminally extended SOX2 protein is depicted. The 85 aa extension (blue) and the 317 aa canonical protein are shown, with known domains of the protein depicted (residue boundaries provided with respect to the extended proteoform). Below, the extension is depicted separately with the three identified, likely functional short linear motifs marked by dark blue and black, respectively. (C) AlphaFold 3 (AF3) models of the proposed domain-motif complexes are provided below the domain maps in cartoon representation. Structural model of human CDC20 WD40 domain (Q12834, residues 160–477) modelled with SOX2 extension peptide residues 9–27, containing the predicted APC/C D box motif. (D) Structural model of human cytoplasmic actin (P60709, residues 1–375) modelled with SOX2 extension peptide residues 17–45, containing the predicted actin-binding WH2 motif. (E) structural model of a CRM1-ran-RanBP complex (chains A, B and C of the PDB:5DIF structure were used) modelled with SOX2 extension peptide residues 24–46, containing the predicted reverse NES motif.

pool of SOX2 also exists [136,137]. Cytosolic SOX2 was reported to interact with ribosomes and to regulate the translation of mRNAs that code for proteins implicated in sugar metabolism, matrix contacts, and development [137]. While the annotated sequence of SOX2 has a known NES that needs to get acetylated (at K75, probably by CBP/p300) for mediating the nuclear export of SOX2 [138], the extended isoform could potentially ensure a cytoplasmic pool of SOX2 independent of acetylation. Thus, decreased stringency of start codon selection has the potential to change the ratio of cytoplasmic and nuclear SOX2 to reprogram gene expression.

Additionally, the 85 residues long extension of SOX2 also contains a predicted Anaphase promoting complex (APC/C) D box degron motif (Figure 5(B)). This is likely functional, as one of the substrate recognition/activator subunits of the APC/C that can recognize this motif, CDH1/FZR1, is

a SOX2 binding partner [139], and because for CDC20 (the other D box-binding activator subunit), a CDC20-APC/SOX2 signalling axis has been proposed to control some key biological properties of glioblastoma stem cells [140]. Since other binding modules enabling binding to the APC/C do not seem to be present in the annotated SOX2 sequence, it is likely that the D box motif of the extension mediates these interactions.

Furthermore, the extension contains a predicted actin-binding WH2 motif (Figure 5(B)), which could also be functional. We have analysed previously performed pull-down assays of SOX2 from the cell lysates of three different cell types [137,140] and found that SOX2 bound cytoplasmic actin (ACTB) in all three cell types along with some proteins involved in the regulation of actin filament assembly, such as cofilin and actin capping protein. The presence of this rather complex, low-probability actin-binding motif and consistently detected SOX2–actin interactions suggest a hitherto undiscovered function of cytoplasmic SOX2 in binding to or regulating the actin cytoskeleton.

We have performed AF3 predictions of the three identified SOX2 extension motifs with the respective motif-binding domains (CDC20 WD40 domain for the D box, Actin for the WH2 motif and the CRM1-Ran-RanBP1 complex for the reverse NES), which resulted in complex structures (Figure 5(C,D,E)) where the SOX2 motifs were bound to the right surface patch/binding groove of the interacting domains with respect to the experimentally determined structures of the same motif-domain interactions (e.g. PDB:4UI9,8A2T [141,142] and PDB:4BH6 [143]; PDB:5YPU [144] and PDB:5DIF [145] for the three motifs, respectively).

## TOP1

Analysis of Topoisomerase I, TOP1 allowed identification of CUG-initiated N-terminal extension (Figure 6(A)) which is supported by 30 PSMs identified with PepQuery2 (Table S2). TOP1 is an enzyme that can relax positive and negative supercoiling as well as torsional tension of the DNA induced by DNA replication and transcription. While performing this task, TOP1 creates single strand breaks (SSBs) that allow relaxation of the torsional tension of DNA and forms covalent DNA-protein crosslinks (DPCs), specifically TOP1 cleavage complexes (TOP1ccs) [147]. DPCs represent a double-edged sword as they are necessary for genome maintenance, but at the same time, when they accumulate due to deregulated removal, they can be detrimental for the cell due to blocking replication and transcription, as well as other processes involving DNA. TOP1ccs are among the most frequently occurring DPCs and therefore even anticancer drugs were developed that aim to kill cancer cells by trapping TOP1ccs [147].

The repair of TOP1ccs is regulated on one hand by the PARylation-dependent TDP1 pathway, wherein TDP1 is capable to cleave the covalent bond between TOP1 and DNA [148], and the ubiquitylation-dependent proteasome pathway that mediated TOP1 degradation [149]. PARylation of the TOP1ccs was demonstrated to enhance USP7-mediated deubiquitination [150], indicating that USP7 can probably reverse RNF4-induced ubiquitylation of TOP1-DPCs [150,151].

The annotated, N-terminal disordered region of TOP1 contains 20 predicted USP7 ubiquitin-like 2 (UBL2) domain docking sites, wherein many lysins involved in the predicted docking motifs (described as KxxxK) are known sumoylation sites of the SUMO ligase PIAS4. When sumoylated, these lysins are masked from UBL2 domain binding and sumoylation was demonstrated to recruit the SUMO-targeted ubiquitin ligase RNF4, which ubiquitinates the TOP-DPCs for subsequent proteasomal degradation [151]. While the annotated fraction of USP7 does not contain predicted docking motifs for the main substrate recognition domain, the MATH domain of USP7, the identified N-terminal extension contains 3 copies of MATH domain docking motifs (2 copies of the type 1 (DOC_USP7_MATH_1) and one copy of the type 2 motif (DOC_USP7_MATH_2), which were reported to bind the same binding groove of the USP7 MATH domain [152]), and thus it is expected to mediate multivalent interactions with USP7 (Figure 6(B)). Therefore, our premise is that the extended TOP1 proteoform could be more efficiently deubiquitinated by USP7 than the normal proteoform.

## Discussion

Differential expression of N-terminally extended proteoforms in cancer cells can be achieved through alterations in certain initiation factors that are involved in start codon stringency. Experiments using reporter constructs and genome-wide CRISPRi screens have shown that dozens of canonical and non-canonical initiation factors can alter initiation at suboptimal start codons in mammals. These 'stringency' factors include eIF1, eIF1A, eIF5, BZW1, BZW2, eIF4G2 and eIF3 [19,153–158]. Certain somatic mutations in these 'stringency' factors and regulators can lead to altered non-AUG initiation in cancer cells. For example, mutations in the unstructured N-terminal tail (NTT) of *EIF1A* have been associated with uveal melanoma [159], and corresponding mutations in yeast eIF1A's NTT can suppress initiation at non-AUG codons [160]. Additionally, several somatic mutations in the *EIF4G2* gene have been identified in primary tumours from cancer patients. Some of these mutations affect the binding of the eIF4G2 protein to interacting proteins and its ability to direct mRNA translation [161].

However, as 'stringency' factors in general do not seem to be frequently mutated in cancer, other mechanisms may contribute to the altered usage of non-optimal start codons. One possible explanation is the differential protein stability of initiation factors such as eIF1 and eIF5. It has been proposed that the half-life of these proteins differs, and therefore, changes in global protein synthesis or protein degradation can alter their ratio which will result in alterations on start codon selection [162]. Changes in 'stringency factors' can also be achieved through modulation of their intracellular localization. Recent study has shown that an increase in the stringency of start-codon selection during mammalian mitosis is mediated by the release of nuclear eIF1 after nuclear envelope breakdown [163]. In addition, post-translational modifications may affect eIF activity. For example, eIF5 can be phosphorylated by casein kinase 2 (CK2), which alters its association with other eIFs. This PTM affects translation

