Trends in Biochemical Sciences

CellPress REVIEWS

Review

N-Terminal Proteoforms in Human Disease

Annelies Bogaert,[1,2] Esperanza Fernandez,[1,2] and Kris Gevaert[1,2,*]

The collection of chemically different protein variants, or proteoforms, by far exceeds the number of protein-coding genes in the human genome. Major contributors are alternative splicing and protein modifications. In this review, we focus on those proteoforms that differ at their N termini with a molecular link to disease. We describe the main underlying mechanisms that give rise to such N-terminal proteoforms, these being splicing, initiation of protein translation, and protein modifications. Given their role in several human diseases, it is becoming increasingly clear that several of these N-terminal proteoforms may have potential as therapeutic interventions and/or for diagnosing and prognosing their associated disease.

Highlights

The chemical space taken by proteoforms is immense. Somewhat overlooked in recent literature are proteoforms that differ at their aminotermini, N-terminal proteoforms.

Diverse molecular mechanisms create N-terminal proteoforms.

Mainly atomistic studies have revealed different functions of N-terminal proteoforms, with several of these proteoforms contributing directly to human diseases.

In the future, specific N-terminal proteoforms might be targets for therapeutic intervention and/or used as proxies for disease diagnosis and prognosis.

Increasing the Proteome Complexity from the Start

In GENCODE, the number of human protein-coding genes is at 20 454 [1]. However, the full collection of chemically different human proteins is impossible to predict. This much higher number of so-called **proteoforms** (see Glossary) compared with protein-coding genes stems from gene transcription (the use of different promoters), processing of immature mRNA molecules (alternative splicing), translation (**alternative translation initiation**, ribosomal frameshifting, and stop codon read-through), and a vast number of protein modifications [2–4] (Figure 1A–C).

One specific type of proteoform are those that stem from the same gene, but differ at their N termini, which we call 'N-terminal' proteoforms. Such proteoforms may hold extended, truncated, and/or modified N termini with respect to the canonical protein (the term 'canonical' is used here as defined in the UniProtKB/Swiss-Prot database) and the mechanisms leading to their production are diverse. In eukaryotes, the canonical mechanism for translation to start involves a ribosome assembling at the 5′ end of a mature mRNA molecule, which then starts scanning for start codons towards the 3′ end. Alternative start codons can be used for translation by various mechanisms. For instance, secondary RNA structures may hinder ribosome scanning and thereby promote translation to initiate from other codons [5]. Translation can also start at internal ribosomal entry sites (IRES, Box 1) [6] and upon leaky scanning, in which the first encountered start codon is embedded in a suboptimal **Kozak consensus sequence** and ribosomes scan for a downstream start codon present in a more optimal context [7] (Figure 1D). Other mechanisms include **translation re-initiation** after a short upstream open reading frame (ORF) [8,9] (Figure 1E), and **5′ mRNA leader sequence recapping** [10]. In addition, alternative splicing may give rise to transcripts that have different 5′ ends and these, possibly combined with alternative translation initiation, may also give rise to N-terminal proteoforms [11,12] (Figure 1).

By using ribosome profiling data to yield an adapted search space for mass spectrometry data, it was shown that 10–20% of all identified protein N termini in several human and mouse cells originated from alternative translation initiation or alternative splicing (Box 2). Furthermore, several of these N-terminal proteoforms were found to be conserved, hinting at a potential biological importance [3,13]. Besides these mechanisms, protein processing may also lead to functional N-terminal proteoforms. Protein processing is involved in protein maturation (e.g., initiator methionine and signal peptide removal) and, more generally, in protein processing. Several

[1]VIB Center for Medical Biotechnology, VIB, B-9000 Ghent, Belgium
[2]Department of Biomolecular Medicine, Ghent University, B-9000 Ghent, Belgium

*Correspondence: kris.gevaert@vib-ugent.be (K. Gevaert).

proteomic studies aimed at mapping new N termini introduced upon protease action (and, thus, new N-terminal proteoforms), for example, N-terminomics studies, have identified novel N-terminal proteoforms. A striking example is a study by Lange and coworkers, which identified a large number of processed protein products that appeared to be stabilized by post-translational N-terminal acetylation in human erythrocytes [14]. Given that this review focuses on (post)transcriptional and co-translational events leading to N-terminal proteoforms, we do not discuss post-translational events giving rise to functional N-terminal proteoforms further.

Until recently, (N-terminal) proteoforms had often been overlooked, but studies on their biological function are now emerging and show that the N terminus of a proteoform may influence the stability, subcellular localization, and functionalities of a protein, and its interactions with other proteins [15–20]. Although a recent review stressed the importance of studying proteoforms in health and disease [21], N-terminal proteoforms arising from alternative start sites were not discussed. Here, we fill this gap and review recent findings on the involvement of N-terminal proteoforms (although not solely arising from alternative translation initiation) in human disease (Table 1, Key Table).

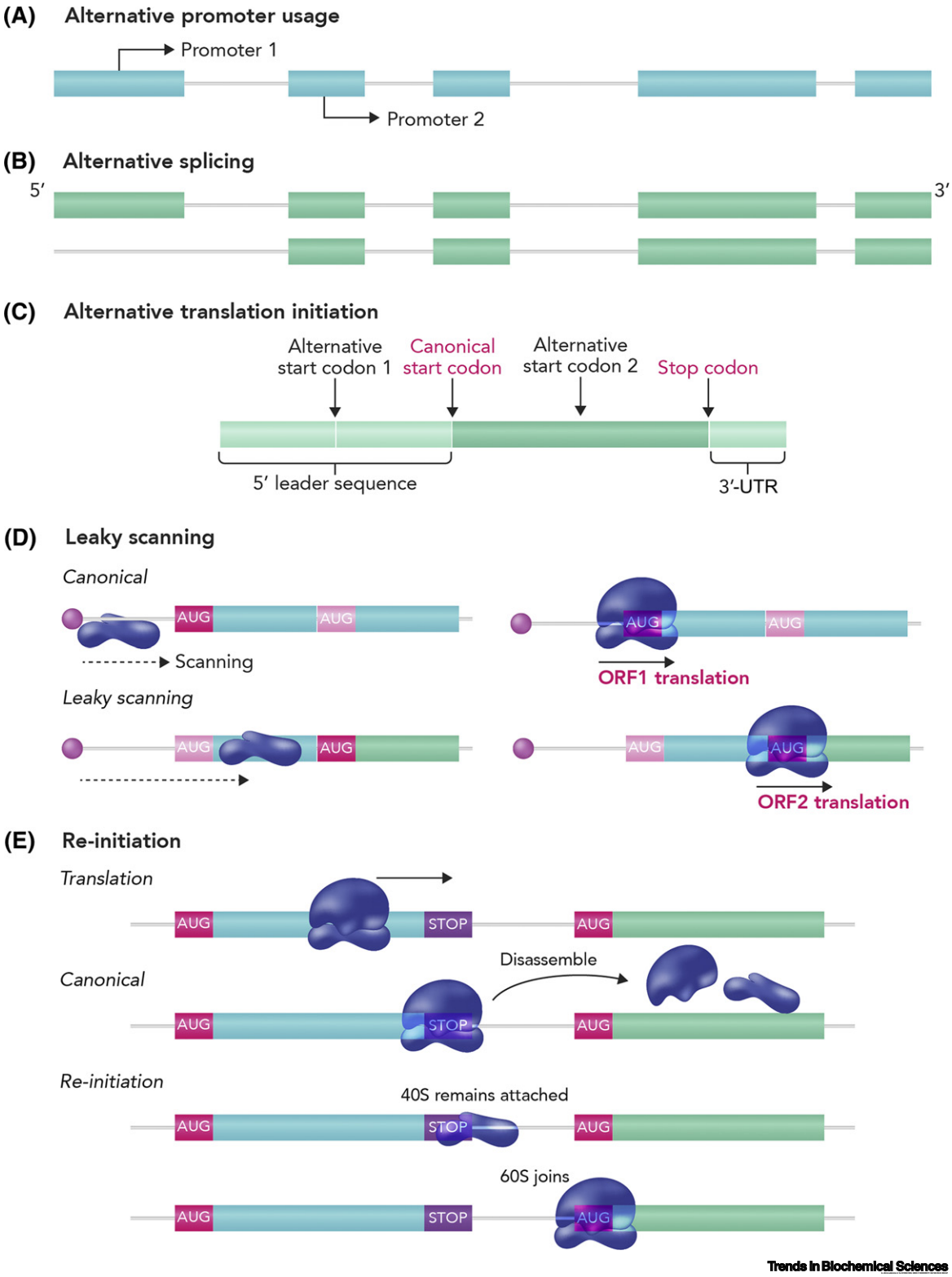
Alternative Splicing Adds a Layer of N-Terminal Complexity

