Alternative translation initiation unraveled by N-terminomics and ribosome profiling

Daria Fijałkowska
Reference:

Dutch translation of the title:
Alternatieve translatie-initiatie ontrafeld door middel van N-terminomics en ribosoom profilering.

Cover illustration:
An artist’s rendition of a ribosome. Credit: C. Bickle/Science.

Printing:
Ridderprint BV, the Netherlands

ISBN:
978-94-6299-512-3

The author and the promoter give the authorisation to consult and copy parts of this work for personal use only. Every other use is subject to the copyright laws. Permission to reproduce any material contained in this work should be obtained from the author.

Author                                         Promoter
Daria Fijałkowska                              Petra Van Damme
Academic year 2016-2017

Alternative translation initiation unraveled by N-terminomics and ribosome profiling

Thesis submitted to fulfill the requirements for the degree of "DOCTOR OF HEALTH SCIENCES"

Daria Fijałkowska

Promoter: Prof. dr. Petra Van Damme
Co-promoter: Prof. dr. Kris Gevaert
VIB Medical Biotechnology Center
Department of Biochemistry
Faculty of Medicine and Health Sciences
Ghent University, Belgium
Members of the examination committee:

Prof. dr. Elfride De Baere  
*Chair*  
Department of Pediatrics and Medical Genetics, Ghent University  
Center for Medical Genetics Ghent, Ghent University Hospital

Prof. dr. Petra Van Damme  
*Promoter*  
VIB Medical Biotechnology Center  
Department of Biochemistry, Ghent University

Prof. dr. Kris Gevaert  
*Co-promoter*  
VIB Medical Biotechnology Center  
Department of Biochemistry, Ghent University

Dr. Maarten Dhaenens  
Department of Pharmaceutics, Ghent University

Prof. dr. Kim De Keersmaecker  
Laboratory for Disease Mechanisms in Cancer, KU Leuven

Dr. Audrey Michel  
School of Biochemistry and Cell Biology, University College Cork

Prof. dr. ir. Kathleen De Preter  
Department of Pediatrics and Medical Genetics, Ghent University  
Center for Medical Genetics Ghent, Ghent University Hospital

Prof. dr. ir. Jo Vandesompele  
Department of Pediatrics and Medical Genetics, Ghent University  
Center for Medical Genetics Ghent, Ghent University Hospital
To my parents and Igor –

in tribute to your inquisitive minds

and loving hearts

“Data! data! data!” he cried impatiently.

"I can't make bricks without clay.”

- Arthur Conan Doyle, The Adventures of Sherlock Holmes
TABLE OF CONTENTS
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>TABLE OF CONTENTS</td>
<td>i</td>
</tr>
<tr>
<td>LIST OF ABBREVIATIONS</td>
<td>v</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>OBJECTIVES</td>
<td>27</td>
</tr>
<tr>
<td>RESULTS</td>
<td>33</td>
</tr>
<tr>
<td>Chapter I: Discovery of novel N-terminal proteoforms</td>
<td>33</td>
</tr>
<tr>
<td>Chapter II: Assessment of N-terminal proteoform stability</td>
<td>67</td>
</tr>
<tr>
<td>Chapter III: Deep integration of OMICS data</td>
<td>105</td>
</tr>
<tr>
<td>Chapter IV: Novel insights into translational control</td>
<td>127</td>
</tr>
<tr>
<td>GENERAL DISCUSSION</td>
<td>163</td>
</tr>
<tr>
<td>BROAD RELEVANCE AND FUTURE PERSPECTIVES</td>
<td>179</td>
</tr>
<tr>
<td>SUMMARY &amp; SAMENVATTING</td>
<td>187</td>
</tr>
<tr>
<td>CURRICULUM VITAE</td>
<td>197</td>
</tr>
<tr>
<td>APPENDIX</td>
<td>203</td>
</tr>
<tr>
<td>ADDITIONAL PUBLICATIONS</td>
<td>223</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>235</td>
</tr>
</tbody>
</table>
LIST OF ABBREVIATIONS
# LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>6-FT</td>
<td>6-frame genome translation</td>
</tr>
<tr>
<td>altORF</td>
<td>alternative open reading frame</td>
</tr>
<tr>
<td>aTIS</td>
<td>alternative/annotated translation initiation site (^1)</td>
</tr>
<tr>
<td>bp</td>
<td>base pair</td>
</tr>
<tr>
<td>CCDS</td>
<td>consensus coding sequence</td>
</tr>
<tr>
<td>CDS</td>
<td>coding sequence</td>
</tr>
<tr>
<td>CHX</td>
<td>cycloheximide</td>
</tr>
<tr>
<td>COFRADIC</td>
<td>COMbined FRActional Diagonal Chromatography</td>
</tr>
<tr>
<td>CORUM</td>
<td>comprehensive resource of mammalian protein complexes</td>
</tr>
<tr>
<td>dbTIS</td>
<td>database-annotated translation initiation site</td>
</tr>
<tr>
<td>DMEM</td>
<td>Dulbecco’s modified Eagle’s medium</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>dTIS</td>
<td>downstream translation initiation site</td>
</tr>
<tr>
<td>eIF</td>
<td>eukaryotic initiation factor</td>
</tr>
<tr>
<td>emPAI</td>
<td>exponentially modified protein abundance index</td>
</tr>
<tr>
<td>FBS</td>
<td>fetal bovine serum</td>
</tr>
<tr>
<td>FDR</td>
<td>false discovery rate</td>
</tr>
<tr>
<td>GEO</td>
<td>Gene Expression Omnibus</td>
</tr>
<tr>
<td>GlutaMAX</td>
<td>alanyl-L-glutamine dipeptide</td>
</tr>
<tr>
<td>GO</td>
<td>gene ontology</td>
</tr>
<tr>
<td>GRAVY</td>
<td>grand average of hydropathy</td>
</tr>
<tr>
<td>GTI-seq</td>
<td>global translation initiation sequencing</td>
</tr>
<tr>
<td>H2G2</td>
<td>HitchHikers Guide to the Genome</td>
</tr>
<tr>
<td>HARR</td>
<td>harringtonine</td>
</tr>
<tr>
<td>HFF</td>
<td>human foreskin fibroblast</td>
</tr>
<tr>
<td>HiRIEF</td>
<td>high resolution isoelectric focusing</td>
</tr>
<tr>
<td>iBAQ</td>
<td>intensity-based absolute-protein-quantification</td>
</tr>
<tr>
<td>IMDM</td>
<td>Iscove’s Modified Dulbecco’s Medium</td>
</tr>
</tbody>
</table>

\(^1\) In the Introduction and Results Chapter 1-3 used to denote alternative TIS; in Results Chapter 4 used to indicate annotated TIS.
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>iMet</td>
<td>initiator methionine</td>
</tr>
<tr>
<td>iMet-tRNA</td>
<td>initiator methionyl transfer RNA</td>
</tr>
<tr>
<td>IRES</td>
<td>internal ribosomal entry site</td>
</tr>
<tr>
<td>KEGG</td>
<td>Kyoto Encyclopedia of Genes and Genomes</td>
</tr>
<tr>
<td>KO</td>
<td>knockout</td>
</tr>
<tr>
<td>LC</td>
<td>liquid chromatography</td>
</tr>
<tr>
<td>lincRNA</td>
<td>long/large intergenic non-coding RNA</td>
</tr>
<tr>
<td>LTM</td>
<td>lactimidomycin</td>
</tr>
<tr>
<td>mESC</td>
<td>mouse embryonic stem cell</td>
</tr>
<tr>
<td>MetAP</td>
<td>methionine aminopeptidase</td>
</tr>
<tr>
<td>miRNA</td>
<td>microRNA</td>
</tr>
<tr>
<td>mRNA-seq</td>
<td>mRNA sequencing</td>
</tr>
<tr>
<td>MS</td>
<td>mass spectrometry</td>
</tr>
<tr>
<td>MS/MS</td>
<td>tandem mass spectrometry</td>
</tr>
<tr>
<td>mTP</td>
<td>mitochondrial targeting peptide</td>
</tr>
<tr>
<td>NAT</td>
<td>N-terminal acetyltransferase</td>
</tr>
<tr>
<td>ncRNA</td>
<td>non-coding RNA</td>
</tr>
<tr>
<td>NMD</td>
<td>nonsense-mediated decay</td>
</tr>
<tr>
<td>NSAF</td>
<td>normalized spectral abundance factor</td>
</tr>
<tr>
<td>nt</td>
<td>nucleotide</td>
</tr>
<tr>
<td>Nt-acetylation</td>
<td>N-terminal acetylation</td>
</tr>
<tr>
<td>PABP</td>
<td>poly(A)-binding protein</td>
</tr>
<tr>
<td>PAGE</td>
<td>polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>Pfam</td>
<td>(database of) protein families</td>
</tr>
<tr>
<td>pl</td>
<td>isoelectric point</td>
</tr>
<tr>
<td>PIC</td>
<td>pre-initiation complex</td>
</tr>
<tr>
<td>PRIDE</td>
<td>PRoteomics IDEntifications (database)</td>
</tr>
<tr>
<td>pSILAC</td>
<td>pulsed stable isotope labelling with amino acids in cell culture</td>
</tr>
<tr>
<td>PSM</td>
<td>peptide-to-spectrum match</td>
</tr>
<tr>
<td>PTC</td>
<td>premature termination codon</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>PUNCH-P</td>
<td>puromycin-associated nascent chain proteomics</td>
</tr>
<tr>
<td>qPCR</td>
<td>quantitative PCR</td>
</tr>
<tr>
<td>ribo-seq</td>
<td>ribosome profiling sequencing</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>RNA-seq</td>
<td>RNA sequencing</td>
</tr>
<tr>
<td>RPF</td>
<td>ribosome-protected fragment</td>
</tr>
<tr>
<td>RPMI</td>
<td>Roswell Park Memorial Institute (medium)</td>
</tr>
<tr>
<td>rRNA</td>
<td>ribosomal RNA</td>
</tr>
<tr>
<td>RT</td>
<td>reverse transcription</td>
</tr>
<tr>
<td>SCX</td>
<td>strong cation exchange</td>
</tr>
<tr>
<td>SILAC</td>
<td>stable isotope labelling with amino acids in cell culture</td>
</tr>
<tr>
<td>siRNA</td>
<td>small interfering RNA</td>
</tr>
<tr>
<td>SNP</td>
<td>single nucleotide polymorphism</td>
</tr>
<tr>
<td>snRNA</td>
<td>small nuclear RNA</td>
</tr>
<tr>
<td>sORF</td>
<td>short open reading frame</td>
</tr>
<tr>
<td>SP</td>
<td>signal peptide</td>
</tr>
<tr>
<td>STAR</td>
<td>spliced transcripts alignment to a reference</td>
</tr>
<tr>
<td>TE</td>
<td>translation efficiency</td>
</tr>
<tr>
<td>TIS</td>
<td>translation initiation site</td>
</tr>
<tr>
<td>TrEMBL</td>
<td>Translated EMBL Nucleotide Sequence Data Library</td>
</tr>
<tr>
<td>tRNA</td>
<td>transfer RNA</td>
</tr>
<tr>
<td>UniProtKB</td>
<td>Universal Protein Knowledgebase</td>
</tr>
<tr>
<td>u-oORF</td>
<td>upstream-overlapping open reading frame</td>
</tr>
<tr>
<td>uORF</td>
<td>upstream open reading frame</td>
</tr>
<tr>
<td>UPR</td>
<td>unfolded protein response</td>
</tr>
<tr>
<td>uTIS</td>
<td>upstream translation initiation site</td>
</tr>
<tr>
<td>UTR</td>
<td>untranslated region</td>
</tr>
</tbody>
</table>
INTRODUCTION
Adapted from:

**Gawron D, Gevaert K, Van Damme P (2014). The proteome under translational control.**
Proteomics, 2014: 14 p. 2647-2662
INTRODUCTION

Abstract

A single eukaryotic gene can give rise to a variety of protein forms (proteoforms) as a result of genetic variation and multi-level regulation of gene expression. In addition to alternative splicing, an increasing line of evidence shows that alternative translation contributes to the overall complexity of proteomes. Identifying the repertoire of proteins and micropeptides expressed by alternative selection of (near-)cognate translation initiation sites and different reading frames however remains challenging with contemporary proteomics. Mass spectrometry enabled identification of proteoforms is expected to benefit from transcriptome and translatome data by the creation of customised and sample-specific protein sequence databases. Here, we focus on contemporary integrative omics approaches that complement proteomics with DNA and/or RNA oriented technologies to elucidate the mechanisms of translational control. Together, these technologies enable to map the translation (initiation) landscape and more comprehensively define the inventory of proteoforms raised upon alternative translation, thus assisting in the (re-)annotation of genomes.
Introduction

Gene expression, especially in higher eukaryotes, is subject to multilevel regulation including gene transcription, mRNA processing, mRNA stability, mRNA translation and finally, protein modification. Contemporary mass spectrometry (MS) based proteomics is indispensable for understanding the outcome of gene expression by providing information complementary to genome and transcript sequencing [1, 2]. The continuous improvement of MS instrumentation and procedures for processing proteome samples lead to a more comprehensive mapping of proteomes in less time [3, 4]. Judged from recently published yeast and mammalian proteome studies [3-6], a complete eukaryotic proteome could soon become accessible, making proteomics a standard practice for gene expression profiling.

Integration of proteomics and transcriptomics for comprehensive gene expression profiling

Contemporary proteomic studies provided estimates on the number of proteins expressed by different cell types from the available collection of about 20,300 human protein-coding genes (according to the GRCh38 GENCODE 20 annotation). Very recently, comprehensive proteome analyses of 30 histologically normal human cell and tissue types resulted in the identification of proteins encoded by 17,294 human genes [6]. From this vast collection, 1,537 genes were found to be tissue-specific [6]. Similarly, the ProteomicDB repository provided proteomics evidence for 18,097 human genes [5]. Overall, a saturation of proteome coverage at 16,000 – 17,000 proteins was demonstrated, consistent with RNA-seq coverage of the transcriptome. Surveying ProteomeDB led to the postulation that the human core proteome consists of at least 10,000 – 12,000 expressed gene products [5] in line with several earlier studies [7-9]. These results indicate that the expressed proteome is omnipresent and the variability between cell types is often achieved by tuning expression levels rather than by the absence or presence of specific proteins [7, 10]. Hence, a quantitative comparison of proteomes under study is needed as protein levels shape cell phenotypes and functions. Recent estimates of protein absolute abundance report on similar dynamic ranges of detectable proteins in mouse fibroblasts [11, 12] and human cells [4, 7-9, 13] spanning six to seven orders of magnitude, with median protein abundance estimated at 10,000 to 21,000 protein copies per human cell [4, 8, 9]. Emerging lines of evidence however, demonstrate that the global proteome diversity might be much larger than originally anticipated and that proteomics has not yet delivered the complete set of expressed molecular forms of protein products - proteoforms [14, 15] (Fig. 1).