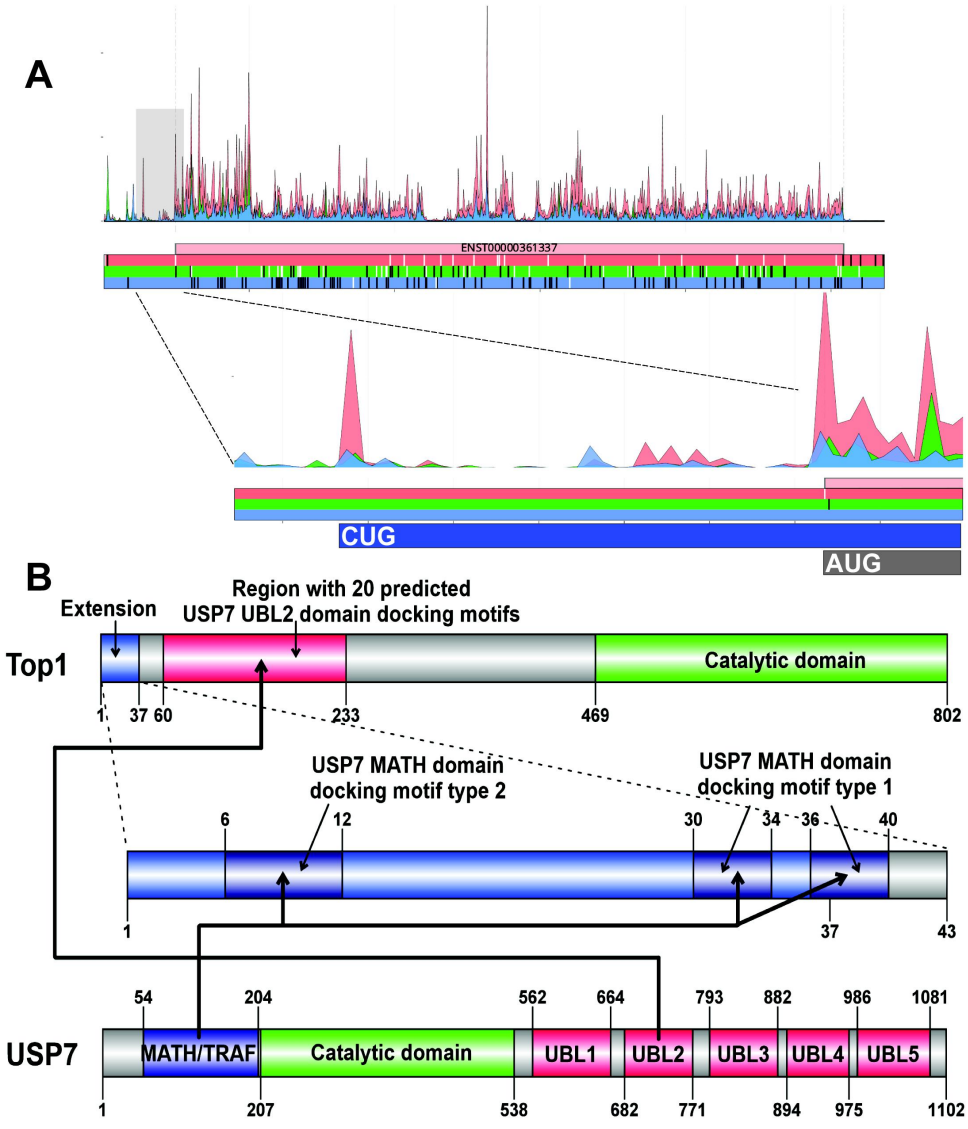



**Figure 6.** A) Aggregated Riboseq data for *TOP1* (ENST00000361337) from the Ribocrypt transcriptome browser (ribocrypt.Org, in preparation). Ribosome footprints are colour-coded according to the translated reading frames shown below the diagram. B) at the top, the domain map of the N-terminally extended TOP1 protein is depicted. The 37 aa extension (blue) and the 765 aa canonical protein are shown, with known domains of the protein depicted (residue boundaries provided with respect to the extended proteoform). Below, the extension is depicted separately with the three identified, likely functional USP7 MATH domain-binding short linear motifs marked by dark blue. At the bottom, the domain map of USP7 is shown, with the domain boundaries provided based on Figure 1 of Valles et al. [146]. Thick black arrows connect the domains of USP7 with the corresponding binding motifs in extended TOP1 that likely interact. The interactions could not be modelled by AF3, because it could not correctly predict the beta-rich domain structure of the USP7 MATH domain.

initiation [164–167]. However, whether CK2-mediated phosphorylation of eIF5 affects the stringency of start codon selection is still an open question. In summary, deregulation of protein homoeostasis, signalling cascades, and cell cycle progression in cancer cells may lead to differential translation from non-AUG codons.

Finally, it should be noted that the probability of initiation at certain non-AUG codons can critically depend on the mRNA sequences upstream and downstream. Therefore, differential translation of specific N-terminally extended proteoforms can be achieved on a gene-specific basis even without any changes to 'stringency' factors [6]. For instance, it has been shown that ribosomal pausing immediately downstream of alternative initiation sites can increase the use of upstream start codons [168]. This pausing is typically relieved by the specialized translation factor eIF5A, and its depletion increases translation from the CUG codon in *MYC* and other transcripts [169]. We expect that the putative N-terminal extensions described in our study may also be under extensive translational control in both normal and cancer cells. This has also been suggested for upstream open reading frames (ORFs) that can impact the main coding sequence and/or produce proteins that contribute to a cancer phenotype [170]. In this study we analysed recently proposed N-terminal extensions of cancer-associated proteins for possible functional readouts. Our analysis highlights that PANTs represent a further layer in protein functional versatility, similar to alternative and tissue-specific splice isoforms [171–173]. We found that the additional sequence regions harbour short linear motifs that confer novel functionalities and regulatory possibilities on the extended proteoforms compared to the canonical ones.

Recent work by Bogaert et al. [174] generated a catalogue of N-terminal extended proteoforms from the cytosol of HEK293T cells and selected 22 N-terminal/canonical protein pairs for interaction mapping using a high-throughput workflow (Virotrap [175]) and yeast two-hybrid screening. From those pairs, three were selected for further experimental validation using affinity purification-mass spectrometry (AP-MS). In agreement with our work, the N-terminal extended proteoforms identified by Bogaert et al. had unique interaction partners when compared to the canonical protein, highlighting the functional diversity created by the presence of the extended proteoforms.

We successfully discovered some very interesting cases where N-terminal extensions (supported by both ribosome footprinting and proteomics evidence) could confer different localization, turnover rates or interactions on proteins classified oncogenic and tumour suppressive. Indeed applications of proteomic approaches are allowing the functional interpretation of proteomics data [176]. Top-down proteomics methods that do not include proteolytic digestion and employ multi-dimensional separation techniques, followed by mass spectrometry [177–179], could be used to comprehensively identify and characterize N-terminally extended proteoforms in different cancers and various cell types. Methods that focus on N-termini, such as COmbined FRActional DIagonal Chromatography (COFRADIC) [180,181], as successfully employed in the work by Bogaert et al. [174], could have expanded use across cancer cell lines. To discover and validate the interactors of N-terminal extensions, AP-MS remains one of the best ways to characterize interactomes [174]; however, it can be labour intensive. Nevertheless, we encourage the scientific community to experimentally validate the discovered interaction sites of the extended proteoforms and elucidate their role in tumorigenesis and cancer progression. The expansion of proteomic methods, in combination with systems and computational biology, will aim to address the vast diversity and dynamics of the proteome and continue to improve our understanding of human biology and disease.

## Methods

### Dataset assembly

Genes encoding proteins with non-annotated N-terminal extensions translated from non-AUG alternative translation initiation sites (TISs) were obtained from the PhyloSET and RiboSET datasets published by Fedorova et al. [51,182]. The two sets were merged to look for an overlap with the list of cancer-associated protein-coding genes downloaded from OncoKB (1164 protein-coding genes as of 8/11/2024) [52,53]. 49 cancer-associated genes could be identified within the merged set of N-terminally extended genes. Of these, 31 belong to RiboSET where the position of the alternative TIS was indicated, thus the extended proteoforms and the extension sequences could be obtained in an automated manner. For the remaining 18 genes belonging to PhyloSET, the N-terminal extension was only predicted by Fedorova *et al.* based on PhyloSCF scores and thus the alternative TISs were not indicated. For these genes, we manually annotated translated N-terminal extensions using RiboCrypt. This manual annotation allowed us to identify translated N-terminal extensions for 9 genes (*TRIM33* (with alternative initiation codon: CUG), *ARID3A* (CUG), *POU2AF1* (CUG), *SPEN* (CUG), *DCUN1D1* (CUG), *UBE2A* (CUG), *MAF* (CUG), *ASXL1* (GUG), *MIB1* (GUG)), while the other 9 genes were removed from the dataset due to the lack of convincing ribosome footprint patterns N-terminal to the annotated TIS. Therefore, we ended up with a list of 40 cancer-associated genes with experimental evidence for an N-terminal extension resulting from an alternative TIS (Table S1), which were subjected to further analysis.