Nuclear factor of activated T cells, cytoplasmic 1 (NFATc1) has two major isoforms called NFATc1-α and NFATc1-β, the expression of which is regulated by different promoters. NFATc1-α is NFAT inducible and results from splicing out the second exon, while NFATc1-β is expressed at basal levels and transcribed from the second exon, which results in two N-terminal proteoforms [22,23]. Upon T cell receptor engagement and NFAT protein activation, NFATc1-α is massively expressed, exceeding the expression levels of NFATc1-β several times [24]. The N termini of these proteoforms have a transactivation domain (TAD) that has several functions [24]. In NIH 3T2 cells, NFATc1-α was found to increase cell proliferation and to induce several hallmarks of cell transformation. By contrast, NFATc1-β reduced cell proliferation and increased cell death (by increasing FasL and TNF-α levels) due to an acidic activation domain (AAD) only present in the TAD of NFATc1-β. When these NFATc1 proteoforms were expressed in mice, NFATc1-α induced large tumors with a high growth rate, whereas the β-proteoform induced smaller tumors with moderate growth rates. Additionally, human peripheral blood mononuclear cells obtained from patients with Burkitt lymphoma showed high levels of NFATc1-α and low levels of NFATc1-β, while the same cells from healthy donors showed lower levels of NFATc1-α compared with NFATc1-β. These data suggest that a specific NFATc1 proteoform is involved in tumor formation. Given that NFATc1-α and NFATc1-β may be involved in different cellular functions, deregulation of proteoform expression could contribute to tumorigenesis. In addition, in Burkitt lymphoma, sustained activity of NFATc1 has been linked to diffuse large B cell lymphoma, T cell acute lymphoblastic leukemia, chronic lymphocytic leukemia, melanoma, and pancreatic and colorectal carcinomas [24].

The cytoplasmic tyrosine-protein kinase BMX, or epithelial and endothelial tyrosine kinase (ETK), is a nonreceptor tyrosine kinase with functions in several cellular processes [25]. ETK/BMX regulates the activity of proteins such as PI3-AKT, TNFR2, PAK1, TP53, PIM, and STAT3, and has an important role in inflammation. It is overexpressed in several tumor types and its inhibition reduces tumor cell proliferation and angiogenesis [26,27]. Ibrutinib, an FDA-approved drug to treat mantle cell lymphoma and chronic leukemia by inhibiting Bruton's tyrosine kinase (BTK), a member of the BMX/ETK family, also inactivates BMX-STAT3 in glioma stem cells and impairs tumor growth [28]. Recently, an N-terminal proteoform of BMX/ETK, resulting from skipping exons 1–8, was found in 21 out of 174 lung adenocarcinoma samples, while it was absent in control samples [29]. This N-terminal proteoform was dominant in tumor samples that had low to no expression of the canonical ETK/BMX proteoform and was associated with the epidermal growth factor

Glossary

5′ mRNA leader sequence recapping: the 5′ end of mRNA molecules can be truncated by endoribonucleases or following the partial degradation by 5′ exoribonucleases by either secondary mRNA structures or RNA-binding proteins. This leads to shorter mRNA molecules, which can be recapped by recapping enzymes. As a consequence, these mRNA molecules can be translated but might differ in their 5′ region, which leads to N-terminal proteoforms.
Alternative translation initiation: initiation of translation on mRNA molecules at start codons that differ from the canonical start codon, or by translation initiation at an internal ribosome entry site. Both lead to the production of N-terminal proteoforms.
Kozak consensus sequence: a nucleotide sequence motif found in eukaryotic mRNA that surrounds the start codon and promotes translation initiation. The sequence is noted as gccRccAUGG, in which the capital letters are highly conserved and R can either be an adenine or guanidine. Lower-case letters represent the most occurring base at that position, but these vary more and, thus, are less conserved.
N-terminal proteoforms: proteoforms that chemically differ at their N termini.
Proteoform: the proteoform definition was introduced to describe all chemically different forms in which the protein products of a single protein-coding gene appear. As such, differences in the actual protein sequence due to genetic variation, alternative promoter usage, alternative splicing, alternative translation initiation, and protein modifications are captured by this definition.
Translation re-initiation: the small ribosomal subunit remains attached to the mRNA following termination of translation. It resumes scanning on the same RNA molecule and can initiate again at a downstream start codon.



(See figure legend at the bottom of the next page.)

receptor mutation (EGFR-L858R). This N-terminal BMX/ETK proteoform was shown to promote tumor cell migration and facilitate cell transformation [29].

The ribosomal protein S6 kinase proteins (S6K1 and S6K2) are serine/threonine kinases that belong to the AGC kinase family and are downstream effectors of the mammalian target of rapamycin (mTOR). mTOR/SK6 targets multiple proteins that control protein translation and regulate pathways such as cell growth and/or size, proliferation, cell cycle progression, and metabolism. mTOR/SK6 has important roles in physiology and diseases, including diabetes, cancer, aging, organ growth, learning, and obesity (reviewed in [30]). S6K1 and S6K2 share 84% identity in their kinase domains and 43% and 59% identity in their N- and C-terminal domains, respectively. The most extensively studied S6K1 has two N-terminal proteoforms, $p70^{S6K1}$ (or S6Kα2), which contains 502 amino acids and is localized in the cytosol, and $p85^{S6K1}$ (or S6Kα1), the longest proteoform, containing an additional 23 N-terminal amino acids including a nuclear localization sequence. $p85^{S6K1}$ can be secreted and enter surrounding cells via its N-terminal 6-arginine repetition motif, which resembles a motif in the HIV TAT protein with similar properties [31]. Secreted $p85^{S6K1}$ increases phosphorylation of its targets, cell size, and cell migration, which are innate phenotypes of cancerous cells. *In vivo*, this proteoform promoted tumor growth of breast cancer cells and lung metastasis, suggesting it as a target to treat breast cancer [31].