MS-based protein identification depends on the completeness of protein sequence databases and is thus largely restricted to annotated proteins. Only now we begin to understand just how many proteoforms can result from allelic variations, different transcripts raised by (epigenetically-controlled) alternative promoter usage, alternative splicing and mRNA editing, upon alternative translation (initiation) and various co- and post-
A single term – proteoforms – may be used to describe the molecular forms of proteins derived from individual genes, thus capturing the complete biological variability and all possible modifications of primary protein structure [14].

translational protein modifications (Fig. 1), amongst others. As an example, our recent N-terminomics studies on mouse and human proteins highlighted the underestimated prevalence of database non-annotated translation start sites [15-17]. Furthermore, N-terminal proteoforms were also found in yeast [18, 19] and fruit fly [20]. Depending on the cell or tissue type, we found that 10-20% of all identified protein N-termini result from alternative translation initiation events, point to incorrectly annotated protein start sites or to alternative splicing. In specific cases, (premature) protein termination may be circumvented by a stop codon readthrough mechanism [21, 22], recently shown to produce C-terminally elongated proteoforms in fly, yeast and human [21]. Ribosomal frameshifting was found to be another mechanism potentially leading to the generation of distinct proteins from one genomic locus [23, 24]. Further, mRNA-seq revealed the enormous diversity of splice junctions used and showed that tissue-specific splicing events occur [25]. Integration of transcriptome data from different sources further suggested that alternative splicing may affect about 95% of all multi-exonic human genes [25]. Importantly, cases of
non-synonymous and cell-specific genetic variation next to RNA editing can be detected by combining proteomics with sample-specific genome and transcriptome-derived database searches [2].

Hence, it is fair to state that all proteoforms expressed under different conditions in a cell type or tissue-specific manner are not yet annotated. An integration of proteomics and transcriptomics data to narrow the search space to a selected set of genes [16, 17, 26] may help to overcome this as we will discuss below.

Proteome data are increasingly integrated with other omics data, thereby opening up new possibilities to discover the relationship between the transcriptome and the proteome. Numerous studies showed that mRNA levels only moderately correlate with protein levels [2, 9, 11, 27, 28]. Instead, individual proteins appear to have a conserved intrinsic rate of mRNA to protein conversion [2, 5]. Also, the remodelling of proteomes in response to various perturbations seems to be largely orchestrated at the post-transcriptional level. In multiple model systems, changes in mRNA levels often poorly correlate with changes in protein levels [28-34], even more so in the parasitic protist Trypanosoma brucei where transcriptional control is lacking and functionally distinct proteins are simultaneously expressed from long polycistronic transcription units [35]. These findings suggest an important role for post-transcriptional regulation of gene expression, including mRNA decay, translational control of protein synthesis and protein degradation. Several integrative omics studies complemented with computational modelling were conducted to assess the accumulation of such effects. In Mycoplasma pneumoniae, regulation of gene expression was found to be largely decoupled from protein dynamics as high protein numbers and long protein half-lives effectively buffer against spikes in gene expression [28]. In T. brucei, translational control and mRNA stability were both found to vary over two order of magnitude and thus to contribute equally to gene expression [35]. In human cells, estimations of the impact of protein turnover on gene regulation [36] revealed that protein synthesis and degradation rates are conserved in functional protein classes and similar among most components of individual protein complexes. Nonetheless, the protein synthesis rate is extensively regulated upon cell differentiation whereas the protein degradation rate overall poorly contributes to protein expression control [36]. In summary, studies in both prokaryotes and eukaryotes show that mRNA translation rates are likely the predominant regulators of protein abundance [11, 28, 36, 37] and have an overall higher impact than transcriptional regulation of mRNA [11], protein turnover [11, 28, 36] and mRNA degradation [11, 37].

**Alternative translation increases the complexity of eukaryotic proteomes**

Molecular mechanisms of translation have been extensively studied [38-42] and showed that translation of most eukaryotic mRNA molecules is initiated by a scanning mechanism.
Mature mRNA molecules are recognised and circularised by a complex of eukaryotic initiation factors (eIFs) 4F, composed of eIF4E, 4A, 4G and 4B, accompanied by the poly(A)-binding protein (PABP). Simultaneously, the ternary complex (composed of an initiator methionyl-tRNA (iMet-tRNA) and eIF2) associates with the 40S ribosome subunit and eIFs 1, 1A, 3 and 5. In this way the pre-initiation complex (PIC) is formed. Once assembled, PIC is recruited to an activated mRNA molecule. The whole complex, referred to as the 48S initiation complex, begins scanning along the 5’ leader until an AUG triplet embedded in a Kozak consensus context is encountered. A series of interactions, initiated by eIF1 dissociation, triggers hydrolysis of GTP bound to eIF2, thereby promoting the release of the initiation factors. This allows the 60S ribosomal subunit to join and proceed to the elongation phase of translation. Synthesis of the polypeptide chain is terminated at stop codons, upon which ribosomal subunits can be recycled.

Translational control involves deviations from this canonical ribosomal scanning, such as leaky scanning, reinitiation and internal ribosome entry [41, 42]. It is presumed that most regulation is exerted during translation initiation [41] via the selection of different start sites, including AUG and near cognate start codons. As for the latter, triplets such as CUG or GUG may interact with the anticodon of iMet-tRNA by an incomplete Watson-Crick interaction [43, 44]. As such, protein biogenesis by ribosomes can generate multiple proteoforms from one mRNA molecule by utilizing alternative translation initiation sites and reading frames (Fig. 1).

Different in-frame initiation sites in the 5’ leader or inside the coding sequence may be selected, leading to increased complexity of proteomes. For instance, proteoforms with N-terminal extensions or truncations may serve different functions [17, 45]. Such N-terminally distinct proteins may lose or gain targeting signals and several proteins were thus reported to reside in different cellular organelles as a consequence of alternative translation initiation [46-51]. In particular, both N-terminally truncated and elongated proteoforms may acquire mitochondrial localisation due to the exposure of a full-length cryptic target signal [51] or by gaining a new targeting sequence at their N-terminus [49], respectively. Other truncated isoforms, which lost their mitochondrial targeting signal, may instead end up in the nucleus [43]. Several of these N-terminal proteoforms are iso-functional but active in different compartments [47, 48]. Others acquire altered properties such as altered sodium permeability in case of the N-terminally truncated form of the potassium channel, subfamily K, member 2 (K2P2.1), which changes neuron responsiveness in specific brain tissues and developmental stages [52]. Alternative translation initiation can also cause a loss of function, like the dominant negative effect of truncated CCAAT/enhancer binding protein beta (C/EBP β) [53], or can functionally modulate growth factors and oncogenes (including fibroblast growth factor (FGF) [45], vascular endothelial growth factor (VEGF) [54] and V-myc avian myelocytomatosis viral oncogene homolog (c-Myc) [55]). Moreover, N-terminal proteoforms may have different cellular stabilities as is the case for the opioid receptor mu 1 (OPRM1) [56]. All of these findings are further enforced by the recent demonstration that alternative
translation start sites at AUG and near-cognate codons are often conserved among eukaryotes, further hinting to a possible functional importance [57, 58].

**Tools for the comprehensive mapping of proteoforms**

Proteomics is indispensable to study the products of mRNA translation as it provides the ultimate evidence for the existence of proteins. Novel N-terminal proteoforms are most efficiently identified by enriching protein N-terminal peptides prior to LC-MS/MS analysis. Our lab pioneered such N-terminomics studies by introducing the N-terminal COCombined FRActional Diagonal Chromatography (COFRADIC) approach [59]. When combined with proteome labelling (e.g. SILAC) and mass tagging of *in vivo* primary amines, it becomes a powerful tool to study alternative translation, as N-termini together with their *in vivo* modification status directly hint to translation initiation events [18, 19, 60, 61] (Fig. 2A). Using the known rules of initiator methionine (iMet) processing by methionine aminopeptidases (MetAPs) and N-terminal acetylation, one of the most common cotranslational protein modifications in eukaryotes, N-terminal COFRADIC data allow to judge if N-terminal peptides can serve as proxies for mRNA translation events or rather are indicative of proteolysis [60, 62].

Of note is that the annotation of protein-coding genes is still incomplete as many such proteoforms are missing in databases, thus hindering comprehensive proteomics and N-terminomics studies. Gene models rely on *ab initio* gene prediction, cDNA sequence determination and sequence conservation amongst others and benefit from proteogenomics strategies to delineate protein-coding sequences and splice junctions [63-65]. Proteogenomics using 6-frame genome translation (6-FT) in combination with splice site predictions allowed for remarkable advances in decoding small genomes such as *A. thaliana* (i.e. by discovering 778 new protein-coding genes) [64]. On the other hand, reducing the complexity of the overall search space is deemed necessary for large genomes to reduce sequence redundancy and to avert the increased false discovery rate [65, 66]. Therefore, in organisms such as mouse, rat and human, further refinement of gene models may be facilitated by partitioning of 6-FT databases [66], accurate prediction of protein-coding loci [2, 65] or by confining the analyses to protein sequences derived from transcriptome databases and RNA-sequencing experiments [2, 26, 67]. Recently, Branca *et al.* [66] combined peptide fractionation by high resolution isoelectric focusing (HiRIEF) and rationalized database searches to achieve an extensive (over 70 fold) reduction of the 6-FT tryptic search space by implementing a isoelectric point (pI) restriction. Overall, their proteogenomics workflow led to the identification of 224 human and 122 mouse peptides as proxies for the expression of previously uncharacterized proteoforms. More specifically, 24 human and 22 mouse peptides pointed to N-terminal protein extensions, 5 human and 5 mouse peptides were derived from upstream open reading frames (uORFs) in mRNA 5’ leaders. Combined with numerous products of frame-shifting in both species, these hint to
Figure 2 – Identification and quantification of proteoforms. A. N-terminomics (i.e. N-terminomics).

B. Gene X

Condition 1

Condition 2

Gene Y

Translation efficiency (TE) = ribosome footprints / mRNA-seq reads

TE of coding sequence (CDS) vs. TE of upstream ORF (uORF)
terminal COFRADIC) combined with ribosome profiling and mRNA sequencing allow for the comprehensive identification of protein translation events [17, 61]. Differential protein labelling of free primary amines with heavy acetyl groups (highlighted in red), next to in vivo N-terminal (Nt)-acetylation (in black), results in blocking of all protein amino termini. Further, diagonal reverse-phase chromatography in combination with specific peptide modifications allows for the isolation of proteome-representative N-terminal peptides, excluding internal and carboxy-terminal peptides. As such, N-terminal COFRADIC allows to study translation initiation and N-terminal protein modifications in a quantitative manner.

Ribosome profiling is based on deep sequencing of ribosome-protected mRNA fragments and thereby enables the study of (alternative) translation with sub- to single-codon resolution [68-70]. Treating cells with translation inhibitors provides a snapshot of different translational events: cycloheximide (CHX) halts elongating ribosomes, thus reflecting the landscape of all translated mRNA regions; harringtonine (HARR) [69] and lactimidomycin (LTM) [70] immobilize initiating ribosomes and allow to map translation start sites. Polysomes (mRNA molecules with attached ribosomes) are digested to yield monosomes and are further isolated by density centrifugation. In parallel, the corresponding mRNA pool is subjected to random fragmentation. Subsequently, ribosome-protected footprints (~30 nt) and mRNA fragments are extracted by PAGE, incorporated into cDNA libraries, deep sequenced and mapped onto the reference genome. The accumulation of ribosomes on initiation sites (upon LTM or HARR treatment) allows delineating the exact translation frames, subsequently converted into customised protein databases that enable the MS-based discovery of yet uncharacterized proteoforms [16, 17].

B. Quantification of gene expression using ribo-seq and mRNA-seq. Ribosome footprints detected upon CHX treatment (blue) are used to quantify translation and thereby estimate gene expression levels. The frequency of ribosomal footprints vs. the overall expression level of mRNA (red) is indicative for the translation efficiency (TE). The right panel shows an upregulated TE of Gene X in Condition 2 as compared to Condition 1. The left panel depicts how TE may deviate between ORFs. Here, translational upregulation of two upstream ORFs (uORFs) in Condition 2 inhibits translation of the annotated protein-coding sequence (CDS) and thus a reduced TE of the CDS as compared to Condition 1.

alternative translation events including the use of non-AUG codons [66]. Another proteogenomic strategy was used by Kim et al. [6]. MS/MS spectra acquired from 30 human tissue types, unmapped to conventional protein databases, were searched against conceptual translation of genes and transcripts, databases of theoretical protein N-termini and predicted signal peptides. This analysis revealed 216 new (splice) variants caused by gene/protein/exon extension and incorporation of new exons. Furthermore, 198 new protein N-termini were identified, pointing to 3 upstream and 195 downstream translation initiation sites. Apart from refining known gene models, Kim et al. [6] identified translation evidence for 44 new ORFs including uORFs, several of which were previously reported [71, 72]. In addition, numerous peptides originating from the so-called pseudogenes and long noncoding RNAs were identified.
Recently, ribosome profiling (or ribo-seq) revolutionised the genome-wide annotation of protein-coding regions by deep sequencing of ribosome-protected mRNA fragments (Fig 2A-B). The comprehensive delineation of (alternative) translation frames by ribosome profiling can further be used to create customized, sample-oriented sequence databases which can serve as reference datasets for the proteome under study (Fig. 2A). Such integrative strategies were already successfully applied to cytomegalovirus, Trypanosoma, zebrafish, mouse and human samples [16, 17, 35, 73, 74] and enabled more efficient discovery of proteins and micropeptides and helped to overcome limitations in identifying N-terminal extensions and non-AUG initiated proteoforms. A recent N-terminomics study reported more than 1,700 alternative translation events in mouse and human [17], of which N-terminally truncated proteoforms represented the largest group with 1,231 and 465 downstream translation initiation sites (dTIS) detected in human and mouse, respectively. Supplementing the MS search space with protein products of reading frames identified by ribosome profiling enabled the identification of additional proteoforms not contained within the Swiss-Prot database [17]. Indeed, 17 N-terminally extended proteoforms, 2 overlapping out-of-frame uORFs and 4 truncated proteoforms were discovered in human. In mouse, 17 N-terminally extended proteoforms, one uORF and one new truncated proteoform were found. Overall, these studies provided evidence at the proteome level that near-cognate start codons, such as GUG, CUG and ACG, are decoded to methionine, instead of their canonical amino acid (valine, leucine and threonine respectively). Further, numerous dTIS pointed to translation initiation events raised upon ribosomal leaky scanning; a scanning model which assumes that ribosomes may omit start codons embedded in poor Kozak consensus sequences and continue scanning until a more favourable translation initiation site is encountered [40]. In many cases where the conventional protein start site was surrounded by a suboptimal Kozak context, translation was initiated at a downstream AUG codon, likely ensuring the effective expression of an isofunctional protein. Of note, alternative N-termini were often positioned in close proximity to the database-annotated initiation sites, consistent with the observations of Kim et al. [6]. Importantly, conservation of dTIS between vertebrate genomes was equally high as conservation of canonical initiation sites, while significantly distinct from a decoy set of putative translationally inactive AUG start sites (based on ribosome profiling and proteome analysis). An observation which was further supported by the identification of 200 orthologous pairs of mouse and human dTIS indicative N-termini [17].