### Computational investigation of the N-terminal extensions

We performed degron and stability predictions on the normal and extended proteoforms of the 40 genes using Degronopedia [54]. The number of predicted degrons, all predicted protein stability index (PSI) scores (N-terminal PSI assuming cleaved initiator Met, N-terminal PSI assuming non-cleaved initiator Met and C-terminal PSI) and the predicted cleavage status of the initiator Met (cleaved/not cleaved (C/NC)) were obtained and compared between the two proteoforms (Table S1).

Prediction of Signal Peptides and their cleavage sites was performed with SignalIP 6.0 [55] web server with the following settings: Organism – Eukarya; Output format – Long output; Model mode – Fast.

The webserver part of the ELM linear motif database [57] was used to obtain predicted short linear motifs (SLiMs) for the extended proteoforms. The predicted linear motifs of the N-terminal extensions (if at least one residue was contributed by the extension sequence to the predicted motif, it was already accepted as an extension motif to be able to identify all possible binding sites contributed or complemented by the extensions) were filtered for those with a pattern probability < 0.015 (this value is provided by the ELM database for annotated motif classes and represents how likely the motif pattern occurs in random sequences) to avoid highly degenerate motifs with excessive false positive prediction rates. For strictly nuclear genes (according to UniProt), the motif search was limited to nuclear motifs, and this was indicated with 'N!' in Table S1 column T before listing the detected motifs. The resulting, predicted extension motifs are listed for all 40 genes (Table S1). Subsequently, we performed extensive literature search and mining of protein–protein interaction databases, e.g. BioGRID [183], to identify the subset of the predicted motifs that are likely functional based on matching the functional landscape of the given protein, i.e. providing an explanation for a known functionality or binding partner of the protein that could not be derived from the normal proteoform.

For a subset of the 40 proteins, mainly those undergoing homo-oligomerization or heterodimerization, the AlphaFold 3 (AF3) web server [56] was used to model the potential structural effect of the N-terminal extensions on oligomerization (Table S1). Similarly, AF3 was used to model the selected domain–motif interactions described in the article and depicted in Figures 2–5. In these AF3

predictions, the motifs were provided in the form of peptides that include ± 5 residues flanking regions around the actual motifs predicted by ELM. Phosphorylated form of the motif was used in the case of phospho-dependent interactions, such as the CKS1 docking motif and the FBXW7 degron motif of SUFU.

### *Identification of proteomics evidence for the existence of N-terminal extensions*

Sequences of N-terminal extensions including the overlap with corresponding CDS until the first trypsin cleavage site were used as input (SPEN: MTVSYEAGEGEPAAAVAGTP PSMVR; SOX2: MITIIGGGRIGQRRREALFLILIPVCLSLFFP QIILRLIFLAEPCAPDTPARLPSSSPPARGPPKVPAGPRVG-GRRRAGPAHSAR; SUFU: MARQCSPRRLPSPVPPVPALRT PMAELRPSGAPGPTAPPAPGPTAPPAFASLFPPGLHAIYG-ECR; TOP1: MRLLEPPESPSARTGRFAVCVSPTPPRLPPRSS LRADMSGDHLHNDSQIEADFR; SFPQ: MASTFPERLLRFC LDRPLTTDM SR;) we searched for proteomic evidence using targeted peptide search engine PepQuery2 [83]. PepQuery analysis was conducted using a web-based application with the following default settings: PepQuery version 2.0.2, Fixed modification -Carbamidomethylation of C, Variable modification – Oxidation of M, Maximal allowed variable modification – 3, Add AA substitution – false, Enzyme – Trypsin, Max Missed cleavages – 1, Precursor mass tolerance – 20.0, Range of allowed isotope peak errors – 0, Precursor ion mass tolerance unit – ppm, Fragment ion mass tolerance – 0.6, Fragment ion mass tolerance unit – Da, Scoring algorithm – Hyperscore, Min score – 12.0, Min peaks – 10, Min peptide length – 7, Max peptide length – 45, Min peptide mass500.0, Max peptide mass – 10000.0, Random peptide number – 1000 in fast searching mode. The reference database used is Gencode_V34_human. The following MS/MS datasets identifiers were used for analysis: PDC000220, PDC000219, MSV000085836, PDC000128, PDC000127, PDC000245, PDC000205, PDC000204, PDC000222, PDC000221, PDC000233, PDC000232, PDC000234, PDC000237, PDC000224, PDC000149, PDC000153, PDC000271, PDC000270, PDC000176, PDC000180, PDC000239, PDC000121, PDC000120, PDC000117, PDC000116, PDC000109, PDC000251, PDC000110, PDC000119, PDC000118, PDC000174, PDC000173, PDC000111, PDC000112, PDC000115, PDC000114, PDC000226, PDC000126, PDC000125, PDX010154, PDX016999, PDC000198, PDC000262, PDC000216, PDC000215, PDC000214.

### *Software tools used for visualizations*

Figure panels showing domain maps were created using DOG2 visualization software [184], while AF3-predicted structural models were depicted using UCSF ChimeraX [185]. The Venn diagram was drawn using a dedicated website: https://bioinformatics.psb.ugent.be/webtools/Venn/.

## Acknowledgments

This article is based upon work from COST Action CA21154 TRANSLACORE, supported by COST (European Cooperation in Science and Technology).

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This research was funded by a grant from the National Research, Development, and Innovation Office (NRDIO/NKFIH) of Hungary [FK-142285 to R.P.]. R.P. is a holder of the János Bolyai Research Fellowship of the Hungarian Academy of Sciences [BO/00174/22] and was supported by a postdoctoral Eötvös Scholarship from the Tempus Public Foundation [no.: 184018]; D.E.A. is supported by the Russian Science Foundation 24-14-00213.

## Author contributions

R. P. Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Validation, Visualization, Writing – original draft, Writing – review and editing

D. E. A. Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Validation, Visualization, Writing – original draft, Writing – review and editing

K. D. Conceptualization, Project Administration, Visualization, Writing – original draft, Writing – review and editing.

All authors have read and approved the final work.

## Data availability statement

Data for this publication is contained in the paper and the supplementary material.

## ORCID

Rita Pancsa http://orcid.org/0000-0003-0849-9312
Dmitry E. Andreev http://orcid.org/0000-0001-5320-3045
Kellie Dean http://orcid.org/0000-0002-3213-8181

## References

[1] Hinnebusch AG. Structural insights into the mechanism of scanning and start codon recognition in eukaryotic translation initiation. Trends Biochemical Sciences. 2017;42(8):589–611. doi: 10.1016/j.tibs.2017.03.004

[2] Pelletier J, Sonenberg N. The organizing principles of eukaryotic ribosome recruitment. Annu Rev Biochem. 2019;88(1):307–335. doi: 10.1146/annurev-biochem-013118-111042

[3] Hinnebusch AG, Ivanov IP, Sonenberg N. Translational control by 5′-untranslated regions of eukaryotic mRNAs. Science. 2016;352(6292):1413–1416. doi: 10.1126/science.aad9868

[4] Dever TE, Ivanov IP, Hinnebusch AG. Translational regulation by uORFs and start codon selection stringency. Genes Dev. 2023;37 (11–12):474–489. doi: 10.1101/gad.350752.123

[5] Kearse MG, Wilusz JE. Non-AUG translation: a new start for protein synthesis in eukaryotes. Genes Dev. 2017;31 (17):1717–1731. doi: 10.1101/gad.305250.117

[6] Andreev DE, Loughran G, Fedorova AD, et al. Non-AUG translation initiation in mammals. Genome Biol. 2022;23(1):111. doi: 10.1186/s13059-022-02674-2