N-Terminal Proteoforms Generated by Alternative Translation Initiation

Caveolin-2 (cav-2α) is involved in insulin signaling and localizes in the plasma membrane, where it recruits the insulin receptor and regulates the initiation of insulin receptor substrate-1 directed signaling. Cav-2β is translated from a downstream AUG start codon (methionine-14) and predominantly expressed in insulin-resistant obese subjects, where it desensitizes the insulin receptor via dephosphorylation, leading to its lysosomal degradation and causing insulin resistance [32] (Figure 2).

Endoplasmic reticulum membrane sensor NFE2L1 (NRF1) is a transcription factor essential for maintaining cellular homeostasis, organ development and growth, and the adaptive response to pathophysiological processes. NRF1 dysfunction is associated with diabetes, liver cancer, and other malignancies [33,34]. Canonical NRF1 (Nrf1α) is yielded by alternative splicing, skipping exon 4 from the main ORF. A shorter form, Nrfβ, is in-frame translated from an internal Kozak sequence located around the four methionine codons between positions 289 and 297 in the sequence of the mouse protein. A third, smaller proteoform, Nrf1γ, is produced by either in-frame translation starting at methionine 584 and/or by proteolytic processing of the longer Nrf1 proteoforms. These Nrf1 proteoforms regulate different sets of homeostatic and developmental target genes, which are involved in several pathological processes [33,34]. Nrf1α and Nrf1β account for the main Nrf1-mediated transcription of downstream genes. Nrf1γ acts in a dominant-negative manner, leading to the downregulation of several key genes, some of which are targets of Nrf1α and Nrf1β. Nrf1γ most likely interferes with the functional assembly of active transcription factors (Nrf1α, Nrf1β, and Nrf2) [34].

Thrombocytopenia 2 is caused by monoallelic mutations in the 5′-untranslated region (UTR) of ANKRD26. Patients with this disease also have an increased risk of developing acute myeloid leukemia (AML). To find the molecular link between both diseases, Marconi *et al.* screened the

Figure 1. Examples of how N-Terminal Proteoforms Are Generated. A single protein-coding gene may give rise to different N-terminal proteoforms by (A) alternative promoter usage, (B) alternative splicing leading to transcripts differing at the 5′ end, and (C) alternative translation initiation. The two major mechanisms leading to alternative translation initiation are (D) leaky scanning, where the first start codon located in a weak Kozak sequence (indicated in light pink) is skipped, and translation starts from a second start codon located in a more optimal Kozak sequence, (indicated by dark pink), and (E) translation re-initiation, in which the small ribosomal subunit remains attached to the mRNA after it passes a stop codon and can re-initiate translation at a downstream start codon. Abbreviations: ORF, open reading frame; UTR, untranslated region.

Box 1. Cellular Stress and Cap-Independent Translation

Cap-dependent initiation of translation, being the assembly of the ribosome at the 5′ end of a mRNA molecule to start scanning towards the 3′ direction for start codons, was long seen as the only possibility to initiate translation of eukaryotic mRNA molecules. However, the discovery of an alternative mode of translation initiation of viral transcripts in eukaryotic cells changed this view [67]. In this cap-independent scanning mode, the 40S ribosomal subunit is directly recruited to an internal ribosome entry site (IRES) to start translation and this gives rise to N-terminal proteoforms [5]. IRES translation is favored when cap-dependent translation is blocked, as, for instance, in various pathophysiological stress conditions (e.g., amino acid starvation, apoptosis, and hypoxia) and during cell cycle progression, cell differentiation, and cell development [68,69]. IRES translation is known to contribute to several pathological states, such as Alzheimer's disease [70], several tumors, and cancers, such as glioblastoma and breast cancer [71]. Given that most papers on IRES-dependent translation are less than recent, besides indicating here a link between IRES and disease, we do not consider this particular aspect of N-terminal proteoform formation further here.

5′UTR and exon 1 of ANKRD26 in 250 patients with AML. Three patients carried two different variants in the 5′ end of the *ANKRD26* coding region, c.3G>A or c.105C>G, resulting in the production of two different N-terminally truncated proteoforms of ANKRD26. These proteoforms maintain ANKDR26 functionality and lead to even stronger activation of the MAPK/ERK pathway. ERK hyperactivation leads to thrombocytopenia and could contribute to the predisposition to myeloid malignancies [35].

Chromosomal aneuploidy and specific gene mutations are early hallmarks of many oncogenic processes. The most common form of aneuploidy is trisomy of chromosome 21, causing Down's syndrome. Patients with Down's syndrome often show hematopoietic abnormalities, such as transient myeloproliferative disorder (TMD), which can progress to AML. These patients with TMD carry GATA1 mutations that lead to the production of a GATA1 short variant (GATA1s) by translation from methionine-84. Trisomy 21 perturbs normal hematopoietic development through the enhanced production of early hematopoietic progenitors and increases the expression of GATA1s, causing excessive aberrant differentiation of megakaryocytes. Several genes (*RUNX1*, *ETS2*, and *ERG*) located at a critical 4-Mb region of chromosome 21, of which expression levels are increased, mediate this effect [36].

Cellular tumor antigen p53 (p53) is one of the best-known proteins associated with cancer. Known for its tumor-suppressor activities, mutant and dysfunctional p53 proteins, as well as

Box 2. Detecting a N-Terminal Proteoform Can Be Challenging

Mass spectrometry-driven proteomics has been a key method to identify N-terminal proteoforms. Of note, whereas ribosomal footprinting led to lists of thousands of rather unexpected, even near-cognate, translation initiation sites, disappointingly, only a minor fraction of these has been detected by means of proteomics. One important aspect to consider is that there is generally just one peptide per N-terminal proteoform that verifies its presence in a sample, that is, its furthest N-terminal peptide. Without any enrichment of such peptides, the presence of higher numbers of non-N-terminal peptides (e.g., in a bottom-up, shotgun proteomics experiment) will hinder the identification of the N-terminal peptides. In addition, especially for cytosolic eukaryotic proteins, protein N termini are frequently acetylated *in vivo*, which makes the corresponding N-terminal peptides less basic, and this may interfere with their ionization, thus lowering the overall sensitivity of their detection. This acetyl group could also interfere with the actual fragmentation of peptide ions in the fragmentation cell of the mass spectrometer, thereby yielding ill-predicted peptide fragmentation patterns that may interfere with the actual matching of a theoretical peptide sequence to an experimental fragmentation spectrum.

Another important aspect to consider is the actual search space that is used to identify the peptides. For instance, in-frame translation starting at downstream or upstream near-cognate start codons yields proteoforms of which the exact identity of their N-terminal parts is not necessarily captured in protein sequence databases. Indeed, such translation starting at downstream start codons is expected to yield a proteoform starting with an initiator methionine, whereas another amino acid might be indicated in the stored protein sequence. One way to overcome this is to merge ribosomal footprinting data to protein sequence data, although this increases the size of the database at the expense of losing overall coverage of N-terminal proteoforms.