**Translational control by short ORFs from a proteomics’ point-of-view**

As explained before, the canonical model of translation does not explain the full diversity of proteins expressed *in vivo*, and besides truncated or elongated N-terminal proteoforms, a remarkable high prevalence of short open reading frames (sORFs) has been postulated throughout eukaryotic genomes, inferred from both *in silico* predictions as well as studies of
translated RNA regions (further referred to as the translatome) [68, 69, 75-78]. Proteomics may help to verify whether such alternative translation events give rise to detectable peptides or merely serve a regulatory role, being rapidly degraded after emerging from the ribosome, if at all resulting in peptide production [79].

The diversity of the documented proteome has clearly expanded over the past years, amongst others by the discovery of sORFs that potentially give rise to polypeptides less than 100 amino acids long [80], classified as micropeptides [81]. sORFs are generally well conserved among species [77, 82] and can reside in many genomic locations [81]. Amongst others, they are found upstream, overlapping or downstream of mRNA coding sequences (CDS) and non-coding RNAs (ncRNA) as well as in intergenic regions. Upstream (overlapping) sORFs (uORFs) play an acknowledged role in steering eukaryotic translation. They reduce the accessibility of the canonical CDS to ribosomes by engaging them for translation in the near proximity of the translation initiation site of the canonical CDS, thus interfering with or completely attenuating synthesis of its corresponding protein [40, 42, 83] (see also Fig. 2B). Both leaky scanning and reinitiation are involved in this type of regulation and allow for multiple different initiation sites to be selected on the same mRNA molecule. Moreover, sORFs may induce ribosome stalling and activate nonsense-mediated decay (NMD) of mRNA molecules [42, 83] to further attenuate protein synthesis. In contrast to the well understood role of sORFs in translation, limited evidence is available on the coding potential of sORFs and the biological activity possibly exerted by their encoded short peptide products [80, 84, 85].

Contemporary proteomics may assist in validating the expression of sORF-derived peptides as new constituents of the proteome [6, 16, 17, 66, 67, 72, 74]. Several features of sORFs may however impede identification of sORF-derived peptides: such short peptides can be rapidly degraded [67, 86] and may thus generally be too scarce. Peptidomic workflows that employ tailored isolation procedures preventing unwanted protein/peptide degradation [71, 86, 87], may here be of use. However, the sequences of sORF-derived peptides are not annotated in sequence databases and alternatives are needed such as customized protein libraries based on translation of genomes [6, 66] or transcriptomes [6] supplemented with RNA-seq [67], using (s)ORF prediction [72] or ribo-seq data [16, 17, 74]. The majority of sORF translation is essentially presumed to be initiated at non-AUG codons [69] translated to methionine [16, 17], thus introducing a mismatch between mRNA and amino acid sequences that needs to be accounted for.

Recently, Slavoff et al. used MS to detect the expression of 90 peptides with an average size of 67 amino acids derived from human sORFs [67], 4 of which were reported in a previous study of Oyama et al. [71]. Conservation analysis in mammalian species indicated that sORFs located in known/expressed transcripts undergo stronger evolutionary selection than introns, but are less conserved than annotated CDSs. In line with previous reports, translation of 57% of sORFs was initiated at near-cognate codons. The expression levels and
range of sORF products (10-2,000 copies per cell) were low compared to human proteins [4, 8, 9]. Using immunofluorescence, the cytoplasmic expression of 10 sORFs localised in presumed 5’ untranslated regions (UTRs) was confirmed, thereby showing that uORFs potentially give rise to peptides while controlling translation of downstream genes. Interestingly, no products of sORFs located in 3’UTRs or internal frame-shifts were observed, the latter proposed to only be expressed from truncated splice variants. Moreover, their dataset provided scarce evidence for translation of long intergenic ncRNAs (lincRNAs), in contrast to the reported translational activity on lincRNAs detected by ribosome profiling [69] and peptide evidence of over 100 lincRNAs in the human proteome draft [6]. Nonetheless, it is important to note that a recent revised opinion on ribo-seq demonstrated that ribosome occupancy alone does not hold sufficient evidence of translation [79], therefore stressing the need for proteome-oriented approaches to univocally establish coding potential.

Additional evidence for short protein and micropeptide synthesis from alternative ORFs (altORFs), residing in UTRs or (partially) overlapping with canonical CDS in different reading frames, was recently provided by Vanderperre et al. [72], who verified AUG translation initiation of unconventional reading frames encoding polypeptides of minimum 40 amino acids. When applied to samples of human cell lines and tissues, their proteogenomics strategy resulted in the detection of 210 alternative proteins, mainly shorter than 100 amino acids. Unexpectedly, these altORFs were rarely found in 5’UTRs, in contrast to other reports [67]. Immunofluorescence and immunoblotting showed that altORFs and reference ORFs were co-expressed to various extents and, in analogy to canonical proteins, were often targeted to different subcellular organelles. The discovery that alternative reading frames might be present in the majority of mammalian mRNAs [69, 72] is especially important when expressing exogenous cDNA constructs, as co-expression of different protein and peptide products may remain unnoticed but might obscure experiments. Strikingly, the identification rate of alternate proteins increased from about 2% when analyzing cells and tissues, to over 50% when analyzing body fluid samples, overall leading to a grand total of 1,259 detected altORF translation products. Upon closer inspection of the data, notable observation was that without any N-terminal enrichment strategy, 889 out of 1,259 altORFs (i.e. 70%) identified, were represented by N-terminally acetylated peptides [72]. This finding might partially be explained by the higher overall chance of identifying N-terminal peptides of alternative proteins with a median size of 57 amino acids compared to 344 amino acids for conventional proteins. However, only 53% of these acetylated peptides were fully compliant with the rules of iMet cleavage and N-terminal acetylation, while 27% peptides violated these rules (including many acetylated proline residues [20]) and 20% were rare cases. For example, in the case of Met-Ser- starting human protein N-termini, iMet processing is expected in more than 95% of the cases [60], however for the alternative proteins identified here, iMet processing was observed in only 28% of the cases. These unexpected findings either hint to alternative modes of translation or potentially, false positive peptide-to-spectrum matches.
A novel ribosome profiling data analysis strategy employed by Bazzini et al. [74] led to the identification of potential sORFs in ncRNAs as well as 5'UTR and 3'UTR regions of protein-coding transcripts in zebrafish and human. sORFs with coding potential were distinguished from non-coding ORFs by a triplet periodicity of ribosome footprints, corresponding to the stepwise translocation of translating ribosomes along reading frames. Matching MS studies in zebrafish could further confirm the expression of 98 Ensembl/RefSeq annotated sORFs, 6 newly predicted sORFs located on ncRNAs in addition to 17 and 10 sORFs in 5'UTR and 3'UTR, respectively [74].

Overall, integrative omics approaches including proteomics and peptidomics provide evidence for sORF translation in a multi-cistronic manner from numerous mRNAs encoding conventional proteins. Likely, alternative translation initiation at sORFs using AUG (or near-cognate) codons is a common mechanism implicated in translational control of neighbouring proteins by regulating ribosome availability (action in cis) but also gives rise to short peptides, potentially exerting their biological functions in trans (see Results Chapter 4).

**Translation regulation by miRNAs**

Another level of post-transcriptional gene repression is orchestrated by microRNAs (miRNAs). These short noncoding RNA species of about 21-23 nucleotides, processed from hairpin-shaped precursors, are widely spread across genomes and involved in the development and metabolism of multicellular organisms, cell differentiation and proliferation, cancer and viral infection [88, 89]. Data repositories like miRBase have collected miRNA loci from 206 different species including algae, plants, animals and viruses [90]. Following the discovery of miRNA targets and pathways responsible for miRNA biogenesis and processing [88, 89], different molecular mechanisms of gene repression by miRNAs were proposed, either via translation repression or mRNA degradation [91]. Quantitative proteomic studies have been used to determine miRNA’s modes of action [92, 93]. Selbach et al. revealed the early and late effects of different miRNAs overexpressed in HeLa cells using pulsed SILAC (pSILAC) labelling in order to distinguish pre-existing proteins from newly synthesised ones [92]. Changes in protein synthesis were measured in parallel with microarray mRNA expression profiling. Individual miRNAs were found to modulate the synthesis of hundreds of proteins, corresponding to the prevalence of miRNA recognition motifs in the 3'UTR of their corresponding mRNAs. However, the impact of miRNA on individual protein levels was generally low and only in rare cases resulted in a 4-fold or higher repression, pointing to the role of miRNAs in the fine-tuning of gene expression. Further, comparing proteomic and mRNA data at early time points of miRNA expression showed that miRNAs first repress translation before they induce mRNA decay.

In another miRNA study, Baek et al. confirmed that miRNAs have an overall modest effect on proteomes despite the fact that a single miRNA may have many target genes [93]. Additionally, this study explored the hierarchy of miRNA efficiency, showing that longer
complementary sequences (8-mers) impact stronger the mRNA target as judged by both mRNA and protein changes. Further, a limited cooperative effect was observed for mRNAs with multiple recognition sites for the same miRNA. Overall, only few proteins displayed translation-only regulation and the degree of this regulation was modest (less than 33% change in expression). In contrast, targets undergoing more substantial regulation were usually destabilized at the mRNA level, putting forward mRNA decay as the proposed primary mechanism of miRNA-induced gene repression.

Further insight into miRNAs’ modes of action was gained by alternative methods, such as ribosome profiling [94, 95] and polysome profiling based on microarray determination of ribosomal occupancy of mRNAs bound to polysomal fractions resolved by density centrifugation [96]. These studies came to the same conclusion that translational control by miRNAs, to whatever extent it occurs, is probably executed during or shortly after translation initiation. Also, the impact of miRNAs on translation and on mRNA degradation might be linked and exerted simultaneously [95, 96]. Especially, studies on the short term miRNA effect [95] suggested that, in line with [92], translation down-regulation is the early response, prior to mRNA deadenylation and thus mRNA decay.

Quantitative proteomics identifies targets of translational control

Translational control of gene expression is involved in the regulation of cellular metabolism and biological responses to stress stimuli, leading to a reduction in global translation, concomitantly allowing for reprogrammed gene expression of stress response proteins [97]. Mediated e.g. by the phosphorylation of eIF2 and eIF4E and their binding partners, such control is mainly executed at the initiation phase of protein synthesis via the efficiency of ternary complex formation and mRNA 5’-cap recognition [42, 97]. Overexpression or continuous activation of various eIFs perturb translational control and are associated with malignant transformation of cells, enhanced cancer progression and poor prognosis [98-104].

Kinase mediated eIF2 phosphorylation (eIF2~P) is elicited upon viral infection (PKR), ER stress (PERK), heme deficiency (HRI) or amino acid deprivation (GCN2) [97]. By reducing translation reinitiation, eIF2~P activates the translation of uORF-repressed genes, including transcription factors GCN4 [105] and ATF4 [106]. The efficiency of cap-dependent translation is further modulated by mTOR-(S6K) signalling and the Ras-MAPK pathway through the phosphorylation of 4E-BP, eIF4G and 4B [42, 107]. Mass spectrometry accompanied by transcriptome and translatome profiling hold promise to study and characterize the coordinated proteome-wide response elicited by these molecular networks.

In this context, transcription and protein synthesis have recently been assessed in yeast subjected to oxidative, heat, DNA damage, rapamycin and osmotic-induced stress [30-32]. Despite the overall concordance between mRNA and translation, changes in protein
expression were usually delayed in time compared to mRNA changes, probably reflecting the
time required for new translation to affect the steady-state levels of proteins [30-32]. As
reported by Fournier et al. [31], in response to rapamycin, yeast cells engage the TOR
pathway to induce translation arrest by repressing ribosome biogenesis and reducing the
activity of eIFs (especially eIF4E), inducing cell cycle arrest at G1/G0 transition and enhanced
autophagy. Under these harsh cellular conditions of protein biosynthesis shut-down, several
proteins increased in abundance while no change in their mRNA levels was observed. An
analogous manner of translational regulation of gene groups linked to stress was observed
by Lackner et al. [30]. Interestingly, ribosome biogenesis was temporarily unaffected at
translation level in heat stress, in contrast to the observed down-regulation during oxidative
stress [30]. Only selected ribosomal proteins were subjected to this regulation [30], hinting
to the proposed specialization of ribosomal proteins described in other yeast species [108].
The notion that a heterogeneous ribosome composition may constitute yet another level of
translation regulation was hypothesized to occur throughout several kingdoms of life [108-
112]. Paralogous genes encoding ribosomal proteins are often exclusively expressed in
specific tissues, developmental stages or in response to environmental stress. Presumably,
when incorporated into ribosome particles, these ribosomal constituents might confer
altered translational properties to the affected particles (i.e. specificity towards mRNAs
harboring certain regulatory elements) [108-110, 112, 113].

In higher eukaryotes, quantitative studies of circadian rhythm revealed another important
post-transcriptional mechanism regulating translation [114] that is reflected in the daily
oscillations in the composition of the mouse liver proteome. The phase of rhythmic changes
in protein abundance was correlated with protein function and often delayed compared to
mRNA cycling or even independent of transcript levels [114]. Interestingly, several energy
consuming metabolic processes were limited to periodical activity and nutrient consumption
by mice. In this context, coordinated expression patterns were observed for a network of
translation initiation factors (eIF1, eIF4A2, eIF4G1, eIF5) and co-cycling translation
elongation factors (eEF2 and eEF1A1). The peak abundance of translation factors in the
nocturnal period coincided with increased translation efficiency of many mRNAs, mainly
ribosomal proteins, findings supported by an independent polysome profiling study in
mouse liver [115]. Additionally, translation initiation factors were temporally activated by
regulatory kinase pathways, such as TOR and AKT, to further enhance translation of
ribosomal proteins [115]. Taken together, these studies hint to an autoregulatory feedback
mechanism established by the synergic regulation of activity and expression levels of various
translation machinery components, thereby controlling protein synthesis independent of
mRNA expression [114, 115].

Translational control may also occur in response to chemotherapy, as recently shown in a
clinically relevant model of human multiple myeloma [37]. In order to identify drug
resistance mechanisms, the researchers performed an advanced system-wide analysis of the
transcriptome, translatome, proteome and N-terminome. Upon proteasomal inhibition by
bortezomib, cancer cells try to evade apoptosis by translational upregulation of components of the unfolded protein response (UPR) and endoplasmic reticulum transport system followed by increased mRNA degradation and translational downregulation of ribosomal proteins. Moreover, enhanced translation activities were reported in the 5’UTR mRNA region of several chaperones, proteasomal subunits, translation elongation and initiation factors as well as ribosomal proteins, hinting to their translational regulation by uORFs, often initiated at non-AUG codons [37].

Recent proteomics studies gave another indication of feedback between UPR and mRNA decay that regulates translation readthrough of premature termination codons (PTC) [22]. Nonsense or frameshift mutations in various hereditary disorders introduce PTC and drug-induced readthrough has been utilized to restore functional full-length protein synthesis. In response to such therapy, UPR triggers translational attenuation via the phosphorylation of eIF2 and reduces nonsense-mediated decay thereby increasing the availability of mRNAs affected by PTC to readthrough treatment [22].