[7] Robichaud N, Sonenberg N, Ruggero D, et al. Translational control in cancer. Cold Spring Harb Perspect Biol. 2019;11(7): a032896. doi: 10.1101/cshperspect.a032896
[8] Bhat M, Robichaud N, Hulea L, et al. Targeting the translation machinery in cancer. Nat Rev Drug Discov. 2015;14(4):261–278. doi: 10.1038/nrd4505
[9] Lee LJ, Papadopoli D, Jewer M, et al. Cancer plasticity: the role of mRNA translation. Trends Cancer. 2021;7(2):134–145. doi: 10.1016/j.trecan.2020.09.005
[10] Chu J, Cargnello M, Topisirovic I, et al. Translation initiation factors: reprogramming protein synthesis in cancer. Trends Cell Biol. 2016;26(12):918–933. doi: 10.1016/j.tcb.2016.06.005
[11] Truitt ML, Ruggero D. New frontiers in translational control of the cancer genome. Nat Rev Cancer. 2016;16(5):288–304. doi: 10.1038/nrc.2016.27
[12] Smith RCL, Kanellos G, Vlahov N, et al. Translation initiation in cancer at a glance. J Cell Sci. 2021;134(1):jcs248476. doi: 10.1242/jcs.248476
[13] Kovalski JR, Kuzuoglu-Ozturk D, Ruggero D. Protein synthesis control in cancer: selectivity and therapeutic targeting. Embo J. 2022;41(8):e109823. doi: 10.15252/embj.2021109823
[14] Fabbri L, Chakraborty A, Robert C, et al. The plasticity of mRNA translation during cancer progression and therapy resistance. Nat Rev Cancer. 2021;21(9):558–577. doi: 10.1038/s41568-021-00380-y
[15] Hann SR, King MW, Bentley DL, et al. 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. 1988;52(2):185–195. doi: 10.1016/0092-8674(88)90507-7
[16] Hann SR, Sloan-Brown K, Spotts GD. Translational activation of the non-AUG-initiated c-myc 1 protein at high cell densities due to methionine deprivation. Genes Dev. 1992;6(7):1229–1240. doi: 10.1101/gad.6.7.1229
[17] Hann SR. Methionine deprivation regulates the translation of functionally-distinct c-Myc proteins. Adv Exp Med Biol. 1995;375:107–116.
[18] Tang L, Morris J, Wan J, et al. Competition between translation initiation factor eIF5 and its mimic protein 5MP determines non-AUG initiation rate genome-wide. Nucleic Acids Research. 2017;45(20):11941–11953. doi: 10.1093/nar/gkx808
[19] Sato K, Masuda T, Hu Q, et al. Novel oncogene 5MP1 reprograms c-Myc translation initiation to drive malignant phenotypes in colorectal cancer. EBioMedicine. 2019 Jun;44:387–402. doi: 10.1016/j.ebiom.2019.05.058
[20] Sherr CJ. Principles of tumor suppression. Cell. 2004 Jan 23;116 (2):235–246. doi: 10.1016/S0092-8674(03)01075-4
[21] Maehama T, Dixon JE. The tumor suppressor, PTEN/MMAC1, dephosphorylates the lipid second messenger, phosphatidylinositol 3,4,5-trisphosphate. J Biol Chem. 1998;273(22):13375–13378. doi: 10.1074/jbc.273.22.13375
[22] Stambolic V, Suzuki A, de la Pompa JL, et al. Negative regulation of PKB/Akt-dependent cell survival by the tumor suppressor PTEN. Cell. 1998;95(1):29–39. doi: 10.1016/S0092-8674(00)81780-8
[23] Ivanov IP, Firth AE, Michel AM, et al. Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences. Nucleic Acids Res. 2011 May;39 (10):4220–4234. doi: 10.1093/nar/gkr007
[24] Hopkins BD, Fine B, Steinbach N, et al. A secreted PTEN phosphatase that enters cells to alter signaling and survival. Science. 2013;341(6144):399–402. doi: 10.1126/science.1234907
[25] Liang H, He S, Yang J, et al. PTENα, a PTEN isoform translated through alternative initiation, regulates mitochondrial function and energy metabolism. Cell Metabolism. 2014;19(5):836–848. doi: 10.1016/j.cmet.2014.03.023
[26] Tzani I, Ivanov IP, Andreev DE, et al. Systematic analysis of the PTEN 5′ leader identifies a major AUU initiated proteoform. Open Biol. 2016 May;6(5):150203. doi: 10.1098/rsob.150203
[27] Liang H, Chen X, Yin Q, et al. PTENβ is an alternatively translated isoform of PTEN that regulates rDNA transcription. Nat Commun. 2017;8(1):14771. doi: 10.1038/ncomms14771
[28] Lee YR, Chen M, Pandolfi PP. The functions and regulation of the PTEN tumour suppressor: new modes and prospects. Nat Rev Mol Cell Biol. 2018 Sep;19(9):547–562. doi: 10.1038/s41580-018-0015-0
[29] Shen SM, Zhang C, Ge MK, et al. PTENα and PTENβ promote carcinogenesis through WDR5 and H3K4 trimethylation. Nat Cell Biol. 2019 Nov;21(11):1436–1448. doi: 10.1038/s41556-019-0409-z
[30] Huang X, Zhang C, Shang X, et al. The NTE domain of PTENα/β promotes cancer progression by interacting with WDR5 via its SSSRRSS motif. Cell Death Dis. 2024;15(5):335. doi: 10.1038/s41419-024-06714-6
[31] Hastie ND. Wilms' tumour 1 (WT1) in development, homeostasis and disease. Development. 2017;144(16):2862–2872. doi: 10.1242/dev.153163
[32] Haber DA, Buckler AJ, Glaser T, et al. An internal deletion within an 11p13 zinc finger gene contributes to the development of Wilms' tumor. Cell. 1990;61(7):1257–1269. doi: 10.1016/0092-8674(90)90690-G
[33] Inoue K, Sugiyama H, Ogawa H, et al. WT1 as a new prognostic factor and a new marker for the detection of minimal residual disease in acute leukemia. Blood. 1994;84(9):3071–3079. doi: 10.1182/blood.V84.9.3071.3071
[34] Oji Y, Ogawa H, Tamaki H, et al. Expression of the Wilms' tumor gene WT1 in solid tumors and its involvement in tumor cell growth. Jpn J Cancer Res. 1999 Feb;90(2):194–204. doi: 10.1111/j.1349-7006.1999.tb00733.x
[35] Huff V. Wilms' tumours: about tumour suppressor genes, an oncogene and a chameleon gene. Nat Rev Cancer. 2011 Feb;11 (2):111–121. doi: 10.1038/nrc3002
[36] Lee KY, Jeon YJ, Kim HG, et al. The CUG-translated WT1, not AUG-WT1, is an oncogene. Carcinogenesis. 2017;38 (12):1228–1240. doi: 10.1093/carcin/bgx108
[37] Yoshitomi H, Lee KY, Yao K, et al. GSK3β-mediated expression of CUG-Translated WT1 is critical for tumor progression. Cancer Research. 2021;81(4):945–955. doi: 10.1158/0008-5472.CAN-20-1880
[38] Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646–674. doi: 10.1016/j.cell.2011.02.013
[39] Florkiewicz RZ, Sommer A. Human basic fibroblast growth factor gene encodes four polypeptides: three initiate translation from non-AUG codons. Proc Natl Acad Sci USA. 1989 Jun;86 (11):3978–3981. doi: 10.1073/pnas.86.11.3978
[40] Prats H, Kaghad M, Prats AC, et al. High molecular mass forms of basic fibroblast growth factor are initiated by alternative CUG codons. Proc Natl Acad Sci USA. 1989 Mar;86(6):1836–1840. doi: 10.1073/pnas.86.6.1836
[41] Arnaud E, Touriol C, Boutonnet C, et al. A new 34-kilodalton isoform of human fibroblast growth factor 2 is cap dependently synthesized by using a non-AUG start codon and behaves as a survival factor. Mol Cell Biol. 1999 Jan;19(1):505–514. doi: 10.1128/MCB.19.1.505
[42] Pintucci G, Quarto N, Rifkin DB. Methylation of high molecular weight fibroblast growth factor-2 determines post-translational increases in molecular weight and affects its intracellular distribution. Mol Biol Cell. 1996 Aug;7(8):1249–1258. doi: 10.1091/mbc.7.8.1249
[43] Dono R, James D, Zeller R. A GR-motif functions in nuclear accumulation of the large FGF-2 isoforms and interferes with mitogenic signalling. Oncogene. 1998;16(16):2151–2158. doi: 10.1038/sj.onc.1201746
[44] Nickle A, Ko S, Merrill AE. Fibroblast growth factor 2. Differentiation. 2024 Sep-Oct;139:100733. doi: 10.1016/j.diff.2023.10.001
[45] Tee MK, Jaffe RB. A precursor form of vascular endothelial growth factor arises by initiation from an upstream in-frame CUG codon. Biochemical Journal. 2001;359(1):219–226. doi: 10.1042/bj3590219
[46] Meiron M, Anunu R, Scheinman EJ, et al. New Isoforms of VEGF are translated from alternative initiation CUG codons located in