Key Table

Table 1. Overview of Recently Discovered N-Terminal Proteoforms with a Link to Disease

Gene	Proteoforms	Link to human disease	Refs
Alternative splicing			
NFATC1	NFATc1-α and NFATc1-β produced by combination of different promoters for transcription and alternative splicing	NFATc1-α is involved in tumor formation, while NFATc1-β acts as an anti-oncogenic factor; involved in several cancers, such as leukemia, lymphoma, and pancreatic and colorectal cancers	[24]
BMX	ΔEx1-8 BMX results from skipping exons 1–8	ΔEx1-8 BMX was dominantly expressed in 21 out of 174 lung adenocarcinoma samples, while it was absent in control samples	[29]
S6K2 $p85^{S6K1}$	S6K2 $p85^{S6K1}$ contains 23 more N-terminal amino acids than the $p70^{S6K1}$ proteoform, including a 6-arginine repetition motif	S6K2 $p85^{S6K1}$ promotes tumor growth of breast cancer cells and lung metastasis	[31]
Alternative translation initiation			
CAV2	Second AUG-codon (Met-14) of Cav-2 is used as internal translational start site. Resulting shorter 18 kDa Cav-2β proteoform lacks 13 amino acids at its N terminus	Caveolin-2β causes insulin resistance	[32]
NRF1	Nrf1α (full length); Nrf1β lacks the first 296 amino acids; and Nrf1γ starts at position 584 resulting from alternative translation initiation; Nrf1γ might also be produced by endoproteolytic processing	These proteoforms transcribe distinct subsets of target genes that are involved in cancer, neurodegeneration, and diabetes. Nrf1γ is dominant negative over the other two proteoforms	[34]
ANKRD26	Mutants c.3G>A and c.105C>G lead to N-terminally truncated variants of ANKRD26	Mutations found in patients with acute myeloid leukemia (AML); cause overexpression of ANKDR26 and hyperactivation of MAPK/ERK signaling pathway, both of which are linked to thrombocytopenia and AML	[35]
GATA1	Mutations lead to GATA1-S, a shorter form starting at methionine-84	Related to Down's syndrome, in which mutations lead to sole production of GATA1-S, in turn leading to RUNX1/ETS2/ERG-mediated hyperproliferation of aberrant megakaryoblasts resulting in transient myeloproliferative disorders	[36]
TP53	Δ40p53 is a N-terminal truncated proteoform resulting from alternative splicing and/or alternative translation initiation at AUG-40	Predominant element of amyloid aggregates and proposed key modulator of p53 tumor suppression and oncogenic activities in endometrial carcinoma	[38]
		Δ40p53 suppresses tumor proliferation, induces cellular senescence, and upregulates p53 expression in hepatocellular carcinoma cells	[39]
PTEN	PTEN-L arising from translation starting at an upstream CUG codon	PTEN and PTEN-L are depleted in renal carcinoma cells; expression of PTEN-L acts as a tumor suppressor and this protein is considered to be a potential therapeutic target because it can enter cells	[42]
SOX-9	Mutation creates upstream translation initiation start codon	Leads to acampomelic campomelic dysplasia	[43]
GRHPR	Two variants in the 5′UTR on separate alleles: c.-4G>A andc.-3C>T occurring in *cis* creating out-of-frame alternative start site	Alternative translation initiation creates 20-amino acid peptide with no relation to GRHPR and causes reduced levels of GRHPR, which results in primary hyperoxaluria type II	[44]
RUNX1	Translation occurs at Met-1 and Met-25, giving rise to full-length RUNX1 and a shorter protein, lacking the first 24 amino acids (RUNX1ΔN24)	Retains hematopoietic activity. Translation initiation from Met-25 may act as a fail-safe mechanism to maintain normal hematopoiesis when production of full-length RUNX1 protein is inhibited by genetic mutations	[45]
N-terminal modifications			
APP	Several N-terminal proteoforms	Proteoforms are enriched in Aβ soluble aggregates, a hallmark of Alzheimer's disease	[46]
CST3	Three N-terminal truncations: desS-, des-SSP, and des-SSPG	des-S and des-SSP proteoforms are enriched in patients with type 2 diabetes mellitus and chronic kidney disease, while des-SSPG has been found in intracystic fluid of pilocytic astrocytoma pediatric brain tumors	[49,50]

Table 1. (continued)

Gene	Proteoforms	Link to human disease	Refs
PPARGC1A	N-terminal acetylation by NAA10	Pgc1α N-terminal acetylation by NAA10 prevents its binding to PPARγ and the subsequent expression of thermogenic genes. Absence of this N-terminal acetyl group prevents diet-induced obesity	[53]
H4	Loss of N-terminal acetylation by Nat4	Absence of N-terminal acetyl group induces specific stress-response genes	[55]
MeCp2e1	Missense mutation in exon 1 (c.5C>T) that results in a Ala2Val	Leads to Rett syndrome	[57]
SNCA	N-terminal acetylation	Reduces aggregation of α-Syn that mostly comprises Lewy bodies, a hallmark of Parkinson's disease	[62–65]

p53 proteoforms, were reported to have oncogenic functions [37]. The expression patterns and localization of p53 proteoforms (canonical p53, Δ40p53, and Δ133p53) were analyzed in endometrial carcinoma (EC) cells and endometrial nontumor cells [38]. Canonical p53 was found to contain the conserved N-terminal TAD. Δ40p53 proteoforms are truncated proteins resulting from alternative splicing of exon 2 and/or from translation initiation at in-frame AUG-40 [38,39] and lack part of the conserved N-terminal TAD. The Δ133p53 proteoforms result from translation initiation at AUG-133, and lack the entire TAD and part of the DNA-binding domain. Δ40p53 and canonical p53 are the main p53 forms in EC tumor cells, while Δ40p53 is the main form in nontumor cells. Δ40p53 mainly localizes to cytoplasmic punctate structures of EC cells, where it is the major component of amyloid aggregates. The N-terminal TAD domain significantly reduces aggregation of the p53 DNA-binding domain, which confirms the higher aggregation tendency of Δ40p53, which partially lacks this domain [38]. Previous studies pointed out that p53 mutants exert a dominant-negative regulatory effect over canonical p53 by converting the latter into aggregated species that acquire a gain-of-function phenotype by loss of their tumor suppressor roles [40].

Another study on Δ40p53 assessed its effect in hepatocellular carcinoma cells (HCC). Expression of Δ40p53 in these cells reduced colony formation and cell survival by inducing cellular senescence (G1 cell cycle arrest). Δ40p53 also increased the expression of several genes (including *MDM2*, *FAS*, and *p21*), indicating that it exerts its tumor-suppressor activity by promoting p53-induced gene expression. Furthermore, Δ40p53 also upregulated expression of canonical p53. These results demonstrate that Δ40p53 exerts tumor-suppressor activity by inducing cellular senescence, at least partly by upregulating p53 target gene expression in HCC [39]. Of note, these studies on Δ40p53 nicely demonstrate that a given proteoform can have opposing roles in different cancers.