Eukaryotic stress responses have been extensively studied at the transcriptome level, but until recently they were rarely complemented with proteomics data. Increasing lines of evidence support the observation that transcript and protein levels do not always correlate and adaptation to stress is partially mediated by translational regulation. Moreover, some functional groups of proteins seem to be affected in response to a variety of stress stimuli, indicating that mechanisms and targets of translational control are often shared. The preferentially targeted ribosomal proteins and translation factors likely orchestrate downstream translational control pathways specifically engaged to counter certain stimuli.

**Future perspectives**

Studies investigating the steady-state of proteomes provide valuable information on the repertoire of expressed proteins and changes in protein levels upon cellular perturbations. Identification of numerous novel N-terminal proteoforms is supported by a strong line of evidence from ribosome profiling and N-terminomics data. However, the dynamic and spatial aspects of protein expression cannot be overlooked. Advances in (post)metabolic strategies of protein labelling enable, amongst others, large-scale studies of protein turnover and subcellular localization [116, 117]. Combining such technologies with N-terminomics may verify the differential regulation of stability and subcellular targeting of such proteoforms on a proteome-wide scale. Enhanced knowledge on protein degradation rates, cellular localisation as well as (co-translational) protein modifications is indispensable for a comprehensive biological description of proteoforms, complementing the systemic view of translation regulation.

Novel proteomic technologies can further advance studies of protein synthesis by monitoring the immediate effect of translational regulation on newly synthesised polypeptides. A recently developed method called puromycin-associated nascent chain
proteomics (PUNCH-P) combines the isolation of nascent polypeptide chains with mass spectrometry [118, 119]. Here, polysome-associated nascent polypeptides are recovered from cells and treated with biotin-puromycin conjugates, which leads to premature translation. C-terminally labelled conjugates can subsequently be recovered by streptavidin affinity purification and subjected to MS. As such, PUNCH-P monitored changes in translation of ~5,000 human proteins during cell cycle progression. Further, PUNCH-P provided evidence for the translational coordinated expression of functionally related proteins, in addition to identifying novel players in cell cycle progression [119].

In contrast to metabolic pSILAC requiring longer duration of pulses to obtain sufficient labelling of proteins and thus reliable SILAC ratios, PUNCH-P monitors translation in a narrow time frame [119]. Protein abundances measured with pSILAC are therefore affected by degradation rates, potentially hampering the capture of unstable proteoforms including micropeptides. In contrast to pSILAC, the ex vivo nature of puromycin labelling is compatible with any type of cell and tissue, thus offering the unique possibility to study temporal proteome remodelling in response to stress or treatment in a tissue-specific manner. An exciting perspective on translation can be inferred by complementing PUNCH-P with shotgun and N-terminomics as well as RNA-seq and ribosome-profiling [119]. Further, targeting selective populations of ribosomes based on their localisation and/or interaction partners by such integrative approaches gives a true possibility to capture a more comprehensive collection of proteoforms and interrogate biological models on a system-wide level.

Nonetheless, some aspects of translation remain difficult to tackle by integrative omics studies. Fidelity of translation elongation may be impaired by defects in loading of tRNAs with cognate amino acids [120, 121] and by mutations in ribosomal subunits that decrease the accuracy in codon-anticodon pairing [122]. The incorporation of incorrect amino acids may also occur by chance (although with low frequency) [121] and was shown to be regulated by eukaryotic elongation factors [123] as well as enhanced under nutritional (such as amino acid starvation) or environmental stress [121]. In such cases, the ambiguity of protein sequence may in principle only be resolved by de novo protein sequencing. Importantly, amino acid availability may have additional consequences for protein synthesis by modulating translation elongation, leading to ribosome pausing or even translation abortion at rate-limiting codons [124]. Natural differences in cellular tRNA levels also left their evolutionary mark on protein coding sequences as evident from the correlation between tRNA abundance and codon frequency [125]. However, the effect size of tRNA availability on protein synthesis during homeostasis is highly debated [124, 126] and needs further elucidation.
References


INTRODUCTION


INTRODUCTION


OBJECTIVES
OBJECTIVES

The canonical mechanism of eukaryotic protein translation assumes that initiation of protein synthesis occurs at the 5’ proximal AUG codon of an mRNA molecule when recognized by a scanning ribosome subunit [1]. However, a certain degree of flexibility in start codon usage appears to be tolerated, leading to the recognition of suboptimal start sites, including near-cognate non-AUG codons [2]. Hence, competition between different start codons within a single mRNA molecule, (extracellular) stimuli and regulatory factors may influence the outcome of the translation process and thereby the repertoire of expressed open reading frames. Combined, and in addition to alternative splicing and protein modifications, such alternative (non-canonical) translation initiation events lead to the generation of multiple proteoforms, adding another, yet poorly studied level of complexity to the proteome.

Alternative translation initiation in eukaryotes is largely controlled by eukaryotic translation initiation factors and involves processes such as leaky scanning and translation re-initiation [3, 4]. However, the biological targets and regulatory modes of action of particular eIFs are insufficiently understood and the scarce available literature on eukaryotic proteoforms arisen from alternative translation initiation indicates that such N-terminal proteoforms may be differentially regulated and localized, display altered stabilities and functionalities [5, 6, 7].

Over the past few years, a rapid development of several state-of-the-art technologies provided the necessary tools for explaining the full diversity of protein variants expressed in vivo. In our laboratory, a positional proteomics technology called N-terminal COFRADIC [8, 9] that enriches for N-terminal peptides, allowed for the first global analysis of protein N-termini as proxies for translation initiation events and stressed the highly underestimated occurrence of alternative translation start sites [10, 11]. By deep sequencing of ribosome-protected mRNA fragments, ribosome profiling further confirmed the omnipresence of alternative translation initiation sites and the existence of novel ORFs throughout genomes [12-14].

In this work we aimed at the:

1. Discovery of novel N-terminal proteoforms

During the course of my PhD project, and building upon our acquired expertise in positional proteomics, we aimed at a comprehensive discovery and systems-level characterization of alternative N-terminal proteoforms. As the analysis of the N-terminome and ribosome profiling data provide complementary information, we further suggested integrating both types of data to expand our current understanding of translation initiation.
2. Assessment of N-terminal proteoform stability

Next, at the proteome-wide level, we pursued to supplement the annotation of (novel) N-terminal proteoforms with measurements of their cellular stability using dynamic SILAC pulse-chase experiments. Since stability has the capacity to reflect certain functional properties of proteins, investigation of isoform-specific protein dynamics might reveal yet another mechanism for biological diversification of the proteome.

3. Deep integration of OMICS data

Our long-term goal is to explore (human) proteome plasticity by incorporating translatomic data into conventional proteomics analysis pipelines. These ambitions were tackled in collaboration with Dr. Gerben Menschaert and his group at BioBix, who provided the necessary knowhow for developing specialized bioinformatics tools which greatly aided in the integration of OMICS endeavors hereby undertaken.

4. Novel insights into translational control

Finally, we aimed at elucidating translation control mechanisms driven by eIF1, an initiation factor considered as a potential molecular switch in the stringency of start codon selection.

Overall, we anticipate that proteome-wide discovery of N-terminal proteoforms, studies on their abundance, stability and regulation are essential to improve our understanding of gene expression and proteome diversity, which may in the longer term facilitate the assignment of biological functions to N-terminal proteoforms.
References:


RESULTS

Chapter I

Discovery of novel N-terminal proteoforms
Adapted from:

RESULTS

Chapter I

Discovery of novel N-terminal proteoforms

Abstract

Usage of presumed 5’UTR or downstream in-frame AUG codons, next to non-AUG codons as translation start codons contributes to the diversity of a proteome as protein isoforms harboring different N-terminal extensions or truncations can serve different functions. Recent ribosome profiling data revealed a highly underestimated occurrence of database non-annotated, and thus alternative translation initiation sites (aTIS) at the mRNA level. N-terminomics data in addition showed that in higher eukaryotes around 20% of all identified protein N-termini point to such aTIS, to incorrect assignments of the translation start codon, translation initiation at near-cognate start codons or to alternative splicing. We here report on more than 1,700 unique alternative protein N-termini identified at the proteome level in human and murine cellular proteomes. Customized databases, created using the translation initiation mapping obtained from ribosome profiling data, additionally demonstrate the use of initiator methionine decoded near-cognate start codons besides the existence of N-terminal extended protein variants at the level of the proteome. Various newly identified aTIS were confirmed by mutagenesis, and meta-analyses demonstrated that aTIS reside in strong Kozak-like motifs and are conserved among eukaryotes, hinting to a possible biological impact. Finally, TargetP analysis predicted that the usage of aTIS often results in altered subcellular localization patterns, providing a mechanism for functional diversification.
Introduction

Eukaryotic protein-coding genes can give rise to multiple translation products of which the expression is regulated at multiple levels. In contrast to transcriptional regulation, protein translational regulation permits for more immediate effects to take place. Initiation, elongation, termination and ribosome recycling constitute the different phases of the eukaryotic translation process, with translation initiation acting as the gate-keeping step by the successive steps of ternary complex recruitment, scanning, AUG codon selection and ribosomal subunit joining. Overall, this process requires over 30 different proteins including the eukaryotic initiation factors (eIFs) [1]. In eukaryotes, the translation start codon is typically found by ribosome scanning, referred to as the canonical mechanism of translation initiation. Here, the 43S pre-initiation complex (PIC) composed of the initiator Met-tRNAi pre-loaded onto the small (40S) ribosomal subunit, binds near the 5′ end of the mRNA molecule in a m7G-cap structure/eukaryotic initiation factors 4 (i.e., eIF4E, 4G and 4A; jointly referred to as the eIF4F complex) mediated fashion. This complex then starts to scan successive triplets of the 5′ untranslated region (5′UTR) in the 3′ direction until an AUG start codon or, alternatively, a near-cognate start codon entered the P (peptidyl) decoding site of the ribosome. Start codon recognition requires base-pairing with the anticodon loop of Met-tRNAi and triggers a scanning arrest and GTP hydrolysis of the eIF2-GTP-Met-tRNAi ternary complex, ultimately leading to the formation of the 48S initiation complex. The latter is then followed by factor displacement, enabling the joining of the large (60S) subunit and assembly of the elongation-competent 80S initiation complex which can now accommodate the second amino acid encoding aminoacyl-tRNA into the aminoacyl site (A-site) and thus formation of the first peptide bond in the process of translation elongation upon recruitment of translation elongation factors.

Secondary RNA structures might influence the processivity and efficiency of scanning and as such regulate translation initiation. mRNAs that contain secondary structures in their 5′UTR require ATP proportional to the degree of secondary structure [2] in addition to helicase activity to enhance 43S PIC binding and scanning.

Although the scanning mechanism of translation initiation is used by most mRNAs, an alternative manner of translation initiation of a specific subset of mRNAs is mediated by internal ribosomal entry sites (IRES). Viruses use internal ribosomal entry as a mechanism of translation, engaging host cell ribosomes while bypassing the need for (a subset of) the limiting eIFs. Internal ribosomal entry sequences are typically long and highly structured elements that mimic the functions of eIFs while requiring trans-acting factors such as the polypyrimidine tract binding protein PTB or the La autoantigen [3]. Several IRESes were also discovered in various cellular mRNAs expressed during apoptosis or mitosis or following cell stress, when cap-dependent translation is known to be impaired [4, 5]. Moreover, other specific mechanisms of translation initiation exist, such as the structural mRNA element driven, cap-dependent and IRES-like mechanism of histone H4 translation initiation, related...
to the fact that the non-canonical histone H4 mRNA features - such as its short 5’ UTR - prevent conventional scanning and translation initiation [6].

Besides IRES, a second common type of alternative translation is leaky ribosomal scanning. Here, the sequence context surrounding the first encountered AUG is sub-optimal, leading to leaky scanning and translational initiation at both this first AUG codon and additional downstream AUG codons [7].

Further, translation reinitiation after a short upstream ORF (uORF) is another common regulatory control mechanism of translation initiation [8, 9] (see also Results Chapter 4). In fact, up to 50% of all mammalian genes encode mRNAs that have at least one short uORF residing upstream of the main protein-encoding ORF and that consists of about 30 codons on average [10]. Here, some translation factors remain associated with the ribosome, thereby enabling scanning after translating the uORF and thus enabling reinitiation of translation at downstream sites.

Finally, 5’ mRNA leader sequence recapping can also give rise to alternative translates [11, 12], and thus contributes to the translational initiation landscape. *Cis-acting* sequence elements steer recognition of the correct initiation codon to ensure the fidelity of translation initiation. Usually, this AUG triplet resides in an optimal context (i.e., gcc[A/G]ccAUGG(not T)), with a purine at position -3 and a guanine at position +4 relative to the A of the AUG codon which is designated as +1 [7]. Control of translation initiation codon recognition and thereby translation initiation can additionally be exerted through various *trans-acting* factors such as elfs, where the conserved elf1 acts as a key determinant. elf1 mutations resulting in premature elf1 dissociation were shown to increase initiation rates at near-cognate start codons [13] and are thus key in maintaining the fidelity of initiation [14] (see also Results Chapter 4). Further, elf1A thought to occupy the A-site, regulates start codon selection in a dual fashion as its N-terminal region decreases the initiation accuracy and promotes elf1 dissociation at AUG codons, whereas its C-terminal region increases the stringency of start codon selection and promotes continued scanning at non-AUG codons [15]. Further, elf2 and elf5 also help to ensure the fidelity of initiation codon selection. In general, phosphorylation of the alpha subunit of elf2 (elf2αP) is known to reduce translation initiation, contradictory however, translational induction of GCN4, a yeast transcriptional activator, has been observed by reducing translation initiation at four uORFs [16], thereby overcoming the inhibitory effect of these uORFs on reinitiation at the GCN4 ORF [17].

Ribosome profiling, a recently developed genomics-based strategy, enables systematic monitoring of protein translation events by deep sequencing of ribosome-protected mRNA fragments. To date, this methodology was applied to study the translomes and the changes thereof in human [18-21], mouse [22], zebrafish [23], nematode [24], plants [25, 26], yeast [17, 27, 28], bacteria (*Escherichia coli* [29, 30] and *Bacillus subtilis* [30]), human cytomegalovirus [31] and bacteriophage lambda [32].
When used in combination with initiation-specific translation inhibitors, this technique allows for the study of (alternative) translation (initiation) with subcodon or even single-nucleotide resolution, the latter referred to as Global Translation Initiation sequencing or GTI-seq [22, 33]. As such, ribosome profiling provided a wealth of information on the mRNA engagement of (initiating) ribosomes and revealed the omnipresence of alternative translation initiation events in human and mice as nearly half the transcripts harbor multiple translation initiation sites or TIS in their sequence [22, 33]. Besides a handful of cases for which alternative TIS selection leads to (functionally) distinct proteins isoforms due to their N-terminal heterogeneity (i.e. protein stability [34], localization [35-39], function [40], etc.), the overall functional outcome of alternative mRNA engagement, the factors and mechanisms involved in TIS selection, and the overall outcome of expanding the proteome diversity remain largely elusive.