its 5′UTR. Biochem Biophys Res Com. 2001;282(4):1053–1060. doi: 10.1006/bbrc.2001.4684
[47] Rosenbaum-Dekel Y, Fuchs A, Yakirevich E, et al. Nuclear localization of long-VEGF is associated with hypoxia and tumor angiogenesis. Biochem Biophys Res Commun. 2005;332 (1):271–278. doi: 10.1016/j.bbrc.2005.04.123
[48] Katsman M, Azriel A, Horev G, et al. N-VEGF, the autoregulatory Arm of VEGF-A. Cells. 2022;11(8):1289. doi: 10.3390/cells11081289
[49] Ingolia NT, Ghaemmaghami S, Newman JRS, et al. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009;324(5924):218–223. doi: 10.1126/science.1168978
[50] Guo H, Ingolia NT, Weissman JS, et al. Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature. 2010;466(7308):835–840. doi: 10.1038/nature09267
[51] Fedorova AD, Kiniry SJ, Andreev DE, et al. Thousands of human non-AUG extended proteoforms lack evidence of evolutionary selection among mammals. Nat Commun. 2022;13(1):7910. doi: 10.1038/s41467-022-35595-6
[52] Suehnholz SP, Nissan MH, Zhang H, et al. Quantifying the expanding landscape of clinical actionability for patients with cancer. Cancer Discovery. 2024;14(1):49–65. doi: 10.1158/2159-8290.CD-23-0467
[53] Chakravarty D, Gao J, Phillips SM, et al. OncoKB: a precision oncology knowledge base. JCO Precis Oncol. 2017;2017: O.17.00011.
[54] Szulc NA, Stefaniak F, Piechota M, et al. DEGRONOPEDIA: a web server for proteome-wide inspection of degrons. Nucleic Acids Res. 2024;52(W1):W221–32. doi: 10.1093/nar/gkae238
[55] Teufel F, Almagro Armenteros JJ, Johansen AR, et al. SignalP 6.0 predicts all five types of signal peptides using protein language models. Nat Biotechnol. 2022 Jul;40(7):1023–1025. doi: 10.1038/s41587-021-01156-3
[56] Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024 Jun;630(8016):493–500. doi: 10.1038/s41586-024-07487-w
[57] Kumar M, Michael S, Alvarado-Valverde J, et al. ELM—the Eukaryotic linear motif resource—2024 update. Nucleic Acids Res. 2024;52(D1):D442–55. doi: 10.1093/nar/gkad1058
[58] He TC, Sparks AB, Rago C, et al. Identification of c-MYC as a target of the APC pathway. Science. 1998;281 (5382):1509–1512. doi: 10.1126/science.281.5382.1509
[59] van de Wetering M, Sancho E, Verweij C, et al. The β-Catenin /TCF-4 complex imposes a crypt progenitor phenotype on colorectal cancer cells. Cell. 2002;111(2):241–250. doi: 10.1016/S0092-8674(02)01014-0
[60] Yu Y, Wu J, Wang Y, et al. Kindlin 2 forms a transcriptional complex with β-catenin and TCF4 to enhance wnt signalling. EMBO Rep. 2012 Aug;13(8):750–758. doi: 10.1038/embor.2012.88
[61] Medici D, Hay ED, Olsen BR, et al. Snail and slug promote epithelial-mesenchymal transition through β-Catenin–T-Cell factor-4-dependent expression of transforming growth factor-β3. Mol Biol Cell. 2008 Nov;19(11):4875–4887. doi: 10.1091/mbc.e08-05-0506
[62] Bao Y, Guo Y, Yang Y, et al. PRSS8 suppresses colorectal carcinogenesis and metastasis. Oncogene. 2019 Jan;38(4):497–517. doi: 10.1038/s41388-018-0453-3
[63] UniProt Consortium T. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2018;46(5):2699. doi: 10.1093/nar/gky092
[64] Knott GJ, Bond CS, Fox AH. The DBHS proteins SFPQ, NONO and PSPC1: a multipurpose molecular scaffold. Nucleic Acids Res. 2016;44(9):3989–4004. doi: 10.1093/nar/gkw271
[65] Peng R, Dye BT, Pérez I, et al. PSF and p54nrb bind a conserved stem in U5 snRNA. RNA. 2002 Oct;8(10):1334–1347. doi: 10.1017/S1355838202022070
[66] Gozani O, Patton JG, Reed R. A novel set of spliceosome-associated proteins and the essential splicing factor PSF bind stably to pre-mRNA prior to catalytic step II of the splicing reaction. Embo J. 1994;13(14):3356–3367. doi: 10.1002/j.1460-2075.1994.tb06638.x
[67] Teigelkamp S, Mundt C, Achsel T, et al. The human U5 snRNP-specific 100-kD protein is an RS domain-containing, putative RNA helicase with significant homology to the yeast splicing factor Prp28p. RNA. 1997 Nov;3(11):1313–1326.
[68] Lee M, Sadowska A, Bekere I, et al. The structure of human SFPQ reveals a coiled-coil mediated polymer essential for functional aggregation in gene regulation. Nucleic Acids Res. 2015;43 (7):3826–3840. doi: 10.1093/nar/gkv156
[69] Lukong KE, Huot ME, Richard S. BRK phosphorylates PSF promoting its cytoplasmic localization and cell cycle arrest. Cell Signal. 2009 Sep;21(9):1415–1422. doi: 10.1016/j.cellsig.2009.04.008
[70] Tyzack GE, Neeves J, Crerar H, et al. Aberrant cytoplasmic intron retention is a blueprint for RNA binding protein mislocalization in VCP-related amyotrophic lateral sclerosis. Brain. 2021;144 (7):1985–1993. doi: 10.1093/brain/awab078
[71] Shav-Tal Y, Zipori D. PSF and p54 nrb /NonO – multi-functional nuclear proteins. FEBS Letters. 2002;531(2):109–114. doi: 10.1016/S0014-5793(02)03447-6
[72] Feng P, Li L, Deng T, et al. NONO and tumorigenesis: more than splicing. J Cell Mol Med. 2020 Apr;24(8):4368–4376. doi: 10.1111/jcmm.15141
[73] Schell B, Legrand P, Fribourg S. Crystal structure of SFPQ-NONO heterodimer. Biochimie. 2022 Jul;198:1–7. doi: 10.1016/j.biochi.2022.02.011
[74] Straub T, Grue P, Uhse A, et al. The RNA-splicing factor PSF/p54 controls DNA-topoisomerase I activity by a direct interaction. Journal Of Biological Chemistry. 1998;273(41):26261–26264. doi: 10.1074/jbc.273.41.26261
[75] Straub T, Knudsen BR, Boege F. PSF/p54 nrb stimulates "jumping" of DNA topoisomerase I between separate DNA helices. Biochemistry. 2000;39(25):7552–7558. doi: 10.1021/bi992898e
[76] Takeiwa T, Ikeda K, Horie K, et al. Role of RNA binding proteins of the Drosophila behavior and human splicing (DBHS) family in health and cancer. RNA Biol. 2024 Jan;21(1):459–475. doi: 10.1080/15476286.2024.2332855
[77] Tsukahara T, Matsuda Y, Haniu H. PSF knockdown enhances apoptosis via downregulation of LC3B in human colon cancer cells. Biomed Res Int. 2013;2013:1–8. doi: 10.1155/2013/204973
[78] Takayama KI, Horie-Inoue K, Katayama S, et al. Androgen-responsive long noncoding RNA CTBP1-AS promotes prostate cancer. Embo J. 2013;32(12):1665–1680. doi: 10.1038/emboj.2013.99
[79] Mitobe Y, Iino K, Takayama KI, et al. PSF promotes ER-Positive breast cancer progression via Posttranscriptional regulation of ESR1 and SCFD2. Cancer Res. 2020;80(11):2230–2242. doi: 10.1158/0008-5472.CAN-19-3095
[80] Pellarin I, Dall'acqua A, Gambelli A, et al. Splicing factor proline- and glutamine-rich (SFPQ) protein regulates platinum response in ovarian cancer-modulating SRSF2 activity. Oncogene. 2020 May;39(22):4390–4403. doi: 10.1038/s41388-020-1292-6
[81] Kauffman EC, Ricketts CJ, Rais-Bahrami S, et al. Molecular genetics and cellular features of TFE3 and TFEB fusion kidney cancers. Nat Rev Urol. 2014 Aug;11(8):465–475. doi: 10.1038/nrurol.2014.162
[82] Rao Q, Shen Q, Xia QY, et al. PSF/SFPQ is a very common gene fusion partner in TFE3 rearrangement-associated perivascular epithelioid cell tumors (PEComas) and melanotic Xp11 translocation renal cancers: clinicopathologic, immunohistochemical, and molecular characteristics suggesting classification as a distinct entity. Am J Surg Pathol. 2015 Sep;39(9):1181–1196. doi: 10.1097/PAS.0000000000000502
[83] Wen B, Zhang B. PepQuery2 democratizes public MS proteomics data for rapid peptide searching. Nat Commun. 2023;14(1):2213. doi: 10.1038/s41467-023-37462-4
[84] Rayner SL, Cheng F, Hogan AL, et al. ALS/FTD-causing mutation in cyclin F causes the dysregulation of SFPQ. Human Mol Genet. 2021;30(11):971–984. doi: 10.1093/hmg/ddab073