Phosphatidylinositol 3,4,5-trisphosphate 3-phosphatase and dual-specificity protein phosphatase (PTEN) is another well-known tumor-associated protein. In 2013, Hopkins *et al.* [41] identified a N-terminally extended proteoform, PTEN-Long (PTEN-L), with a 173-amino acid extension that holds a secretion signal, a 6-arginine repetition similar to that of the HIV TAT protein mentioned earlier. PTEN-L was found to be secreted and imported by other cells, and had the same functionalities as canonical PTEN. In a mouse tumor model, PTEN-L uptake inhibited the PI3K pathway and tumor growth. Based on this study, Wang *et al.* analyzed the role and therapeutic significance of PTEN-L in renal cell carcinogenesis (RCC) [42]. Akt was demonstrated to control the balance between cell survival and apoptosis, and most likely has a role in carcinogenesis and progression in RCC by its elevated activity. Furthermore, compared with normal tissue, in RCC cells, PTEN and PTEN-L levels were reduced or even not detected (with the levels of PTEN-L being more reduced than those of PTEN) and the levels of activated Akt were

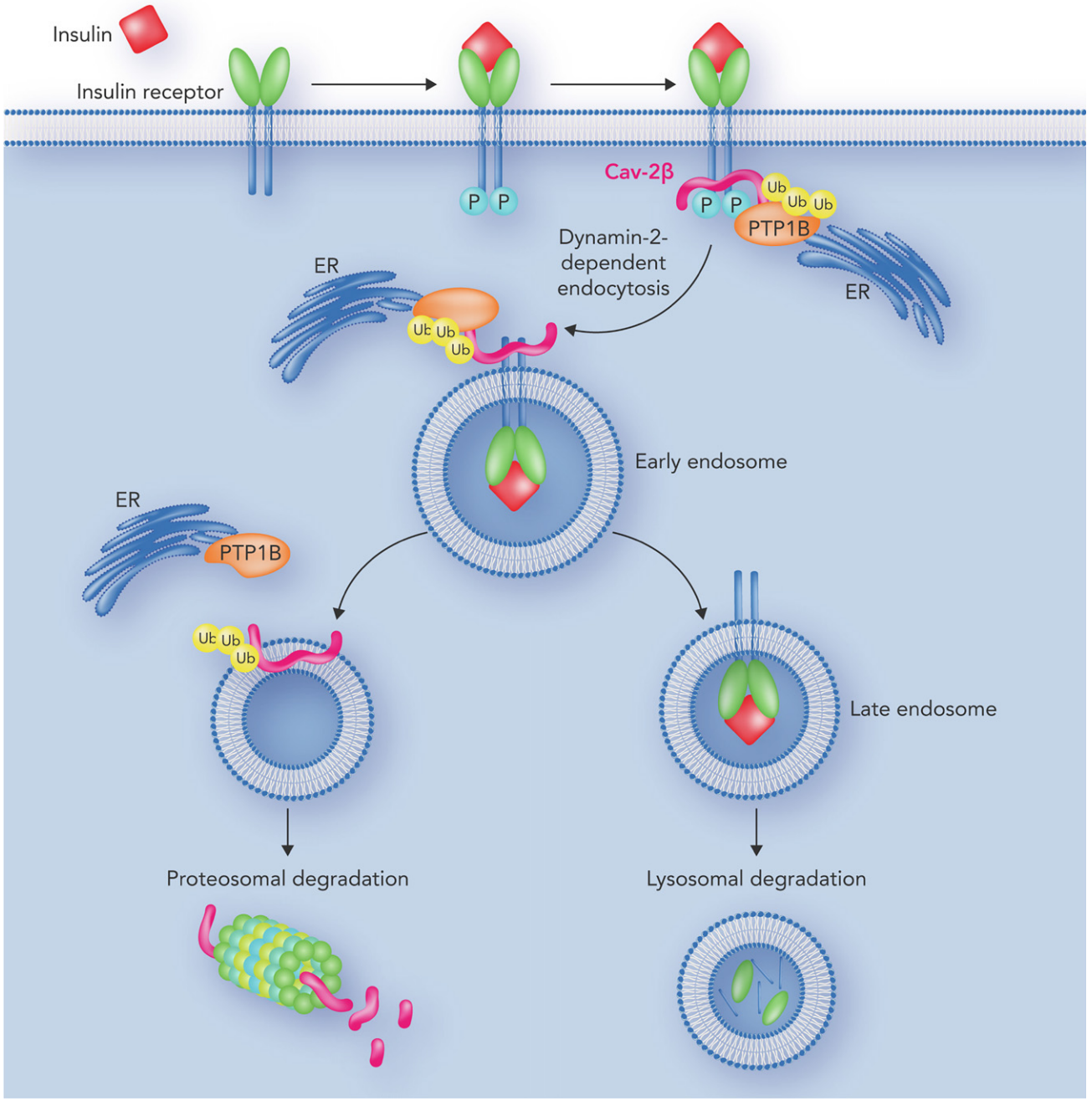


Figure 2. An N-Terminal Proteoform of Caveolin-2 (Cav-2α), Caveolin-2β, Causes Insulin Resistance. Cav-2α is involved in insulin signaling and is localized in the plasma membrane, where it recruits the insulin receptor and regulates the initiation of insulin receptor substrate-1-directed signaling. Cav-2β, the main Cav-2 form in insulin-resistant cells, is a shorter proteoform that desensitizes the insulin receptor via dephosphorylation by protein tyrosine phosphatase 1B (PTP1B), leading to its lysosomal degradation and causing insulin resistance. Abbreviation: ER, endoplasmic reticulum.

significantly increased. Of note, PTEN-L or PTEN expression in RCC cells reduced activated Akt levels by 90%, inhibited cell proliferation, migration, and invasion, and induced apoptosis. Furthermore, mice carrying tumors that were injected with PTEN-L showed tumor regression

after 4 days, while PTEN had no effect. Together, these results hint at a potential protein drug capacity of PTEN-L against RCC [42].

The transcription factor SOX-9 regulates the expression of genes involved in skeleton development and sex determination. Approximately 70 mutations in SOX-9 have been linked to (a)campomelic dysplasia or (A)CD, a disease characterized by several skeletal malformations and bending, respiratory distress, and hearing loss, and causing a high mortality rate in neonates and infants. ACD appears as a milder form of CD with less severe symptoms and better prognosis. Most mutations causing CD prevent SOX-9 production or lead to the production of a protein product with impaired functionality. In ACD, it is presumed that residual SOX-9 is produced. Evidence for this comes from a single case study that reported on a patient with ACD with no mutations in the SOX-9 coding sequence or intron/exon boundaries, but with a mutation in the 5′UTR sequence, a heterozygous mutation, c.-185G>A. This mutation gives rise to a novel upstream AUG (uAUG) translation initiation codon, also surrounded by a better Kozak sequence, 185 nucleotides before the normal AUG start codon. This uAUG creates a novel upstream ORF (uORF) of 62 codons out-of-frame with the normal ORF and terminates just after the wild-type start codon. Although this new starting site was mainly used for translation, still some, albeit strongly reduced, expression of canonical SOX-9 was observed, leading to the milder CD form [43].

Another example is glyoxylate reductase/hydroxypyruvate reductase (GRHPR). Mutations in this gene lead to the recessive genetic disorder primary hyperoxaluria type II. A genetic screen on a patient revealed two variants in the 5′UTR of GRHPR on two separate alleles (c.-4G>A and c.-3C>T). When these two occur in *cis*, a new out-of-frame translation initiation start site is created. This start site is embedded in a more ideal Kozak sequence and highly translated, leading to a 20-amino acid peptide with no relation to GRHPR. As a consequence, GRHPR translation is strongly reduced [44].