Upon ribosome emergence, nascent protein chains (i.e. 30 to 50 amino acid long protein N-termini) can be subjected to various co-translational modification events, including proteolysis (removal of the initiator methionine (iMet) by the MetAPs (methionine aminopeptidases) [41, 42]) and N-terminal (de)blocking modifications (N-terminal acetylation (Nt-acetylation) [43-46] or deformylation [47]); ubiquitous modifications in eukaryotes and prokaryotes respectively. 50% of all soluble yeast proteins and 80-90% of all soluble human proteins are modified by acetylation of the α-amino group of the amino-terminal residue (Nt-acetylation) [48-51]. The utmost N-terminal amino acid(s) identity is the major determining factor whether or not a given protein is Nt-acetylated and by which N-terminal acetyltransferase (NAT) this occurs [52], although some redundancy amongst the different NATs can be observed [50, 53]. Since Nt-acetylation is considered to mainly occur co-translationally [54], in vivo acetylated protein N-termini can thus be considered as proxies of translation initiation, though read out at the proteome-wide level.

In this study N-terminal proteomics was used to map the TIS landscape in human and mouse cells. Overall, more than 20% of all identified protein N-termini point to aTIS and we report on more than 1,700 unique alternative protein N-termini next to the more than 4,500 database annotated protein N-termini identified in the proteomes of study, thereby linking about one-third of the uniquely identified protein N-termini to alternative translation initiation events.
Discovery of novel N-terminal proteoforms

Results

Mapping of the translation initiation landscape in human and mice using N-terminomics reveals numerous alternative translation initiation events.

In this study, a proteome-wide map of the translation initiation landscape in human and mouse was created by mass spectrometry assisted analysis of protein N-termini isolated by N-terminal COFRADIC [55]. A TIS compilation was made from previously generated N-terminal proteomics data ([50, 56-58] and unpublished data) derived from the proteomes of the human HeLa, HCT116, A-431, THP-1, K-562 and Jurkat cell lines in addition to primary B-cells as well as the mouse cell lines Mf4/4 and YAC-1 next to primary dendritic mouse cells. Here, prior to tryptic digestion, all primary amines, and thus free protein N-termini, were mass tagged by acetylation using non-natural, stable isotope encoded groups such as trideutero-acetate. In this way, in vivo Nt-acetylated and in vivo free N-termini can be distinguished and the degree of Nt-acetylation determined [59]. After tryptic digestion, all protein N-termini will thus be blocked, while all other internal peptides will have a newly generated primary α-amine. Subsequently, N-terminal peptides are enriched for by means of strong cation exchange (SCX) step at low pH and further segregated from remaining internal peptides using a diagonal chromatography strategy. Selected protein N-termini are subsequently identified following LC-MS/MS analysis [60]. Identified protein N-termini were grouped by their TIS context. First, protein N-termini with a Swiss-Prot database annotated protein start position (i.e., N-termini starting at protein position one or two in the protein sequence) are referred to as database annotated TIS or dbTIS. Overall, 2,879 human and 1,771 mouse dbTIS-indicative N-termini originating from 2,723 and 1,708 unique Swiss-Prot protein entries were identified (Appendix Table I.1).

Second, based on the co-translational nature of N-terminal acetylation of protein N-termini [48] by the NATs, the near universal requirement of a Met-encoding initiator codon (iMet) and the co-translational processing of iMet by methionine aminopeptidases (MetAPs), all in vivo free and/or Nt-acetylated peptides with start positions downstream the database annotated TIS were grouped. In this way, 1,231 human (1060 proteins) and 465 mouse (418 proteins) N-termini hinted to protein N-termini originating from the usage of in-frame, downstream TIS (dTIS), thus giving rise to N-terminal truncated protein isoforms. The N-termini hinting to dTIS were further subdivided into two sub-categories (See Table 1 and Experimental procedures); the high confident dTIS encompassing all (partially) in vivo Nt-acetylated peptides amongst others. In vivo unmodified dTIS compliant with the rules of iMet-processing and Nt-acetylation (the latter for example considering that, without exception, (X)-Pro-starting N-termini are unmodified [61]) were withheld as low confident dTIS and this only when their protein start position did not overlap with any proteolytic cleavage event reported in public repositories [62-65] (i.e., protein signal processing sites or reported proteolytic cleavage sites after or before a Met residue). Finally, and whenever
ribo-seq or orthologue mapping hinted to translation initiation events, low confident dTIS were re-catalogued as high confident dTIS (see below and Appendix Table I.1).

<table>
<thead>
<tr>
<th>Human dTIS</th>
<th>Mouse dTIS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>confidence level</td>
</tr>
<tr>
<td></td>
<td>H</td>
</tr>
<tr>
<td>identified peptides</td>
<td>858</td>
</tr>
<tr>
<td>mapped peptides</td>
<td>850</td>
</tr>
<tr>
<td>Ensembl TIS</td>
<td></td>
</tr>
<tr>
<td>exon 1</td>
<td>41</td>
</tr>
<tr>
<td>exon &gt;1</td>
<td>146</td>
</tr>
<tr>
<td>subtotal</td>
<td>187</td>
</tr>
<tr>
<td>no Ensembl TIS</td>
<td></td>
</tr>
<tr>
<td>exon 1</td>
<td>301</td>
</tr>
<tr>
<td>exon &gt;1</td>
<td>362</td>
</tr>
<tr>
<td>subtotal</td>
<td>663</td>
</tr>
<tr>
<td>Swiss-Prot isoform TIS</td>
<td></td>
</tr>
<tr>
<td>no Swiss-Prot isoform TIS</td>
<td>17</td>
</tr>
<tr>
<td>TrEMBL TIS</td>
<td>159</td>
</tr>
<tr>
<td>no TrEMBL TIS</td>
<td>691</td>
</tr>
<tr>
<td>ribo-seq TIS</td>
<td>96</td>
</tr>
<tr>
<td>no ribo-seq TIS</td>
<td>754</td>
</tr>
<tr>
<td>identified peptides(+meta)</td>
<td>900</td>
</tr>
<tr>
<td>mapped peptides(+meta)</td>
<td>892</td>
</tr>
</tbody>
</table>

Table 1 - All experimentally observed alternative translation initiation sites (aTIS: 1231 human, 465 mouse) were mapped onto their corresponding reference genomes (Ensembl human GRCh37 and mouse NCBIM37). The dTIS locations were detected throughout several mass spectrometry analyses (i.e. data obtained from previously generated N-terminal proteomics data ([50, 56-58] and previously unpublished data) and a compilation was extracted from the in-house ms-lims system [66] (Appendix Table I.1). The overlap with annotated Ensembl, Swiss-Prot or TrEMBL annotated TIS is also provided. For the Ensembl mapping an extra subdivision was done based on TIS location in the first (Exon1) or the consecutive exons (Exon>1). Also, a comparison was made with the TIS identified in two ribosome profiling studies on HEK293 and mESC cell lines (18,19). See the “selection of N-termini” paragraph within the material and methods section for more explanation on the subdivision based on confidence level (either H or L). ‘(+ meta)’ indicates that isoform, transcript, ribo-seq and/or orthologues dTIS metadata is available for dTIS originally assigned as low confidence (see also Appendix Table I.1).

For dbTIS, the discrepancy between the numbers of identified protein N-termini and the actual proteins is only due to the observed incompleteness of iMet-processing (i.e., cases where both the iMet processed and unprocessed N-termini were identified) and thus
heterogeneity of the N-terminal proteoforms, while in the case of dTIS, multiple dTIS were observed for several proteins.

Further, the identified N-termini were grouped by mapping them to the first or a subsequent exon, with the former category hinting to alternative translation events by leaky scanning or reinitiation, whereas the latter, besides representing putative dTIS, might point to TIS originating from alternative splicing (Table 1 and Appendix Table I.1). Overall 1,220 human dTIS (out of the 1,231 Swiss-Prot N-terminal identified) and all 465 mouse dTIS N-termini could be mapped onto their corresponding reference genome and their confidence level is given in Table 1 and Appendix Table I.1. Meta data related to these d(b)TIS identifications are made available as visualization tracks in the H2G2 genome browser (http://h2g2.ugent.be/biobix.html, see also Appendix File I).

Of the dTIS N-termini identified, 18% (n=223) and 15% (n=71) of the Swiss-Prot non-annotated human and mouse TIS mapped to Swiss-Prot isoform entries and/or indicative transcripts in TrEMBL and/or Ensembl, validating our selection strategy for identifying dTIS, as these have been experimentally proven to give rise to N-terminally truncated protein isoforms (Appendix Table I.1). Overall these numbers are indicative for the fact that our dataset is of high quality and thus holds numerous hitherto unreported dTIS sites, here discovered at the level of the proteome.

**TIS sequence context analyses**

A survey of the sequence context flanking the dbTIS and dTIS of the Exon1 and Exon>1 categories using WebLogo [67] revealed a preference of the most crucial Kozak context elements being a purine at position -3 and guanine at position +4 [68] in the Kozak motif gcc[A/G]ccAUGG(not U) with the dbTIS and dTIS_{exon1} equally well conserved, followed by the dTIS_{exon>1} category. Only d(b)TIS locations were taken into account where the complete flanking region does not span any splice junction. For the dbTIS, additionally the higher GC content in the flanking nucleotide context becomes noticeable [69, 70] (see Figure 1 A. and B.).

A detailed analysis of the human Swiss-Prot dbTIS for which we report an alternative start site located in the first exon (i.e. the dTIS_{exon1} category) additionally revealed an increased occurrence of suboptimal start codon contexts (with a pyrimidine in position -3 upstream of AUG instead of purine [71]). As compared to the start codon contexts of all identified dbTIS, an increased suboptimal versus optimal measure of 35.2% versus 19.5% is observed (as deduced from input data used in Fig. 1 and Appendix Fig. I.1). According to the leaky scanning model [72], the 40S ribosomal subunits can miss an AUG codon in a suboptimal context and initiate translation at downstream AUG(s) which is in corroboration with the data obtained from GTI-seq data, showing that the strongest Kozak consensus sequence was observed in the gene group with no detectable dTIS but with dbTIS initiation, while this context was largely absent in the group of genes lacking a detectable translation initiation at
Figure 1 - A. Homo sapiens TIS WebLogos. The flanking sequences (12 bases upstream, 9 bases downstream) of the corresponding dbTIS for which a dTIS has been identified, the experimentally observed dTIS, located in exon 1 (n = 374), and the experimentally observed dTIS, located in subsequent exons (n = 791) are used to create WebLogos. B. Mus musculus TIS WebLogos. The flanking sequences (12 bases upstream, 9 bases downstream) of the corresponding dbTIS for which a dTIS has been identified, the experimentally observed dTIS, located in exon 1 (n = 197), and the experimentally observed dTIS, located in subsequent exons (n = 251) are used to create WebLogos. Both the probability and bits values are plotted. The discrepancy between the numbers given above and the numbers in Table 1 are because splice site spanning flanking sequences were not used to create the WebLogo analysis.

dbTIS [33]. The downstream flanking sites of these downstream start sites in both the dTIS\textsubscript{exon1} and dTIS\textsubscript{exon>1} categories were further investigated using the AUG_Hairpin software, enabling the prediction of downstream secondary structure influencing translation start site recognition [71, 73-75]. Following the strategy of Kochetov \textit{et al.}, only those dTIS that show a stable stem-loop structure (E\textsubscript{tot} < -20 kcal/mol) located between 13 and 19 nucleotides downstream from the start site were retained. Average energies of eligible stem-loop
Discovery of novel N-terminal proteoforms structures \((E_{\text{tot}})\) were -32.2 kcal/mol and -32.6 kcal/mol for the dTIS with suboptimal and optimal start codon contexts respectively (also the distributions of \(E_{\text{tot}}\) values proved not to be significantly different according to a Kolmogorov-Smirnov two-sample test). Overall, the presence of the Kozak sequence context in all categories is further indicative for real TIS events.

**Conservation analysis**

To assess the possibility of evolutionary conservation of the identified dTIS and their flanking sequences as compared to their corresponding dbTIS, the orthologous positions in various vertebrate genomes were extracted using phastCons [76, 77] and scored in a multiple sequence alignment, thereby generating a metagenic conservation plot (Fig. 2). Also, an analysis was made between the identified dTIS and a set of 5,000 randomly chosen, BioMart [78] annotated complete CDS (CCDS) translation initiation sites (serving as a proxy for the global dbTIS landscape, Appendix Fig. I.2). Only d(b)TIS locations were taken into account where the complete flanking region does not span any splice junction. In general, the phastCons score (between 0 and 1) gives a probability that each nucleotide belongs to a conserved element (see Material and Methods for detailed explanation). Overall, the human and mouse conservation plot indicated that the dTIS are highly conserved, with a mean conservation score of 0.97 (+/- 0.002, 95% confidence interval) and 0.97 (+/- 0.005) for respectively the Exon1 and Exon>1 groups compared to the dbTIS with a mean conservation score of 0.96 (+/- 0.01) and are thus indicative for the fact that the dTIS translation start sites are very well conserved within eukaryotic genomes in analogy to what was previously reported by Bazykin et al. [79] using in silico predictive analyses and using in vivo GTI-seq experiments [33].

Further, the conservation scores of the dTIS flanking regions of both the Exon1 and Exon>1 groups are high, ranging from 0.9 to 1. Here, next to the translation start codon, other Kozak hallmarks such as the guanine at position +4 and purine at position -3 are well conserved. Also notable in the dTIS conservation plots - and expected given the higher coding potential of the first two nucleotides - is the slightly higher conservation of the first two nucleotides of the coding triplets in the translated sequence [80]. This feature is most pronounced in the human dTIS\(_{\text{Exon>1}}\) plot. As opposed to the dTIS plots, the flanking 5′ upstream sequences in the dbTIS plots score significantly lower as these presumably contain untranslated sequence (UTR) in contrast to the 5′ upstream region of the dTIS that contain translated sequence encoding for N-terminal protein extensions. No significant differences are obvious between the dTIS conservation plots of the Exon1 and Exon>1 groups.

Statistical testing was performed to assess the sequence conservation surrounding the Kozak motif and to increase confidence that the identified sites (dTIS) are genuine translation initiation sites. For that purpose we compiled a dataset of decoy sites meeting the following criteria: (I) Consensus Coding Sequences (CCDS) were scanned for downstream Kozak sequence motifs [A/G]ccAUGG, (II) the identified Kozak sequence motif sites that overlap
Figure 2 - Conservation plots for the flanking exonic regions of human and mouse TIS.