[85] Siebert A, Gattringer V, Weishaupt JH, et al. ALS-linked loss of cyclin-F function affects HSP90. Life Sci Alliance. 2022;5(12): e202101359. doi: 10.26508/lsa.202101359
[86] Williams KL, Topp S, Yang S, et al. CCNF mutations in amyotrophic lateral sclerosis and frontotemporal dementia. Nat Commun. 2016;7(1):11253. doi: 10.1038/ncomms11253
[87] Lim YW, James D, Huang J, et al. The emerging role of the RNA-Binding protein SFPQ in neuronal function and neurodegeneration. IJMS. 2020;21(19):7151. doi: 10.3390/ijms21197151
[88] D'Angiolella V, Donato V, Vijayakumar S, et al. SCF(Cyclin F) controls centrosome homeostasis and mitotic fidelity through CP110 degradation. Nature. 2010;466(7302):138–142. doi: 10.1038/nature09140
[89] Klein DK, Hoffmann S, Ahlskog JK, et al. Cyclin F suppresses B-Myb activity to promote cell cycle checkpoint control. Nat Commun. 2015;6(1):5800. doi: 10.1038/ncomms6800
[90] Yuan R, Liu Q, Segeren HA, et al. Cyclin F-dependent degradation of E2F7 is critical for DNA repair and G2-phase progression. Embo J. 2019;38(20):e101430. doi: 10.15252/embj.2018101430
[91] Dankert JF, Rona G, Clijsters L, et al. Cyclin F-Mediated degradation of SLBP limits H2A.X accumulation and apoptosis upon genotoxic stress in G2. Mol Cell. 2016 Nov 3;64(3):507–519. doi: 10.1016/j.molcel.2016.09.010
[92] Lowe ED, Tews I, Cheng KY, et al. Specificity determinants of recruitment peptides bound to phospho-CDK2/cyclin a. Biochemistry. 2002;41(52):15625–15634. doi: 10.1021/bi0268910
[93] Wakula P, Beullens M, Ceulemans H, et al. Degeneracy and function of the ubiquitous RVXF motif that mediates binding to protein phosphatase-1. J Biol Chem. 2003;278(21):18817–18823. doi: 10.1074/jbc.M300175200
[94] Liu L, Xie N, Rennie P, et al. Consensus PP1 binding motifs regulate transcriptional corepression and alternative RNA splicing activities of the steroid receptor coregulators, p54nrb and PSF. Mol Endocrinol. 2011 Jul;25(7):1197–1210. doi: 10.1210/me.2010-0517
[95] Hirschi A, Cecchini M, Steinhardt RC, et al. An overlapping kinase and phosphatase docking site regulates activity of the retinoblastoma protein. Nat Struct Mol Biol. 2010 Sep;17 (9):1051–1057. doi: 10.1038/nsmb.1868
[96] Mikami S, Kanaba T, Takizawa N, et al. Structural insights into the recruitment of SMRT by the corepressor SHARP under phosphorylative regulation. Structure. 2014;22(1):35–46. doi: 10.1016/j.str.2013.10.007
[97] Ariyoshi M, Schwabe JWR. A conserved structural motif reveals the essential transcriptional repression function of spen proteins and their role in developmental signaling. Genes Dev. 2003;17 (15):1909–1920. doi: 10.1101/gad.266203
[98] Shi Y, Downes M, Xie W, et al. Sharp, an inducible cofactor that integrates nuclear receptor repression and activation. Genes Dev. 2001;15(9):1140–1151. doi: 10.1101/gad.871201
[99] Appel LM, Franke V, Benedum J, et al. The SPOC domain is a phosphoserine binding module that bridges transcription machinery with co- and post-transcriptional regulators. Nat Commun. 2023;14(1):166. doi: 10.1038/s41467-023-35853-1
[100] Oswald F, Winkler M, Cao Y, et al. RBP-Jκ/SHARP recruits CtIP/CtBP corepressors to silence notch target genes. Mol Cell Biol. 2005 Dec;25(23):10379–10390. doi: 10.1128/MCB.25.23.10379-10390.2005
[101] Nardini M, Spanò S, Cericola C, et al. CtBP/BARS: a dual-function protein involved in transcription co-repression and Golgi membrane fission. EMBO Journal. 2003;22 (12):3122–3130. doi: 10.1093/emboj/cdg283
[102] Lum L, Beachy PA. The hedgehog response network: sensors, switches, and routers. Science. 2004;304(5678):1755–1759. doi: 10.1126/science.1098020
[103] Stone DM, Murone M, Luoh S, et al. Characterization of the human suppressor of fused, a negative regulator of the zinc-finger transcription factor Gli. J Cell Sci. 1999 Dec;112(Pt 23):4437–4448. doi: 10.1242/jcs.112.23.4437
[104] Kogerman P, Grimm T, Kogerman L, et al. Mammalian suppressor-of-fused modulates nuclear–cytoplasmic shuttling of GLI-1. Nat Cell Biol. 1999 Sep;1(5):312–319. doi: 10.1038/13031
[105] Merchant M, Vajdos FF, Ultsch M, et al. Suppressor of fused regulates Gli activity through a dual binding mechanism. Mol Cell Biol. 2004 Oct;24(19):8627–8641. doi: 10.1128/MCB.24.19.8627-8641.2004
[106] Murone M, Luoh SM, Stone D, et al. Gli regulation by the opposing activities of fused and suppressor of fused. Nat Cell Biol. 2000 May;2(5):310–312. doi: 10.1038/35010610
[107] Jiang J. Hedgehog signaling mechanism and role in cancer. Semin Cancer Biol. 2022 Oct;85:107–122. doi: 10.1016/j.semcancer.2021.04.003
[108] Taylor MD, Liu L, Raffel C, et al. Mutations in SUFU predispose to medulloblastoma. Nat Genet. 2002 Jul;31(3):306–310. doi: 10.1038/ng916
[109] Lee Y, Kawagoe R, Sasai K, et al. Loss of suppressor-of-fused function promotes tumorigenesis. Oncogene. 2007;26 (44):6442–6447. doi: 10.1038/sj.onc.1210467
[110] Bonilla X, Parmentier L, King B, et al. Genomic analysis identifies new drivers and progression pathways in skin basal cell carcinoma. Nat Genet. 2016 Apr;48(4):398–406. doi: 10.1038/ng.3525
[111] Tate JG, Bamford S, Jubb HC, et al. COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res. 2019;47(D1): D941–7. doi: 10.1093/nar/gky1015
[112] Raducu M, Fung E, Serres S, et al. SCF (Fbxl17) ubiquitylation of Sufu regulates hedgehog signaling and medulloblastoma development. Embo J. 2016;35(13):1400–1416. doi: 10.15252/embj.201593374
[113] Zhuang T, Zhang B, Song Y, et al. Sufu negatively regulates both initiations of centrosome duplication and DNA replication. Proc Natl Acad Sci USA. 2021;118(28):e2026421118. doi: 10.1073/pnas.2026421118
[114] Popov N, Wanzel M, Madiredjo M, et al. The ubiquitin-specific protease USP28 is required for MYC stability. Nat Cell Biol. 2007 Jul;9(7):765–774. doi: 10.1038/ncb1601
[115] Koepp DM, Schaefer LK, Ye X, et al. Phosphorylation-dependent ubiquitination of cyclin E by the SCFFbw7 ubiquitin ligase. Science. 2001;294(5540):173–177. doi: 10.1126/science.1065203
[116] Koo J, Wu X, Mao Z, et al. Rictor undergoes glycogen synthase kinase 3 (GSK3)-dependent, FBXW7-mediated ubiquitination and proteasomal degradation. J Biol Chem. 2015;290 (22):14120–14129. doi: 10.1074/jbc.M114.633057
[117] Hao B, Oehlmann S, Sowa ME, et al. Structure of a Fbw7-Skp1-cyclin E complex: multisite-phosphorylated substrate recognition by SCF ubiquitin ligases. Mol Cell. 2007;26(1):131–143. doi: 10.1016/j.molcel.2007.02.022
[118] Hao B, Zheng N, Schulman BA, et al. Structural basis of the Cks1-dependent recognition of p27(Kip1) by the SCF(Skp2) ubiquitin ligase. Mol Cell. 2005;20(1):9–19. doi: 10.1016/j.molcel.2005.09.003
[119] Stolzenburg S, Rots MG, Beltran AS, et al. Targeted silencing of the oncogenic transcription factor SOX2 in breast cancer. Nucleic Acids Res. 2012 Aug;40(14):6725–6740. doi: 10.1093/nar/gks360
[120] Yang Z, Pan X, Gao A, et al. Expression of Sox2 in cervical squamous cell carcinoma. J Buon. 2014 Jan-Mar;19(1):203–206.
[121] Hütz K, Mejías-Luque R, Farsakova K, et al. The stem cell factor SOX2 regulates the tumorigenic potential in human gastric cancer cells. Carcinogenesis. 2014 Apr;35(4):942–950. doi: 10.1093/carcin/bgt410
[122] Herreros-Villanueva M, Zhang JS, Koenig A, et al. SOX2 promotes dedifferentiation and imparts stem cell-like features to pancreatic cancer cells. Oncogenesis. 2013;2(8):e61. doi: 10.1038/oncsis.2013.23
[123] Jia X, Li X, Xu Y, et al. SOX2 promotes tumorigenesis and increases the anti-apoptotic property of human prostate cancer cell. J Mol Cell Biol. 2011 Aug;3(4):230–238. doi: 10.1093/jmcb/mjr002
[124] Chen S, Li X, Lu D, et al. SOX2 regulates apoptosis through MAP4K4-survivin signaling pathway in human lung cancer cells.