Runt-related transcription factor 1 (RUNX1) is essential for hematopoiesis and its disruption leads to hematopoietic diseases. For instance, in leukemia, RUNX1 has a growth-promoting role. A N-terminal RUNX1 proteoform starting at methionine-25 is produced by alternative translation initiation. This proteoform has an enhanced stability, retains functionality, and its expression is competitive with that of canonical RUNX1. Therefore, this proteoform might act as a fail-safe mechanism to maintain normal hematopoiesis when, due to mutations, production of full-length RUNX1 is inhibited [45].

On the Influence of N-Terminal Modifications

Soluble and insoluble aggregates of amyloid-beta (Aβ) protein have a role in the pathogenesis of Alzheimer's disease (AD). Analysis of aggregates from human AD brains by nano-liquid chromatography tandem mass spectrometry pointed to a heterogeneous population of Aβ proteoforms, including proteoforms starting at seven different positions [46]. N-terminal truncations were found to be more abundant in the insoluble aggregates due to the removal of the hydrophilic region located in the N-terminal part, while C-terminal proteoforms were enriched in the soluble aggregate fraction. Although soluble aggregates are the most toxic forms of Aβ [47], additional investigation is needed to determine the biological function and clinical relevance of all these proteoforms in AD.

Cystatin C (CysC) is a cysteine proteinase inhibitor used as a proxy for kidney function. In a study of 500 human plasma samples from a control population, two CysC proteoforms were reported as consequence of N-terminal truncations, one missing the N-terminal serine (des-S) and another lacking three N-terminal residues (des-SSP) [48,49]. Quantitative mass spectrometry immunoassays of CysC proteoforms in a cohort of patients with chronic kidney disease (CKD) with or without type 2 diabetes mellitus, showed that the levels of des-S and des-SSP were greater in the diabetic CKD group and could be used as markers for CDK progression [50]. A third CysC

proteoform, a N-terminal truncation lacking four terminal residues (des-SSPG), was reported in a proteomic study of intracystic fluid from a pilocytic astrocytoma pediatric brain tumor [49].

Removal of the initiator methionine and N-terminal acetylation are highly conserved and widespread modifications that mostly occur co-translationally. Mutations in the *NAA10* gene, which encodes N-alpha-acetyltransferase 10 (ARD1), which has the highest substrate repertoire, lead to NAA10 syndrome. The first described NAA10 mutation was the Ser37Pro mutation [51], which impairs the interaction of NAA10 with Naa15 and Naa50 and leads to decreased N-terminal acetylation of several substrates [52]. This NAA10 mutation was linked to the lethal Ogden syndrome. NAA10 also has a role in high-fat diet (HFD)-induced obesity [53]. Naa10 knockout (KO) mice fed a HFD showed reduced weight gain and reduced body fat compared with wild-type mice. Further studies revealed that a Naa10 KO resulted in increased expression of thermogenic genes and this was linked to the absence of N-terminal acetylation of Pgc1α. The N terminus of Pgc1α is normally acetylated by NAA10, which prevents its interaction with PPARγ. Given that this interaction normally leads to increased expression of thermogenic genes, the absence of the N-terminal acetyl group explains why a NAA10 KO results in an increased expression of such genes [53]. In addition, NAA10 overexpression is also found in several cancers, where it correlates with low survival rate and the aggressiveness of tumors [54].

Post-translational modifications on histones are part of the epigenetic system that regulates chromatin structure and condensation, and DNA replication, repair, and transcription. Loss of N-terminal acetylation of histone H4, due to low levels of N-terminal acetyltransferase 4 (Nat4), was directly related to longevity driven by calorie restriction. Calorie restriction reduces the levels of Nat4, resulting in reduced levels of N-terminal acetylated H4, which induces specific stress-response genes [55].

The methyl CpG-binding protein 2 (MeCp2) is a chromatin-associated transcription factor with two described isoforms, MeCp2e1 (MeCp2B) and MeCp2e2 (MeCp2A), due to the switch between translation start codons in exons 1 and 2, respectively. Among the mutations in the *MECP2* gene that lead to clinical manifestations of Rett Syndrome, there is a missense mutation in exon 1 (c.5C>T) that results in a Ala2Val mutation [56]. This residue is conserved throughout evolution and located in a region with multiple binding sites for the transcription factor SP1. Sheikh *et al.* found that the mutated proteoform MeCp2e1-Ala2Val was less stable than the wild-type protein due to reduced initiator methionine cleavage and reduced N-terminal acetylation of either methionine or valine [57].

The N-terminal acetylation of α-synuclein (α-Syn or NACP) was extensively characterized in view of the functional and pathological role of this protein [58]. Intracellular inclusions of aggregated and misfolded α-Syn comprising the Lewy bodies are a hallmark of Parkinson's and other neurodegenerative diseases [59]. N-terminal acetylation stabilizes the N-terminal domain helicity, which confers increased affinity for membrane interaction [60,61], and decreases α-Syn aggregation by blocking the formation of hydrogen bonds, which contribute to α-Syn oligomerization [62–65].

Concluding Remarks

N-terminal proteoforms arise from different transcription and translation-related mechanisms, in addition to mutations leading to differences in transcript sequences. Besides increasing the overall chemical complexity of a proteome, some N-terminal proteoforms have been linked to diseases, as we have discussed (see Outstanding Questions). The roles of such proteoforms in a given disease may also be subject to other cellular factors. For instance, the N-terminal proteoform of p53 arising from translation starting at methionine-40 causes aggregation of

Outstanding Questions

To what extent is the collection of N-terminal proteoforms found in human cell lines or even primary cells identical to that found in human tissues?

Are there different N-terminal proteoforms, originating from the same gene, in different tissues?

Which N-terminal proteoforms have critical roles in biological or pathological processes, and which are simply the result of differences in transcription, splicing, translation, and modifications and are not causing any harm to the cells or tissues?

Which (external) factors cause the generation of N-terminal proteoforms?

Box 3. Global and Functional Analysis of N- and C-Terminal Proteoforms

A recent study in mouse brain set out to map proteoforms in a global way. By combining ribosome footprinting and liquid chromatography tandem mass spectrometry (LC-MS/MS), N-terminal proteoforms resulting from alternative translation initiation and C-terminal proteoforms resulting from stop codon read-through were identified. Interestingly, these proteoforms were found to be differentially expressed in different cell types (neuron and glia cells) and upon neuron depolarization [66]. Among the proteoforms detected, an Uchl1 proteoform lacking 15 N-terminal amino acids was identified. Uchl1 is a deubiquitinating protease with protective roles in neurodegeneration and AD. For the C-terminal proteoform of AQP4, differential regulation of the normal and extended proteins in pathological conditions (involving gliosis) was found [66].

full-length p53, removing its tumor suppressor role in endometrial carcinoma cells [38], whereas it exerts tumor suppressor activity in HCC [39].

The fact that N-terminal proteoforms appear to hold important functions in different diseases may open avenues for new therapies. As an example, specific inhibition of NFATc1-α with a tumorigenic signaling role, and not of NFATc1- β with tumor-suppressive abilities, appears to be an interesting route for future anticancer drug design [24]. In addition, some N-terminal proteoforms could be seen as potential biomarkers and, thus, one could envision applied methods, such as targeted proteomics and antibodies, specifically recognizing N-terminal epitopes for diagnostic purposes.