Conservation plots for the flanking (12 bp upstream, 9 bp downstream) exonic regions of human (left pane) and mouse d(b)TIS (right pane) are plotted. The conservation measure averages the phastCons score at every position of all flanking sequences within the subgroups (corresponding dbTIS, dTIS located in exon 1 (Exon1 subgroup), and dTIS located in subsequent exons (Exon>1 subgroup)) after alignment based on their translation start site. For all flanking positions, the mean is provided together with its 95% confidence interval. The upper panels show the conservation plot of the flanking region of corresponding dbTIS for which a dTIS has been identified. The middle and lower panel are respectively based on flanking regions of dTIS of the Exon1 (n = 374 for human, n = 197 for mouse) and Exon>1 (n = 791 for human, n = 251 for mouse) subgroups. Only d(b)TIS locations were taken into account where the complete flanking region does not span any splice junction

with dTIS identified in the NT-COFRADIC mass spectrometry analysis described in this study were discarded, (III) the ones showing an overlap with dTIS identified within the ribosome profiling experiments re-analyzed in this study (human (33) and mouse (22)) were also discarded. This group of decoy sites was compared with the different categories of TIS described in the study: database annotated TIS (dbTIS), downstream TIS located in exon1 (dTIS_{exon1}) and downstream TIS located in further downstream exons (dTIS_{exon>1}). The PhastCons conservation scores at positions (-3,+1,+2,+3,+4), the most crucial Kozak context positions, were averaged for further calculation. A low P-value (4.701e^{-12} and <2.2e^{-16} for respectively the human and mouse datasets) in a Welsh one-way ANOVA is indicative for a difference amongst the four TIS groups (after testing for heteroscedasticity using the Levene’s test). In order to determine which particular group of TIS deviates the most, a Tukey’s Honestly Significant Difference (Tukey-HSD) post-hoc test (accounting for heteroscedasticity by using a heteroscedastic consistent covariance estimation) was performed showing a clear difference between the decoy group and the three other sets (dbTis, dTIS_{exon1} and dTIS_{exon>1}, p-value < 0.001) at a 95% confidence level.
Finally, and to further analyze the degree of conservation of dTIS between our human and mouse datasets, the experimentally identified mouse and human dTIS were compared. In total, of 200 orthologous dTIS pairs, both the human and mouse N-termini could be identified (i.e. for 43% of all mouse dTIS identified, the human orthologous N-termini could be identified).

Of these, 29 human and 31 mouse dTIS were originally classified as low confident dTIS. Interestingly, and based on the MS/MS-based evidence of Nt-acetylation of these orthologous N-termini, 3 human and 9 murine dTIS could now be re-catalogued under the reliability class 1 (Appendix Table I.1).

**TargetP analysis**

To have a first approximation of the functional impact of alternative TIS usage, a TargetP analysis was performed predicting the subcellular location of the full-length proteins (i.e. proteins translated starting from their dbTIS) versus their N-terminally truncated counterparts identified (dTIS\(_{\text{exon1}}\) and dTIS\(_{\text{exon>1}}\)) [81]. Figure 3 (upper and lower panel for respectively human and mouse) show that although only a small percentage of the dbTIS protein products is predicted to contain a mitochondrial targeting or signal peptide (i.e., most likely an underrepresentation, since in most cases signal or transit peptide maturation has occurred), a noticeable decrease of secreted or mitochondrial targeted proteins can be observed when assessing their N-terminally truncated counterparts (see Figure 3 pie charts, Chi-squared test of Independency, p-value < 2.2e\(^{-16}\) for both the human and mouse datasets). The bar plots within Figure 3 give a more detailed view, making an extra subdivision based on (I) the reliability of the TargetP prediction (class 1 to 5) and (II) whether the dTIS is localized in exon1 or in an exon downstream exon1 (exon>1). The more detailed bar plots also show a significant drop for the mitochondrial and secretory pathway localization categories (“M” and “S”) independent of the reliability classes (1-5) in both the exon1 and exon>1 groups (green versus blue bars).

Overall, the TargetP output strengthens the idea that dTIS usage has an impact on protein subcellular localization [35, 36, 82], which was also hypothesized and computationally investigated by Cai et al. [83] and in fact proven for a variety of N-terminal protein isoforms generated by means of alternative translation initiation.

**Ribosome profiling data provide independent experimental support for N-terminomics data**

Interestingly, of the here identified TIS, complementary ribo-seq TIS profiling data are available for 861 of the 1,755 transcript-matching mouse dbTIS (49%), 69 of the 465 mouse dTIS (15%), 1,150 of the 2,841 human dbTIS (40%) and 105 of the 1,220 human dTIS (9%) (Appendix Table I.1), thereby providing evidence that these represent genuine translation initiation sites in mouse and human transcripts [22], and thus that these N-termini are representative for N-terminal protein variants.
Figure 3 - TargetP analysis of the protein products generated by dbTIS versus dTIS usage. TargetP predicts both N-terminal mitochondrial targeting peptide (mTP) and signal peptides (SP) processing, respectively reflecting mitochondrial and secretory pathway localization. Both human (upper pane) and mouse (lower pane) are plotted. The pie charts depict the overall localization patterns of the dTIS and dbTIS translation products. The more detailed bar charts make an extra subdivision based on (I) the reliability of the TargetP prediction (class 1 to 5, where 1 indicates the strongest prediction) and (II) whether the dTIS is localized in exon1 or in an exon downstream exon1 (exon>1). The green and blue bars respectively correspond to N-terminal isoforms raised upon dbTIS and dTIS usage, dark and clear bars represent the Exon1 and Exon>1 group respectively. The X-axis shows the predicted localizations (“M” stands for mitochondrial, “S” for secretory pathway) and the reliability of that prediction (class 1 to 5). The rightmost bars depict the combination of all reliability classes (“All M” and “All S”). The Y-axis corresponds to the total number of TIS events falling within the groups depicted in the X-axis.

As such, the experimental evidence obtained by ribosome profiling-assisted TIS identification categorizes 47 extra mouse dTIS and 66 extra human dTIS to reliability class 1, thereby increasing the percentage of validated dTIS to 27% (n=125) and 22% (n=272) in mouse and human respectively (Appendix Table I.1).
Discovery of novel N-terminal proteoforms

Overall, and when taking into account the available isoform, transcript, ribo-seq and orthologues dTIS metadata, 24 extra mouse dTIS and 42 extra human dTIS originally assigned as low confidence, are now classified as highly likely genuine dTIS, respectively summing up to 73% (n=900) and 64% (n=298) of all identified human and mouse dTIS having data that support their translation initiation potency (Appendix Table I.1).

Since the translatome (i.e. the ORF delineation and the translation initiation landscape) can be specifically delineated using ribosome profiling data, usage of this type of data does not necessitate translation into its three reading frames, hence decreasing the search space tremendously. Alongside, non-canonical codons serving as alternate initiation codons are depicted. These near cognate translation initiation codons can either be decoded as the expected initiator methionine residue or alternatively to their coding-matching amino acid as for example leucine-decoded CUG starts of translation initiations have been reported [84, 85]. For these reasons, customized sample-oriented and ribosome profiling derived protein databases were created which served as custom-made reference datasets for the proteomes under study [86].

In the mouse proteomes, besides the additional full-length protein N-termini identified (i.e., N-termini not enclosed in the Swiss-Prot database), 17 N-termini indicating N-terminal protein extensions were identified (Appendix Table I.1). Four N-terminal extensions were generated upon translation initiation at an AUG codon, in addition, 13 were produced by translation initiation at near cognate start codons; being GUG (6 N-termini), CUG (4) and ACG (3) normally encoding for respectively Val, Leu and Thr, but decoded to Met as evident from the iMet-retaining N-termini identified. Besides, the N-termini of an uORF and a dTIS protein product could be identified using respectively GUG and CUG as start codon.

In human, 17, 4 and 2 N-termini respectively hinted to N-terminal protein extensions, N-terminally truncated and overlapping uORF protein products, 22 of which were generated upon translation initiation at a near-cognate start codon (Appendix Table I.1). Besides, these database searches led to the identification of some additional dbTIS, not contained in the Swiss-Prot database (Appendix Table I.1).

**TIS mutagenesis analyses reveal that N-terminal protein isoforms of the class dTIS_exon1 are generated by means of leaky ribosomal scanning**

To further verify whether some of the alternative TIS products identified are raised by alternative translation initiation, *in vitro* translation studies of (mutagenized) dTIS holding coding sequences (CDS) flanked by (a part of) their presumed 5’UTR were performed using coupled *in vitro* transcription/translation assays. In general, mutation of AUG to CUG typically abolished or greatly diminished translation initiation at these sites. In all cases distinct protein bands corresponding to the short N-terminal protein isoform(s) (i.e. dTIS products), identified by proteomics means, and the database annotated variant could be observed (Fig. 4).
Figure 4 - *In vitro* translation of TIS-mutagenized constructs reveal the existence of the by proteomics identified N-terminal protein variants. Wild type and d(b)TIS mutagenized pOTB7 constructs encoding N-terminal variants of aminoacyl tRNA synthase complex-interacting multifunctional protein 2 (AIMP2), inhibitor of kappa light polypeptide gene enhancer in B-cells kinase gamma (NEMO), Zinc finger protein 296 (ZN296), splicing factor 3a subunit 3 (SF3A3), cytoplasmic aspartate-tRNA ligase (SYDC), cytokine receptor-like factor 3 (CRLF3), Nucleosome assembly protein 1-like 1 (NP1L1) and pyridoxamine 5’-phosphate oxidase (PNPO) were *in vitro* transcribed and translated. Following SDS-PAGE and electroblotting, radiolabeled proteins were visualized by radiography. Assignments of the precursor band corresponding to the database annotated protein sequences (black arrowhead) and protein products raised upon dTIS usage (orange and dashed arrowheads) were verified by mutating their respective initiator methionine. A grey arrow points to translation initiation events at dTIS which were not identified by proteomics means. An asterisk is indicative of an unspecific protein band produced in the control *in vitro* transcription and translation reaction (i.e. a reaction without input DNA). In each case the theoretical molecular weights of the identified N-terminal protein variants are indicated.

Further, since the mutation of the canonical initiation site affected the production of the truncated isoform(s) and vice versa (i.e. some truncated N-terminal isoforms were only detected when the dbTIS were mutagenized), our results strongly indicate that the dTIS products are produced by alternative translation initiation via leaky ribosomal scanning at internal translation start sites. Besides, in some cases a deviation from the 5’ polarity of scanning could be observed for closely spaced AUG codons (up to 16-19 nt). As a result of the proposed reverse directionality of scanning (3’ to 5’), a lower initiation frequency at a 5’ proximal AUG could be observed in the presence of (a) nearby downstream AUG(s), suggestive of downstream nucleotides inferring a restricted relaxation to the forward directionality of scanning of the proximal AUG [87].
**Discussion**

The most acknowledged mechanism of protein diversification in mammalian genomes is alternative splicing, where different mRNAs are derived from the same nascent transcript. Only recently, alternative translation initiation from a single mature transcript was recognized as an important and wide-spread mechanism of protein diversification, further highlighting the importance of gene functionality [22, 33].

As previously postulated, targeted analysis of protein N-termini is ideally suited to study N-terminal protein isoform diversity [35].

In this study, we report on more than 1,700 alternative translation initiation events in mouse and human cell lines by applying stringent rules for mass spectrometric based identifications of N-terminal peptides. Besides, our detailed understanding of the specificity of the Nt-acetyltransferases and co-translational processes in general assists in judging if such peptides indeed report protein translation events and thereby allow for the functional (re-) annotation of genomes. For a significant fraction of the here reported TIS, available metadata from transcripts, ribosome profiling, TIS sequence context and conservation analyses served as evidence that our N-terminal selection provides a very powerful strategy to map the TIS landscape in higher eukaryotes. In addition, and as is the case for the high mobility group proteins HMGB1, HMGB2 and HMGB3, next to the orthologues matching dTIS sites, corresponding dTIS sites in all three homologues could be identified (Appendix Table I.1). These observations are further strengthened by the fact that previously reported dTIS products of the Insulin-like growth factor 2 mRNA-binding protein 2 (IGF2BP2) [88], Glucocorticoid receptor (GCR) [89], Insulin-degrading enzyme [90], Regulator of G-protein signaling 2 [91] and the BAG family molecular chaperone regulator 1 [92] - of which the N-terminal isoform expression was shown to display a stage- and site-specific expression profile during mouse development - were also identified in this study (Appendix Table I.1).

Ribo-seq and GTI-seq in mammalian cells revealed that only half of the TIS codons made use of AUG as the translation initiation codon. However, in the study of Lee et al. [33], TIS codon usage was shown to be distinct when residing in the presumed 5'UTR (uTIS) as opposed to the annotated CDS. When outside the dbTIS/dTIS reading frame, uTIS are mostly associated with short ORFs and were mostly non-AUG codons, while dTIS, typically encoding for N-terminal truncated protein variants predominantly made use of AUG codons, an observation which is in line with our data and the fact that only 37 protein products raised upon translation initiation at near-cognate start codons, typically giving rise to N-terminal protein extensions, were identified in our mouse and human datasets based on available ribo-seq datasets [22, 33]. Further, a recent computational analysis of ribosome profiling data calculating the efficiencies of individual translation initiation sites, revealed that despite the high frequency of non-AUG translation initiation sites identified by means of ribosome profiling, the probability of initiation at non-AUG codons was found to be considerably lower.
than at AUG codons (data presented by Pavel Baranov at the EMBO Conference Series ‘Protein Synthesis and Translational Control’, Heidelberg, Germany 2013 [93]), likely explaining their general underrepresentation in the N-terminomics datasets here presented.

In addition to translation reinitiation for alternative translation initiation, GTI-seq demonstrated that leaky scanning was the major contributing mechanism leading to TIS selection since the strongest Kozak consensus sequences were observed in the gene group with dbTIS selection but no detectable dTIS, while dTIS selection was observed when a weak or no consensus sequence context of the dbTIS was present, enabling for an estimation of the leakiness of the first AUG codon and again confirming our proteome data.

Our TargetP analyses as well as other multiple lines of evidence indicate that alternative translation initiation can give rise to iso-functional though localization-specific N-terminal protein variants making translation initiation a very attractive mechanism of regulating protein localization as previously reported for the p43 Component of the Multisynthetase Complex [37] where a mitochondrial targeting sequence is lost when translation initiation proceeds via a dTIS, while in contrast, translation initiation at a dTIS in Flap endonuclease 1 (FEN-1) exposes a cryptic mitochondrial targeting signal [38], two cases contributing to the increased complexity of the mitochondrial proteome [39].

Although translation initiation at dTIS found in close proximity to dbTIS are more likely to yield isofunctional and localization non-distinct N-terminal isoforms, thereby providing a potential fail-safe mechanism for translation initiation to occur, it is noteworthy that the N-terminal identity of a protein is critically important in determining protein stability. More specifically, the N-end rule relates the regulation of the in vivo half-life of a protein to the identity of (the N-terminal modification of) its N-terminal residue. Therefore, the loss of even a single N-terminal amino acid or its modification can significantly impact protein stability [94] (see also Results Chapter 2 [95]), influence protein localization [96] and protein complex formation [97] amongst others. Finally, our TargetP analyses shows a clearly noticeable change in localization in both exon1 and exon>1 groups for both the human and mouse datasets (Fig. 3) indicative that the findings of altered subcellular localization/functional diversification also applies to the broader set of alternative translation events identified in this study.