Carcinogenesis. 2014 Mar;35(3):613–623. doi: 10.1093/carcin/bgt371
[125] Alonso MM, Diez-Valle R, Manterola L, et al. Genetic and epigenetic modifications of Sox2 contribute to the invasive phenotype of malignant gliomas. PLOS ONE. 2011;6(11):e26740. doi: 10.1371/journal.pone.0026740
[126] Girouard SD, Laga AC, Mihm MC, et al. SOX2 contributes to melanoma cell invasion. Lab Invest. 2012 Mar;92(3):362–370. doi: 10.1038/labinvest.2011.188
[127] Han X, Fang X, Lou X, et al. Silencing SOX2 induced mesenchymal-epithelial transition and its expression predicts liver and lymph node metastasis of CRC patients. PLOS ONE. 2012;7(8):e41335. doi: 10.1371/journal.pone.0041335
[128] Lou X, Han X, Jin C, et al. SOX2 targets fibronectin 1 to promote cell migration and invasion in ovarian cancer: new molecular leads for therapeutic intervention. OMICS. 2013 Oct;17(10):510–518. doi: 10.1089/omi.2013.0058
[129] Hussenet T, Dali S, Exinger J, et al. SOX2 is an oncogene activated by recurrent 3q26.3 amplifications in human lung squamous cell carcinomas. PLOS ONE. 2010;5(1):e8960. doi: 10.1371/journal.pone.0008960
[130] Annovazzi L, Mellai M, Caldera V, et al. SOX2 expression and amplification in gliomas and glioma cell lines. Cancer Genomics Proteomics. 2011 May-Jun;8(3):139–147.
[131] Hussenet T, du Manoir S. SOX2 in squamous cell carcinoma: amplifying a pleiotropic oncogene along carcinogenesis. Cell Cycle. 2010;9(8):1480–1486. doi: 10.4161/cc.9.8.11203
[132] Bass AJ, Watanabe H, Mermel CH, et al. SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. Nat Genet. 2009 Nov;41(11):1238–1242. doi: 10.1038/ng.465
[133] Maier S, Wilbertz T, Braun M, et al. SOX2 amplification is a common event in squamous cell carcinomas of different organ sites. Hum Pathol. 2011 Aug;42(8):1078–1088. doi: 10.1016/j.humpath.2010.11.010
[134] Rudin CM, Durinck S, Stawiski EW, et al. Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. Nat Genet. 2012 Oct;44(10):1111–1116. doi: 10.1038/ng.2405
[135] Freier K, Knoepfle K, Flechtenmacher C, et al. Recurrent copy number gain of transcription factor SOX2 and corresponding high protein expression in oral squamous cell carcinoma. Genes Chromosomes Cancer. 2010 Jan;49(1):9–16. doi: 10.1002/gcc.20714
[136] van Schaijik B, Davis PF, Wickremesekera AC, et al. Subcellular localisation of the stem cell markers OCT4, SOX2, NANOG, KLF4 and c-MYC in cancer: a review. J Clin Pathol. 2018 Jan;71(1):88–91. doi: 10.1136/jclinpath-2017-204815
[137] Schaefer T, Mittal N, Wang H, et al. Nuclear and cytosolic fractions of SOX2 synergize as transcriptional and translational co-regulators of cell fate. Cell Reports. 2024;43(10):114807. doi: 10.1016/j.celrep.2024.114807
[138] Baltus GA, Kowalski MP, Zhai H, et al. Acetylation of sox2 induces its nuclear export in embryonic stem cells. Stem Cells. 2009 Sep;27(9):2175–2184. doi: 10.1002/stem.168
[139] Fukushima H, Ogura K, Wan L, et al. SCF-mediated Cdh1 degradation defines a negative feedback system that coordinates cell-cycle progression. Cell Reports. 2013;4(4):803–816. doi: 10.1016/j.celrep.2013.07.031
[140] Mao DD, Gujar AD, Mahlokozera T, et al. A CDC20-APC/SOX2 signaling axis regulates human Glioblastoma stem-like cells. Cell Reports. 2015;11(11):1809–1821. doi: 10.1016/j.celrep.2015.05.027
[141] Chang L, Zhang Z, Yang J, et al. Atomic structure of the APC/C and its mechanism of protein ubiquitination. Nature. 2015;522(7557):450–454. doi: 10.1038/nature14471
[142] Davey NE, Morgan DO. Building a regulatory network with short linear sequence motifs: lessons from the Degrons of the anaphase-promoting complex. Mol Cell. 2016;64(1):12–23. doi: 10.1016/j.molcel.2016.09.006
[143] He J, Chao WCH, Zhang Z, et al. Insights into degron recognition by APC/C coactivators from the structure of an Acm1-Cdh1 complex. Mol Cell. 2013;50(5):649–660. doi: 10.1016/j.molcel.2013.04.024
[144] Scipion CPM, Ghoshdastider U, Ferrer FJ, et al. Structural evidence for the roles of divalent cations in actin polymerization and activation of ATP hydrolysis. Proc Natl Acad Sci USA. 2018;115(41):10345–10350. doi: 10.1073/pnas.1806394115
[145] Fung HYJ, Fu SC, Brautigam CA, et al. Structural determinants of nuclear export signal orientation in binding to exportin CRM1. Elife. 2015;4:e10034. doi: 10.7554/eLife.10034
[146] Valles GJ, Bezsonova I, Woodgate R, et al. USP7 is a master regulator of genome stability. Front Cell Dev Biol. 2020;8:717.
[147] Pommier Y, Sun Y, Huang SYN, et al. Roles of eukaryotic topoisomerases in transcription, replication and genomic stability. Nat Rev Mol Cell Biol. 2016 Nov;17(11):703–721. doi: 10.1038/nrm.2016.111
[148] Debéthune L, Kohlhagen G, Grandas A, et al. Processing of nucleopeptides mimicking the topoisomerase I-DNA covalent complex by tyrosyl-DNA phosphodiesterase. Nucleic Acids Res. 2002;30(5):1198–1204. doi: 10.1093/nar/30.5.1198
[149] Sun Y, Saha LK, Saha S, et al. Debulking of topoisomerase DNA-protein crosslinks (TOP-DPC) by the proteasome, non-proteasomal and non-proteolytic pathways. DNA Repair (Amst). 2020 Oct;94:102926. doi: 10.1016/j.dnarep.2020.102926
[150] Sun Y, Chen J, Huang SYN, et al. Parylation prevents the proteasomal degradation of topoisomerase I DNA-protein crosslinks and induces their deubiquitylation. Nat Commun. 2021;12(1):5010. doi: 10.1038/s41467-021-25252-9
[151] Sun Y, Miller Jenkins LM, Su YP, et al. A conserved SUMO pathway repairs topoisomerase DNA-protein cross-links by engaging ubiquitin-mediated proteasomal degradation. Sci Adv. 2020 Nov;6(46):eaba6290. doi: 10.1126/sciadv.aba6290
[152] Saridakis V, Sheng Y, Sarkari F, et al. Structure of the p53 binding domain of HAUSP/USP7 bound to Epstein-Barr nuclear antigen 1 implications for EBV-mediated immortalization. Molecular Cell. 2005;18(1):25–36. doi: 10.1016/j.molcel.2005.02.029
[153] Loughran G, Sachs MS, Atkins JF, et al. Stringency of start codon selection modulates autoregulation of translation initiation factor eIF5. Nucleic Acids Res. 2012 Apr;40(7):2898–2906. doi: 10.1093/nar/gkr1192
[154] Ivanov IP, Loughran G, Sachs MS, et al. Initiation context modulates autoregulation of eukaryotic translation initiation factor 1 (eIF1). Proc Natl Acad Sci USA. 2010;107(42):18056–18060. doi: 10.1073/pnas.1009269107
[155] Loughran G, Firth AE, Atkins JF, et al. Translational autoregulation of BZW1 and BZW2 expression by modulating the stringency of start codon selection. PLOS ONE. 2018;13(2):e0192648. doi: 10.1371/journal.pone.0192648
[156] Fijalkowska D, Verbruggen S, Ndah E, et al. eIF1 modulates the recognition of suboptimal translation initiation sites and steers gene expression via uORFs. Nucleic Acids Research. 2017;45(13):7997–8013. doi: 10.1093/nar/gkx469
[157] Zhang J, Pi SB, Zhang N, et al. Translation regulatory factor BZW1 regulates preimplantation embryo development and compaction by restricting global non-AUG initiation. Nat Commun. 2022;13(1):6621. doi: 10.1038/s41467-022-34427-x
[158] She R, Luo J, Weissman JS. Translational fidelity screens in mammalian cells reveal eIF3 and eIF4G2 as regulators of start codon selectivity. Nucleic Acids Research. 2023;51(12):6355–6369. doi: 10.1093/nar/gkad329
[159] Martin M, Maßhöfer L, Temming P, et al. Exome sequencing identifies recurrent somatic mutations in EIF1AX and SF3B1 in uveal melanoma with disomy 3. Nat Genet. 2013 Aug;45(8):933–936. doi: 10.1038/ng.2674
[160] Martin-Marcos P, Zhou F, Karunasiri C, et al. eIF1A residues implicated in cancer stabilize translation preinitiation complexes and favor suboptimal initiation sites in yeast. Elife. 2017;6:e31250. doi: 10.7554/eLife.31250