Of note, most N-terminal proteoforms were discovered in studies focused on single proteins. To date, there is only one widescale study that maps N-terminal proteoforms in mouse brain and their relation to functionality (Box 3 and [66]). Thus, this emerging field appears to be gaining attention. Along this line, it is somewhat disappointing that the term 'proteoform' is not yet widely used, rather one mentions isoforms. Therefore, it is not straightforward to mine public data for functionalities of N-terminal proteoforms and, thus, we encourage researchers to consider when to use the term 'proteoform' and when to use the term 'isoform'. In fact, in studies reporting on peptide and protein sequences, we do recommend the use of the term 'proteoform' as defined by Smith and Kelleher [2] because it embraces all different chemical structures of the protein products from a single gene. As such, it includes possibly different proteins that originate from genetic variation, RNA transcript variation, and protein modification. This is different from the term 'isoform' as used by popular protein sequence databases, such as UniProt. This term only includes alternative protein sequences arising from alternative promoter usage, alternative splicing, alternative translation initiation and/or ribosomal frameshifting and, hence, does not consider protein modifications.

Acknowledgments

K.G. acknowledges support from EPIC-XS, project number 823839, funded by the Horizon 2020 Programme of the European Union and from The Research Foundation - Flanders (FWO), project number G008018N.

References

1. Cunningham, F. *et al.* (2019) Ensembl 2019. *Nucleic Acids Res.* 47, D745–D751
2. Smith, L.M. *et al.* (2013) Proteoform: a single term describing protein complexity. *Nat. Methods* 10, 186–187
3. Van Damme, P. *et al.* (2014) N-terminal proteomics and ribosome profiling provide a comprehensive view of the alternative translation initiation landscape in mice and men. *Mol. Cell. Proteomics* 13, 1245–1261
4. Gawron, D. *et al.* (2014) The proteome under translational control. *Proteomics* 14, 2647–2662
5. Jackson, R.J. *et al.* (2010) The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat. Rev. Mol. Cell Biol.* 11, 113–127
6. Vagner, S. *et al.* (2001) Irresistible IRES. Attracting the translation machinery to internal ribosome entry sites. *EMBO Rep.* 2, 893–898
7. Kozak, M. (1991) Structural features in eukaryotic mRNAs that modulate the initiation of translation. *J. Biol. Chem.* 266, 19867–19870
8. Kozak, M. (1987) Effects of intercistronic length on the efficiency of reinitiation by eukaryotic ribosomes. *Mol. Cell. Biol.* 7, 3438–3445
9. Gaba, A. *et al.* (2001) Physical evidence for distinct mechanisms of translational control by upstream open reading frames. *EMBO J.* 20, 6453–6463
10. Schoenberg, D.R. and Maquat, L.E. (2009) Re-capping the message. *Trends Biochem. Sci.* 34, 435–442
11. Alberts, B. *et al.* (2008) *Molecular Biology of the Cell* (5th edn), Garland Science
12. Yang, X. *et al.* (2016) Widespread expansion of protein interaction capabilities by alternative splicing. *Cell* 164, 805–817