Besides steering protein localization and protein stability, TIS selection can also regulate protein expression levels. For example, hypo- or hypermorphic point mutations introducing premature translation termination codons in the first exon can result in the production of truncated protein variants through translation initiation at in-frame methionines downstream the nonsense mutation [98-100]. Two such examples for which various dTIS sites were identified in this study include the nuclear factor kappa B (NF-κB) essential modulator (IKKγ/NEMO) and the NF-κB inhibitor IκBα [98-100] (Fig. 4 and Appendix Table I.1). Premature translation termination codons in these genes - although leading to the residual production of a truncated variant sufficient for the non-lethality observed during
development - have been shown to underlie specific cases of the genetic disorder anhidrotic ectodermal dysplasia with immune deficiency. The (presumed) dTIS of both reported truncated protein variants have here been identified as being in vivo Nt-acetylated (Ac-M_{38}LHLPSEQAPETLQR (NEMO) and Ac-M_{37}KDEEYEQMVKELQEIR (IκBα)), indicative for the fact that, as demonstrated to be the case for wild type NEMO transfected cells, a (limited) translation initiation of IκBα at these alternative methionines can also occur under normal cellular conditions. The dTIS product of IκBα lacks the two amino-terminal IκB kinase (IKK) phosphorylation sites known to be essential for targeting IκBα for proteasomal degradation, and as a result the degradation-resistant variant acts as a dominant negative regulator of NF-κB activity. The mutation in IKKy/NEMO, the scaffolding subunit of the IKK complex, gives rise to a truncated but functional variant that is produced in limited, insufficient amounts for the development of protective immune responses. In vitro transcription/translation assays independently confirmed that the 44 kDa isoform of NEMO is the product of alternative translation initiation at the internal Met38. Here, close to the canonical start codon, other start codons are located downstream of the database annotated iMet₁ (i.e. GxxAUGG and GxxAUGC), corresponding to the (surrounding) nucleotide motifs of methionines 13 and 38. The likelihoods of these AUG codons to act as translation initiation sites were estimated 0.29 and 0.27 versus 0.54 for the first AUG codon (TxxAUGA) (translation initiation prediction at http://atgpr.dbcls.jp). As such, various mutagenized constructs were made (Fig. 4) in which the presumed initiator Met1, the internal Met13 and/or Met38 were mutagenized to monitor translation initiation at these sites. Expression from the coding sequence alone resulted in the production of two clearly distinct protein forms (Fig. 4, WT sample), of which the expression of shorter isoform was lower as compared to its full-length counterpart, indicative for leaky ribosome scanning. Further, mutagenesis of Met13 also resulted in two distinct bands of which the higher MW band runs approximately 1 kDa lower as compared the highest MW band of the WT construct and thus probably indicative for the fact that the higher MW band observed encompasses the products of translation at Met1 and Met13, an observation which is confirmed when Met 13 or alternatively Met 38 is mutated. The fact that translation initiation may occur at Met 1 and Met 13 is further supported by the observation that cellular expression of the WT and M38A mutant resulted in a somewhat smeared out precursor band [100] and that the TrEMBL and Ensembl databases hold preliminary entries of this Met13 initiated protein variant. The combined Met 13/38 mutant seems to express the 44 kDa form besides residual translation initiation at the first non-canonical mutant CUG codon. Although, in vitro translation from the triple AUG to CUG mutant construct is significantly impaired, residual translation initiation (mainly at the near-cognate start codon decoded to Met38) can still be observed. Overall, we observed translation at three different AUG codons in NEMO, of which the translation product starting with Met38 was identified using N-terminal proteomics in the proteomes of K562, THP-1 and B-cells.
Figure 5 - Representation of the single exon encoding mouse and human JUNB protein in the H2G2 genome browser. Several information tracks are presented. From top to bottom: (I) Ensembl Gene annotation, (II) Ensembl Transcript annotation, (III) ribosome profile data of control human HEK293t A. and mESC cell line sample B. showing profiles alongside the CDS, (IV) ribosome profile data of lactimidomycin (LTM) treated human HEK293t A. and harringtonine treated mESC cell line sample B. showing profiles alongside the CDS, (V) ribo-seq predicted TIS, (VI) the translation product prediction based on the two aforementioned tracks, (VII) TrEmbl annotated TIS, (VIII) Ensembl annotated TIS and (IX) dbTIS and (X) alternative TIS identified using N-terminal COFRADIC. Two zoomed figures (in black boxes) are also depicted, representing a more detailed view of the genomic region around the translation initiation site of the N-terminal truncated JUNB protein isoform in human A. and mouse B. identified using N-terminal COFRADIC, clearly demonstrating the use of an alternative initiation site (ATG) and accumulated ribo-seq signals at this start site in human and mouse. Furthermore an alignment of the mouse and human N-terminal protein sequences of the transcription factor jun-B and d(b)TIS conservation is presented together with the C. autoradiograph showing translation initiation of *in vitro* transcribed JUNB at the AUG start codons encoding M₁ and M₅.

Further, ribosome profiling data where the utmost 5’ AUG triplet resides in a suboptimal context (absence of a purine at -3 and/or G at +4) suggested that more than one quarter of
the human transcripts showed clear evidence of downstream translation initiation and thus could display a bi- or multicistronic behavior [33], meaning that these mRNAs could produce more than one polypeptide through leaky scanning. Many of these hypothetical cases would be expected to involve the production of small peptides or N-terminal truncated protein isoforms that have routinely been excluded from database sequence annotations.

Despite the multitude of alternative TIS here identified, linking over 30% of the protein N-termini to alternative translation initiation, for the majority of them their spectral counts (Appendix Table I.1) hint to a general lower translation efficiency and thus likely lower concentration of protein products expressed from such secondary initiation codons. Nonetheless, potent functions of such possibly lower abundance dTIS protein products have been demonstrated, as in the case of the mitogenic osteogenic growth peptide (OGP) translated by leaky scanning from mammalian histone H4 mRNA [101]. Strikingly however, when TIS are relatively closely spaced (i.e. resulting in N-terminal protein isoforms differing less than about 6 amino acids), spectral counts and our in vitro translation results hint to a more balanced expression, likely linking this to a proximity effect previously shown to modify the strict sequential constraint of regular leaky scanning into a more competitive feature [87]. In this respect, it is important to note that such protein products can easily be overlooked by conventional detection technologies such as western blotting despite their potent expression, as here shown to be the case by mutagenesis studies enabling visualization of the N-terminal AIMP2, JUNB, UCHL1 and CRLF3 isoforms generated by translation initiation at closely spaced AUG start codons (Fig. 4-6).

Further, the bioinformatics-assisted integration of positional proteomics and available ribosome profiling data enabled for a more sensitive and comprehensive protein discovery, thereby enabling a global (re-)annotation of the translation initiation landscape [31, 86]. Interestingly, the high overlap of alternative translation products identified in this comprehensive study with those from a previous positional proteomics study that focused on the TIS-landscape in mouse embryonic stem (mESC) cells, further hints to their functional importance and conservation. In contrast however to the majority of translation products linked to alternative translation initiation as observed in ribosome profiling, the repertoire reported in this study is mainly confined to translation products raised upon downstream AUG besides some upstream non-cognate codon usage, meaning that the products of uORF and out-of-frame translation generally remain undetected. Although the sensitivity of ribosome profiling to detect TIS sites and translation products remains unprecedented, the bioinformatics-oriented approaches used to assign TIS in ribosome profiling in some cases appear ineffective (e.g. the dTIS in JUNB here identified, clearly provided a discrete LTM and Harringtonin signal in the human and mouse ribo-seq datasets [22, 33], though remained undetected by the strict training algorithm used in the original Ingolia et al. study (Fig. 5)). This, as well as the occurrence of co-translational protein modification events and the recent finding that mRNA ribosome occupancy does not necessarily hints to effective translation [102], necessitate the need for proteomics endeavors to identify the (mature) translation
products. As such, we foresee that the complementary use of proteomics approaches and ribosome profiling will further assist in the comprehensive cataloguing of TIS ultimately leading to detectable and functional translation products.

Figure 6 - Representation of the human UCHL1 protein in the H2G2 genome browser. A. Information tracks are presented as in Fig. 5. Furthermore an alignment of the N-terminal UCHL1 mouse and human protein sequences and d(b)TIS conservation is presented together with the autoradiograph (B.) showing translation initiation of in vitro transcribed ubiquitin carboxyl-terminal hydrolase isozyme L1 (UCHL1) at the AUG start codons encoding $M_1$ and $M_6$ and $M_{12}$. 
Materials and Methods

Cell culture

Human epithelial colon (HCT116) and human epidermoid (A-431) cells were grown in Dulbecco’s Modified Eagle’s Medium (DMEM) medium supplemented with 10% fetal bovine serum and penicillin/streptomycin. Human monocytic THP-1 cells and human primary B-cells were grown in HyClone Roswell Park Memorial Institute (RPMI) medium supplemented with 10% fetal bovine serum and penicillin/streptomycin (all Invitrogen). Primary mouse dendritic cells [57], mouse macrophage Mf4/4 [58], mouse lymphoblast YAC-1, human lymphoblast K-562 and Jurkat [56] and human epithelial cervix HeLa [50] cells were cultured as described previously. All cell lines were purchased from ATCC.

N-terminal COFRADIC and LC-MS/MS analysis

N-terminal COFRADIC analyses were performed as described previously [103]. To enable the assignment of in vivo Nt-acetylation events, all primary protein amines were blocked making use of a (stable isotopic encoded) N-hydroxysuccinimide ester at the protein level (i.e. NHS esters of \(^{13}\)C\(_2\)D\(_3\) or D\(_3\) acetate).

LC-MS/MS analysis was performed as described previously ([50] and [48]). The generated MS/MS peak lists were searched with Mascot using the Mascot Daemon interface (version 2.2.0, Matrix Science). Searches were performed in the Swiss-Prot database with taxonomy set to human or mouse (UniProtKB/Swiss-Prot database version 2012_03, containing 20,254 human and 16,513 mouse entries (535,248 sequence entries in total)) or using custom databases (combination of UniProtKB/Swiss-Prot and ribo-seq derived translation sequences [86]).

\(^{13}\)C\(_2\)D\(_3\) or D\(_3\)-acetylation at lysines, carbamidomethylation of cysteine and methionine oxidation to methionine-sulfoxide were set as fixed modifications for the N-terminal COFRADIC analyses. Variable modifications were \(^{13}\)C\(_2\)D\(_3\)-acetylation, D\(_3\)-acetylation and acetylation of protein N-termini. Pyroglutamate formation of N-terminal glutamine was additionally set as a variable modification. Endoproteinase Arg-C/P (Arg-C specificity with arginine-proline cleavage allowed) was set as enzyme allowing no missed cleavages. The mass tolerance on the precursor ion was set to 10 ppm, 0.2 Da and 0.5 Da, and on fragment ions to 0.5 Da, 0.1 Da, 0.5 Da for the Orbitrap, Q-TOF Premier and Ion Trap analyses respectively. The peptide charge was set to 1+, 2+, 3+ and instrument setting was put to ESI-TRAP for Orbitrap and Ion Trap analyses and to ESI-QUAD-TOF for Q-TOF Premier analyses. Only peptides that were ranked one and scored above the threshold score, set at 99% confidence, were withheld. The estimated false discovery rate by searching decoy databases (a shuffled version of the yeast Swiss-Prot database made by the DBToolkit algorithm [104]) was found to lie below 1.5% on the spectrum level. All annotated highest scoring spectra of the N-terminal peptides reported in Appendix Table I.1 are provided as supplemental data.
Selection of N-termini

From the mouse and human N-terminal datasets, N-terminally blocked peptides were selected and classified. The high confident TIS encompass: i.) all (partially) in vivo Nα-acetylated N-termini and in vivo unmodified N-termini of which the start position corresponded with a Swiss-Prot isoform, Ensembl and/or TrEMBL annotated TIS site, ii.) iMet processed or iMet retaining counterparts of in vivo Nα-acetylated N-termini, iii.) N-termini matching TIS previously identified by ribosome profiling, iv.) N-termini annotated as dbTIS in (a) prior Swiss-Prot release(s), v.) N-termini for which the iMet processed and/or iMet retaining orthologous N-terminal peptide (HomoloGene) was identified as being (partially) Nα-acetylated in vivo.

Sequence logo analysis

All experimentally observed alternative N-termini were aligned based on their translation start codon. The N-terminal peptides lacking the initiator methionine (iMet) were preceded with the iMet to rule out codon shifts in the sequence logo creation. Afterwards, all peptides were mapped to their coding sequence (Perl scripting using the Ensembl API). Sequence logos were created based on the aligned transcript sequences (12 bp upstream and 9 bp downstream) using WebLogo 3 ([http://weblogo.threeplusone.com](http://weblogo.threeplusone.com), [105]). Sequence logos were plotted using both the residue probability and information content in bits as measure. Sequence logos were created for both the dTIS and corresponding dbTIS flanking regions. Also, an extra positive control to the dTIS sequence logos was generated based on 5,000 randomly selected coding sequences corresponding to annotated translation initiation sites from CCDS [106] proteins were aligned for nucleotide context logo creation.

Ribosome profiling and genome-wide visualization

Raw sequencing reads of the mESC ribosome profiling data [22] were downloaded from the Gene Expression Omnibus (dataset GSE30839). All reads from the control (cycloheximide treated, also referred to as CHX treated, sample GSM765292) and harringtonine treated (sample GSM765295) were remapped using bowtie (v.0.12.7) on the mouse genome (assembly version 37) using the protocol described [107]. All HEK293 cell line GTI-seq data [33] was downloaded from the NCBI Sequence Read Archive (accessions: SRX172392, SRX172361, SRX172360, SRX172315) and processed in the same way.

Genome-wide visualization of the experimental data, in combination with publicly available data, was accomplished using an in-house developed genome browser (H2G2, [http://h2g2.ugent.be/biobix.html](http://h2g2.ugent.be/biobix.html)). Information tracks containing the ribosome profile/GTI-seq mappings of the CHX treated samples (generating a translation profile all over the coding mRNA) and harringtonine/lactimidomycin treated samples (translation profile accumulation at the TIS) are available (see [22] for more information). Furthermore, an information track is constructed showing the predicted translation products, based on the TIS-predictions from Ingolia et al. [22] and Lee et al. [33] (for respectively the mESC and HEK293 cell line sample)
and the UCSC transcript annotation. The genomic locations of the N-terminal peptides identified by means of N-terminal COFRADIC are also visualized in the H2G2 browser. Other visualization tracks include genomic information from a local Ensembl [108] instance NCBIM37.66, transcript tracks holding the annotated TIS within the UniProtKB and Ensembl database and a conservation track based on the phastCons [76, 77] conservation scores amongst others. More information on how to use the H2G2 browser can be found in the Appendix File I.1.