[161] Meril S, Bahlsen M, Eisenstein M, et al. Loss-of-function cancer-linked mutations in the EIF4G2 non-canonical translation initiation factor. Life Sci Alliance. 2024 Mar;7(3):e202302338. doi: 10.26508/lsa.202302338

[162] Ivanov IP, Saba JA, Fan CM, et al. Evolutionarily conserved inhibitory uORFs sensitize Hox mRNA translation to start codon selection stringency. Proc Natl Acad Sci USA. 2022;119 (9):e2117226119. doi: 10.1073/pnas.2117226119

[163] Ly J, Xiang K, Su KC, et al. Nuclear release of eIF1 restricts start-codon selection during mitosis. Nature. 2024 Nov;635 (8038):490–498. doi: 10.1038/s41586-024-08088-3

[164] Gamble N, Paul EE, Anand B, et al. Regulation of the interactions between human eIF5 and eIF1A by the CK2 kinase. Curr Res Struct Biol. 2022;4:308–319. doi: 10.1016/j.crstbi.2022.09.003

[165] Paul EE, Lin KY, Gamble N, et al. Dynamic interaction network involving the conserved intrinsically disordered regions in human eIF5. Biophys Chem. 2022 Feb;281:106740. doi: 10.1016/j.bpc.2021.106740

[166] Homma MK, Wada I, Suzuki T, et al. CK2 phosphorylation of eukaryotic translation initiation factor 5 potentiates cell cycle progression. Proc Natl Acad Sci USA. 2005;102 (43):15688–15693. doi: 10.1073/pnas.0506791102

[167] Dennis MD, Person MD, Browning KS. Phosphorylation of plant translation initiation factors by CK2 enhances the in vitro interaction of multifactor complex components. Journal Of Biological Chemistry. 2009;284(31):20615–20628. doi: 10.1074/jbc.M109.007658

[168] Ivanov IP, Shin BS, Loughran G, et al. Polyamine control of translation elongation regulates start site selection on antizyme inhibitor mRNA via ribosome queuing. Molecular Cell. 2018;70 (2):254–64.e6. doi: 10.1016/j.molcel.2018.03.015

[169] Manjunath H, Zhang H, Rehfeld F, et al. Suppression of ribosomal pausing by eIF5A is necessary to maintain the fidelity of start codon selection. Cell Reports. 2019;29(10):3134–46.e6. doi: 10.1016/j.celrep.2019.10.129

[170] Dasgupta A, Prensner JR. Upstream open reading frames: new players in the landscape of cancer gene regulation. NAR Cancer. 2024 Jun;6(2):zcae023. doi: 10.1093/narcan/zcae023

[171] Weatheritt RJ, Davey NE, Gibson TJ. Linear motifs confer functional diversity onto splice variants. Nucleic Acids Res. 2012 Aug;40(15):7123–7131. doi: 10.1093/nar/gks442

[172] Weatheritt RJ, Gibson TJ. Linear motifs: lost in (pre)translation. Trends Biochem Sci. 2012 Aug;37(8):333–341. doi: 10.1016/j.tibs.2012.05.001

[173] Buljan M, Chalancon G, Eustermann S, et al. Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. Molecular Cell. 2012;46 (6):871–883. doi: 10.1016/j.molcel.2012.05.039

[174] Bogaert A, Fijalkowska D, Staes A, et al. N-terminal proteoforms may engage in different protein complexes. Life Sci Alliance. 2023 Aug;6(8):e202301972. doi: 10.26508/lsa.202301972

[175] Moschonas GD, De Meyer M, De Sutter D, et al. Virotrap: trapping protein complexes in virus-like particles. Methods Mol Biol. 2023;2718:53–71.

[176] Monti C, Zilocchi M, Colugnat I, et al. Proteomics turns functional. J Proteomics. 2019;198:36–44. doi: 10.1016/j.jprot.2018.12.012

[177] Brown KA, Melby JA, Roberts DS, et al. Top-down proteomics: challenges, innovations, and applications in basic and clinical research. Expert Rev Proteomics. 2020 Oct;17(10):719–733. doi: 10.1080/14789450.2020.1855982

[178] Cupp-Sutton KA, Wang Z, Yu D, et al. RPLC-RPLC-MS/MS for proteoform identification. Methods Mol Biol. 2022;2500:31–42.

[179] Guo Y, Cupp-Sutton KA, Zhao Z, et al. Multidimensional separations in top–down proteomics. Anal Sci Adv. 2023 Jul;4(5--6):181–203. doi: 10.1002/ansa.202300016

[180] Staes A, Van Damme P, Helsens K, et al. Improved recovery of proteome-informative, protein N-terminal peptides by combined fractional diagonal chromatography (COFRADIC). Proteomics. 2008 Apr;8(7):1362–1370. doi: 10.1002/pmic.200700950

[181] Staes A, Impens F, Van Damme P, et al. Selecting protein N-terminal peptides by combined fractional diagonal chromatography. Nat Protoc. 2011;6(8):1130–1141.

[182] Fedorova AD, Kiniry SJ, Andreev DE, et al. Addendum: thousands of human non-AUG extended proteoforms lack evidence of evolutionary selection among mammals. Nat Commun. 2024;15(1):228.

[183] Oughtred R, Rust J, Chang C, et al. The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci. 2021 Jan;30(1):187–200. doi: 10.1002/pro.3978

[184] Ren J, Wen L, Gao X, et al. DOG 1.0: illustrator of protein domain structures. Cell Res. 2009 Feb;19(2):271–273. doi: 10.1038/cr.2009.6

[185] Meng EC, Goddard TD, Pettersen EF, et al. UCSF ChimeraX: tools for structure building and analysis. Protein Sci. 2023 Nov;32 (11):e4792. doi: 10.1002/pro.4792