13. Bazykin, G.A. and Kochetov, A.V. (2011) Alternative translation start sites are conserved in eukaryotic genomes. *Nucleic Acids Res.* 39, 567–577
14. Lange, P.F. *et al.* (2014) Annotating N termini for the human proteome project: N termini and Nalpha-acetylation status differentiate stable cleaved protein species from degradation remnants in the human erythrocyte proteome. *J. Proteome Res.* 13, 2028–2044
15. Kazak, L. *et al.* (2013) Alternative translation initiation augments the human mitochondrial proteome. *Nucleic Acids Res.* 41, 2354–2369
16. Kobayashi, R. *et al.* (2009) Targeted mass spectrometric analysis of N-terminally truncated isoforms generated via alternative translation initiation. *FEBS Lett.* 583, 2441–2445
17. Gawron, D. *et al.* (2016) Positional proteomics reveals differences in N–terminal proteoform stability. *Mol. Syst. Biol.* 12, 858
18. Calligaris, R. *et al.* (1995) Alternative translation initiation site usage results in two functionally distinct forms of the GATA-1 transcription factor. *Proc. Natl. Acad. Sci. U. S. A.* 92, 11598–11602
19. Thomas, D. *et al.* (2008) Alternative translation initiation in rat brain yields K2P2.1 potassium channels permeable to sodium. *Neuron* 58, 859–870
20. Claus, P. *et al.* (2003) Differential intranuclear localization of fibroblast growth factor-2 isoforms and specific interaction with the survival of motoneuron protein. *J. Biol. Chem.* 278, 479–485
21. Aebersold, R. *et al.* (2018) How many human proteoforms are there? *Nat. Chem. Biol.* 14, 206–214
22. Chuvpilo, S. *et al.* (2002) Autoregulation of NFATc1/A expression facilitates effector T cells to escape from rapid apoptosis. *Immunity* 16, 881–895
23. Serfling, E. *et al.* (2000) The role of NF-AT transcription factors in T cell activation and differentiation. *Biochim. Biophys. Acta* 1498, 1–18
24. Lucena, P.I. *et al.* (2016) NFAT2 isoforms differentially regulate gene expression, cell death, and transformation through alternative N-terminal domains. *Mol. Cell. Biol.* 36, 119–131
25. Molina-Cerrillo, J. *et al.* (2017) Bruton's tyrosine kinase (BTK) as a promising target in solid tumors. *Cancer Treat. Rev.* 58, 41–50
26. Guryanova, O.A. *et al.* (2011) Nonreceptor tyrosine kinase BMX maintains self-renewal and tumorigenic potential of glioblastoma stem cells by activating STAT3. *Cancer Cell* 19, 498–511
27. Holopainen, T. *et al.* (2012) Deletion of the endothelial Bmx tyrosine kinase decreases tumor angiogenesis and growth. *Cancer Res.* 72, 3512–3521
28. Shi, Y. *et al.* (2018) Ibrutinib inactivates BMX–STAT3 in glioma stem cells to impair malignant growth and radioresistance. *Sci. Transl. Med.* 10, eaah6816
29. Wang, Y. *et al.* (2017) A novel BMX variant promotes tumor cell growth and migration in lung adenocarcinoma. *Oncotarget* 8, 33405–33415
30. Magnuson, B. *et al.* (2012) Regulation and function of ribosomal protein S6 kinase (S6K) within mTOR signalling networks. *Biochem. J.* 441, 1–21
31. Zhang, J. *et al.* (2018) The p85 isoform of the kinase S6K1 functions as a secreted oncoprotein to facilitate cell migration and tumor growth. *Sci. Signal.* 11, eaao1052
32. Kwon, H. *et al.* (2018) Alternative translation initiation of Caveolin-2 desensitizes insulin signaling through dephosphorylation of insulin receptor by PTP1B and causes insulin resistance. *Biochim. Biophys. Acta Mol. basis Dis.* 1864, 2169–2182
33. Zhang, Y. and Xiang, Y. (2016) Molecular and cellular basis for the unique functioning of Nrf1, an indispensable transcription factor for maintaining cell homoeostasis and organ integrity. *Biochem. J.* 473, 961–1000
34. Wang, M. *et al.* (2019) Distinct isoforms of Nrf1 diversely regulate different subsets of its cognate target genes. *Sci. Rep.* 9, 2960
35. Marconi, C. *et al.* (2017) 5'UTR point substitutions and N–terminal truncating mutations of ANKRD26 in acute myeloid leukemia. *J. Hematol. Oncol.* 10, 18
36. Banno, K. *et al.* (2016) Systematic cellular disease models reveal synergistic interaction of Trisomy 21 and GATA1 mutations in hematopoietic abnormalities. *Cell Rep.* 15, 1228–1241
37. Bykov, V.J.N. *et al.* (2018) Targeting mutant p53 for efficient cancer therapy. *Nat. Rev. Cancer* 18, 89–102
38. Melo Dos Santos, N. *et al.* (2019) Loss of the p53 transactivation domain results in high amyloid aggregation of the Delta40p53 isoform in endometrial carcinoma cells. *J. Biol. Chem.* 294, 9430–9439
39. Ota, A. *et al.* (2017) Delta40p53alpha suppresses tumor cell proliferation and induces cellular senescence in hepatocellular carcinoma cells. *J. Cell Sci.* 130, 614–625
40. Silva, J.L. *et al.* (2018) Targeting the prion-like aggregation of mutant p53 to combat cancer. *Acc. Chem. Res.* 51, 181–190
41. Hopkins, B.D. *et al.* (2013) A secreted PTEN phosphatase that enters cells to alter signaling and survival. *Science* 341, 399–402
42. Wang, H. *et al.* (2015) Relevance and therapeutic possibility of PTEN-long in renal cell carcinoma. *PLoS ONE* 10, e114250
43. von Bohlen, A.E. *et al.* (2017) A mutation creating an upstream initiation codon in the SOX9 5' UTR causes acampomelic campomelic dysplasia. *Mol. Genet. Genomic Med.* 5, 261–268
44. Fu, Y. *et al.* (2015) A mutation creating an out-of-frame alternative translation initiation site in the GRHPR 5'UTR causing primary hyperoxaluria type II. *Clin. Genet.* 88, 494–498
45. Goyama, S. *et al.* (2019) Alternative translation initiation generates the N-terminal truncated form of RUNX1 that retains hematopoietic activity. *Exp. Hematol.* 72, 27–35
46. Wildburger, N.C. *et al.* (2017) Diversity of amyloid-beta proteoforms in the Alzheimer's disease brain. *Sci. Rep.* 7, 9520
47. McLean, C.A. *et al.* (1999) Soluble pool of Abeta amyloid as a determinant of severity of neurodegeneration in Alzheimer's disease. *Ann. Neurol.* 46, 860–866
48. Trenchevska, O. *et al.* (2014) Delineation of concentration ranges and longitudinal changes of human plasma protein variants. *PLoS ONE* 9, e100713
49. Inserra, I. *et al.* (2014) Proteomic study of pilocytic astrocytoma pediatric brain tumor intracystic fluid. *J. Proteome Res.* 13, 4594–4606
50. Yassine, H.N. *et al.* (2016) The association of plasma cystatin C proteoforms with diabetic chronic kidney disease. *Proteome Sci.* 14, 7
51. Rope, A.F. *et al.* (2011) Using VAAST to identify an X-linked disorder resulting in lethality in male infants due to N-terminal acetyltransferase deficiency. *Am. J. Hum. Genet.* 89, 28–43
52. Myklebust, L.M. *et al.* (2015) Biochemical and cellular analysis of Ogden syndrome reveals downstream Nt-acetylation defects. *Hum. Mol. Genet.* 24, 1956–1976
53. Lee, C.C. *et al.* (2019) Naa10p inhibits beige adipocyte-mediated thermogenesis through N-alpha–acetylation of Pgc1alpha. Mol. *Cell* 76, 500–515
54. Kalvik, T.V. and Arnesen, T. (2013) Protein N-terminal acetyltransferases in cancer. *Oncogene* 32, 269–276
55. Molina-Serrano, D. *et al.* (2016) Loss of Nat4 and its associated histone H4 N-terminal acetylation mediates calorie restriction-induced longevity. *EMBO Rep.* 17, 1829–1843
56. Saunders, C.J. *et al.* (2009) Novel exon 1 mutations in MECP2 implicate isoform MeCP2_e1 in classical Rett syndrome. *Am. J. Med. Genet. A* 149A, 1019–1023
57. Sheikh, T.I. *et al.* (2017) MeCP2_E1 N-terminal modifications affect its degradation rate and are disrupted by the Ala2Val Rett mutation. *Hum. Mol. Genet.* 26, 4132–4141
58. Burre, J. (2015) The synaptic function of alpha-synuclein. *J. Parkinsons Dis.* 5, 699–713
59. Goedert, M. (2001) Alpha-synuclein and neurodegenerative diseases. *Nat. Rev. Neurosci.* 2, 492–501
60. Maltsev, A.S. *et al.* (2012) Impact of N-terminal acetylation of alpha-synuclein on its random coil and lipid binding properties. *Biochemistry* 51, 5004–5013
61. Dikiy, I. and Eliezer, D. (2014) N-terminal acetylation stabilizes N-terminal helicity in lipid- and micelle-bound alpha-synuclein and increases its affinity for physiological membranes. *J. Biol. Chem.* 289, 3652–3665
62. Rossetti, G. *et al.* (2016) Conformational ensemble of human alpha-synuclein physiological form predicted by molecular simulations. *Phys. Chem. Chem. Phys.* 18, 5702–5706
63. Iyer, A. *et al.* (2016) The impact of N-terminal acetylation of alpha-synuclein on phospholipid membrane binding and fibril structure. *J. Biol. Chem.* 291, 21110–21122
64. Kang, L. *et al.* (2012) N-terminal acetylation of alpha-synuclein induces increased transient helical propensity and decreased aggregation rates in the intrinsically disordered monomer. *Protein Sci.* 21, 911–917

65. Bu, B. *et al.* (2017) N-terminal acetylation preserves alpha-synuclein from oligomerization by blocking intermolecular hydrogen bonds. *ACS Chem. Neurosci.* 8, 2145–2151
66. Sapkota, D. *et al.* (2019) Cell-type-specific profiling of alternative translation identifies regulated protein isoform variation in the mouse brain. *Cell Rep.* 26, 594–607
67. Komar, A.A. and Hatzoglou, M. (2011) Cellular IRES-mediated translation: the war of ITAFs in pathophysiological states. *Cell Cycle* 10, 229–240
68. Komar, A.A. and Hatzoglou, M. (2005) Internal ribosome entry sites in cellular mRNAs: mystery of their existence. *J. Biol. Chem.* 280, 23425–23428
69. King, H.A. *et al.* (2010) The role of IRES trans-acting factors in regulating translation initiation. *Biochem. Soc. Trans.* 38, 1581–1586
70. Beaudoin, M.E. *et al.* (2008) Regulating amyloid precursor protein synthesis through an internal ribosomal entry site. *Nucleic Acids Res.* 36, 6835–6847
71. Walters, B. and Thompson, S.R. (2016) Cap-independent translational control of carcinogenesis. *Front. Oncol.* 6, 128