**Genomic annotation of the identified dbTIS and aTIS**

All Swiss-Prot annotated N-termini (dbTIS) and alternative N-termini (aTIS) were mapped to their corresponding reference genome (GRCh37 for human and NCBIM37 for mouse) based on the UniProt-KB/Swiss-Prot accession number (or alternatively the Swiss-Prot gene name) and the N-terminal peptide sequence (PeptideMapper script based on the BioMart [78] and Ensembl [109] API, version 66). The genomic locations of the experimental aTIS and dbTIS locations were made available as a visualization track in the H2G2 genome browser (see above). Two projects are made available (named TIS Human and TIS Mouse) using a public login (see Appendix File I.1 for more details). These projects hold several visualization tracks (see fig 5 and 6 for examples): (A) an Ensembl gene track (B) an Ensembl transcript track (visible after mouse-click on the gene track), and tracks holding (C) the annotated TIS within the UniProtKB database split into Swiss-Prot and trEMBL (D) the annotated TIS within the Ensembl database (E) the reported aTIS and dbTIS locations (F) a conservation track based on the phastCons [76, 77] conservation scores based on alignment of 45 and 29 vertebrate genomes respectively for human and mouse. Furthermore all genes where an aTIS or dbTIS has been identified by means of the N-terminal COFRADIC experiments are listed in “GeneDigest” reports within the H2G2 genome browser environment.

The aTIS “GeneDigest” report lists all genes wherefore an alternative start site is reported, whereas the dbTIS “GeneDigest” report lists all genes of which a Swiss-Prot database annotated TIS has been identified. A third “Genedigest” report lists extra translation start sites identified from the N-terminomics experiments searching a protein product database constructed based on ribosome profiling sequence information [22] (see above). Further, experimental ribo-seq data [22] are also presented as custom tracks, allowing manual inspection of co-occurrence of N-terminal COFRADIC and ribo-seq experimental evidence.

**Conservation analysis**

To assess the evolutionary conservation potential of the identified d(b)TIS and their flanking sequences as compared to 5000 randomly chosen, BioMart [78] annotated CCDS translation initiation sites, their orthologous positions in various vertebrate genomes were extracted using phastCons [76, 77] and scored in a multiple sequence alignment. The phastCons program computes conservation scores based on a phylo-HMM, a type of probabilistic model that describes both the process of DNA substitution at each site in a genome and the
way this process changes from one site to the next. The value plotted at each site is the posterior probability that the corresponding alignment column was “generated” by the conserved state of the phylo-HMM.

**TargetP analysis**

To categorize the subcellular location of the proteins translated from their annotated versus their alternative N-termini, a TargetP prediction (v1.1b, [110]) was performed. The location assignment is based on a predicted N-terminal presequence: a mitochondrial targeting peptide, or a secretory pathway signal peptide.

**Generation of TIS mutagenized CDS and in vitro translation assays**

pOTB7 vectors (RZPD Imagenes, Germany) encoding aminoacyl tRNA synthase complex-interacting multifunctional protein 2 (AIMP2), inhibitor of kappa light polypeptide gene enhancer in B-cells kinase gamma (NEMO), Zinc finger protein 296 (ZN296), splicing factor 3a subunit 3 (SF3A3), cytoplasmic aspartate-tRNA ligase (SYDC), ubiquitin carboxyl-terminal hydrolase isozyme L1 (UCHL1), transcription factor jun-B (JUNB), cytokine receptor-like factor 3 (CRLF3), Nucleosome assembly protein 1-like 1 (NP1L1) and pyridoxamine 5’-phosphate oxidase (PNPO) served as templates for site-directed PCR-mutagenesis (QuickChange, Stratagene) according to the manufacturer’s instructions using the primer pairs indicated in Appendix Table I.2. The correctness of all (mutant) cDNA sequences generated was confirmed by DNA-sequencing.

The mutagenized constructs were used as templates for in vitro coupled transcription/translation in a rabbit reticulocyte lysate system according to the manufacturer’s instructions (IVTT; Promega) to generate $[^{35}\text{S}]$methionine labeled protein products. 5 µl of the translate reaction was diluted 10-fold in 10 mM Tris pH 8.0. NuPAGE LDS Sample Buffer (Invitrogen) was added and the samples heated for 10 min at 70 °C. Samples were separated on 4-12% NuPAGE Bis-Tris gradient gels (1.0 mm x 12 well) (Invitrogen) using MOPS Buffer. Subsequently, proteins were transferred onto a PVDF membrane, air-dried and exposed to a film suitable for radiographic detection (ECL Hyperfilms, Amersham Biosciences, Buckinghamshire, UK). Radiolabeled proteins were visualized by radiography.
Author contributions

P.V.D. and G.M. designed research; P.V.D., D.G. and G.M. performed research; P.V.D. and G.M. analyzed data; P.V.D. wrote the paper; W.V.C. supervised research.
References

Discovery of novel N-terminal proteoforms


RESULTS
Chapter II
Assessment of N-terminal proteoform stability
Adapted from:

RESULTS Chapter II

Assessment of N-terminal proteoform stability

Abstract

To understand the impact of alternative translation initiation on a proteome, we performed a proteome-wide study on protein turnover using positional proteomics and ribosome profiling to distinguish between N-terminal proteoforms of individual genes. By combining pulsed SILAC with N-terminal COFRADIC we monitored the stability of 1,941 human N-terminal proteoforms, including 147 N-terminal proteoform pairs that originate from alternative translation initiation, alternative splicing or incomplete processing of the initiator methionine. N-terminally truncated proteoforms were less abundant than canonical proteoforms and often displayed altered stabilities, likely attributed to individual protein characteristics, including intrinsic disorder, but independent of N-terminal amino acid identity or truncation length. We discovered that the removal of initiator methionine by methionine aminopeptidases reduced the stability of processed proteoforms, while susceptibility for N-terminal acetylation did not seem to influence protein turnover rates. Taken together, our findings reveal differences in protein stability between N-terminal proteoforms and point to a role for alternative translation initiation and co-translational initiator methionine removal, next to alternative splicing, in the overall regulation of proteome homeostasis.
Introduction

Protein stability has been suggested to reflect the individual properties and biological functions of proteins [1, 2]. In eukaryotic cells, a wide variety of proteins with roles in cell cycle progression, signal transduction and metabolic regulation are quickly degraded by the proteasome [3], whereas some constituents of abundant protein complexes (such as the ribosome and the spliceosome) are extremely stable [1, 2]. Short-lived proteins are predicted to have an increased aggregation potential, which often places them central in protein deposition diseases correlated with ageing and reduced proteasome activity [4]. Estimations of the impact of protein turnover on gene regulation revealed that protein synthesis and degradation rates are comparable for proteins within functional protein classes and among individual components of macromolecular protein complexes [5]. As such, protein stability is undoubtedly a key component that regulates the outcome of gene expression and shapes cellular phenotypes. Large-scale studies of proteome dynamics in living cells were made possible by applying stable isotope labeling, including SILAC (stable isotope labeling by amino acids in cell culture), in combination with mass spectrometry. To measure protein turnover, cells are usually pulsed with media containing stable isotopic variants of essential amino acids and the rate of incorporation of these newly added isotopes then enables calculating the protein half-life [1, 6, 7].

In several recent N-terminomics analyses using mouse and human cell lines, we showed that the prevalence of translation start sites not annotated in protein sequence databases was underestimated. More specifically, we found that 10-20% of all protein N-termini resulted from alternative translation initiation events, pointed to inadequately annotated protein start sites or to alternative splicing [8-10]. Scarce evidence indicates that alternative translation initiation can produce proteins variants – proteoforms [11] – that differ in stability when compared to their database-annotated counterparts [12]. As an example, alternative translation initiation of the opioid receptor that results in an N-terminal protein extension rich in lysine residues and prone to ubiquitination, contributes to the high degradation rate of this proteoform [12].

Triggered by such atomistic studies, we set out to study on a more global scale if N-terminal proteoforms derived from the same gene showed different stabilities. In contrast to other proteome studies that do not distinguish between N-terminal proteoforms, we here combine SILAC pulse-labeling [2] (pSILAC) with N-terminal COFRADIC (COmbined FRActional Dlagonal Chromatography) [13] to study the stability and turnover rates of proteoforms with heterogeneous N-termini. Important here is that we rely on known (previously identified [9]) N-terminal modifications to unambiguously assign an identified peptide as a proxy for a protein’s N-terminus. Indeed, N-termini of eukaryotic proteins typically undergo co-translational modifications, which can be used to discriminate N-termini pointing to translation initiation events from those pointing to proteolytic cleavage using mass spectrometry [14]. For example, methionine aminopeptidases typically remove the initiator
Met (iMet) if the side chain of the next amino acid has a small gyration radius like found in alanine (Ala), valine (Val), serine (Ser), threonine (Thr), cysteine (Cys), glycine (Gly) or proline (Pro) [15]. Subsequently, the alpha-amino group of the resulting N-terminus, with the exception of Pro, are frequently modified by N-terminal acetylation (Nt-acetylation) catalysed by N-terminal acetyltransferases (Nt-acetyltransferases or NATs) [16]. iMet-retaining N-termini may also be Nt-acetylated, largely depending on the sequence specificity of the NAT involved [17, 18]. Taken together, up to 90% of all human proteins undergo either partial or complete Nt-acetylation [19]. For several cases Nt-acetylation was shown to alter protein interactions, functions and structure [20-23]. Moreover, protein N-termini have other unique properties, such as increased disorder (the lack of a fixed dimensional structure) [24]. Surprisingly, the length of N-terminal disorder strongly correlates with the probability of protein Nt-acetylation, with highly disordered sequences being more frequently Nt-acetylated. As a consequence, it was suggested that Nt-acetylation may play a role in structural stabilization. More specifically, Nt-acetylation of alpha-synuclein was shown to stabilize the helicity of its N-terminus [22] and loss of Nt-acetylation seems to inhibit aggregation of prion proteins [24]. Moreover, Nt-acetylation may have important roles in protein stability. Although it was long believed that Nt-acetylation may prevent protein degradation by the ubiquitin system [25], other studies indicated that Nt-acetylated proteins are not significantly more stable or shielded from proteolytic degradation [26]. The N-end rule further explores the relationship between the in vivo protein half-life and the identity of a protein’s N-terminal amino acid(s). In summary, the N-end rule links the removal of protein substrates covering the whole spectrum of N-terminal identities, except of Pro- and Gly-starting N-termini. Degradation signals encoded by N-terminal residues, referred to as N-degrons, may be recognised by specialised E3 ubiquitin ligases and the N-end rule consists of two major branches, being the Arg/N-end rule and the Ac/N-end rule pathways. The Arg/N-end rule targets unmodified iMet-starting N-termini followed by a hydrophobic amino acid [27] and neo-N-termini generated upon proteolytic cleavage [27, 28]. The acetylation-dependent branch of the N-end-rule (Ac/N-end rule) mediates a targeted degradation of proteins with Nt-acetylated Met, Arg, Val, Ser, Thr and Cys residues [29]. Interestingly, Nt-acetylation of two N-terminal proteoforms raised by alternative initiation of SNC1 in Arabidopsis thaliana had an opossing effect on proteoform stability dependent on the NAT involved [30]. The longest form was destabilised by NatA, whereas NatB mediated Nt-acetylation increased the stability of the truncated proteoform. On the other hand, recent studies of human Ogden syndrome caused by a NatA loss-of-funtion mutation indicated that Nt-acetylation was involved in the increased stability of the NatA substrate THOC7 [31]. Next to the N-end rule, additional mechanisms have been shown to repress undesirable protein degradation. For example, N-degrons may be shielded by proper folding of the mature protein or by integration into macromolecular complexes [21, 32, 33].

To summarize, alternative translation initiation and co-translational modifications seem to contribute substantially to the diversity of protein N-termini and are implicated in the
regulation of protein turnover. Evidently, a comprehensive characterisation of N-terminal proteoforms linked to their stability is of great importance.

Results

We combined N-terminal COFRADIC, a positional proteomics technology that enables the enrichment of N-terminal peptides, with pSILAC for the identification of N-terminal proteoforms and characterisation of their cellular stability in human Jurkat T-lymphocytes. Following database searching, we retained N-termini starting at position 1 or 2 that point to “database-annotated” translation initiation sites (dbTIS). Next, we inspected N-terminal peptides starting beyond position 2 and selected those peptides that comply with the rules of iMet processing and Nt-acetylation [16, 18]. Proteoforms linked to these N-terminal peptides were expected to be mainly derived from the use of alternative translation initiation sites (aTIS).

Overall, we identified 2,578 N-terminal peptides (Appendix Table II.1A), including 2,135 dbTIS and 443 aTIS. Our study demonstrates that N-terminal proteoforms raised upon translation initiation and alternative iMet processing contribute substantially to the diversity of the human proteome. Indeed, we identified 335 N-terminal proteoforms derived from 133 genes displaying multiple N-termini raised from different translation initiation sites (Fig 1A). Additionally, we observed 81 proteoforms resulting from partial iMet processing (Fig 1A). Of note, 63% of all aTIS were identified in the absence of their canonical counterparts.

Diversity of N-terminal proteoforms

Deviations from the canonical translation mechanism such as ribosome leaky scanning and reinitiation [34, 35] allow for selection of alternative translation initiation sites within one transcript and, together with mRNA splicing, contribute to the N-terminal proteoform diversity observed in our study. Leaky scanning is a mechanism by which a fraction of ribosomes omits less favourable start codons to initiate translation downstream [36, 37]. Ribosome profiling previously demonstrated [38] that leaky scanning is responsible for most TIS selection events and that the efficiency of downstream initiation is heavily dependent on the strength of the translation initiation context surrounding dbTIS, a finding corroborated by our previously obtained N-terminomics data [9]. To study the contribution of (alternative) translation initiation and splicing to N-terminal proteoform expression in more detail, we performed ribosome profiling experiments in Jurkat cells. Additionally, the N-terminal peptides identified in our proteomics experiment in the Swiss-Prot database were mapped to alternative databases; human Swiss-Prot Isoform, UniProt TrEMBL and Ensembl.

Ribosome profiling, or ribo-seq, enables studying in vivo protein synthesis by deep sequencing of ribosome-associated mRNA fragments, thereby providing a genome-wide snapshot of actively translated mRNAs. Additionally, (alternative) translation initiation can
Assessment of N-terminal proteoform stability

be studied with sub-codon to single-nucleotide resolution through the use of antibiotics such lactimidomycin (LTM), which exclusively inhibit initiating ribosomes [38].

Figure 1 - The origin of N-terminal proteoforms. A. 2,578 proteoforms were identified including 2,135 dbTIS and 443 aTIS. For the majority of proteins we found a unique translation initiation site (grey). Additionally, 335 N-terminal peptides pointed to proteins with multiple N-termini created by alternative translation initiation (blue), whereas 81 N-termini were generated by alternative iMet processing (green). B-C. Proteomics-derived translation initiation events were verified at the transcript level using ribo-seq data obtained in Jurkat cells and HCT116 cells (data unpublished). Although in this study, the degree of in vivo Nt-acetylation could not be directly determined, all retained N-termini were compliant with the rules of N-terminal processing [9]. In the case N-termini were found to be in vivo Nt-acetylated (previous MS/MS-based evidence [9]) this information is indicated. dbTIS and aTIS peptides were further mapped to protein N-termini annotated in Swiss-Prot Isoform, UniProt TrEMBL and Ensembl databases and neo-N-termini (proteolytic products) of the knowledgebase TopFind 3.0. The number of matching N-termini is indicated in dark grey. N-terminal peptides supported by at least one source of metadata pointing to translation (Nt-acetylation, ribo-seq, Swiss-Prot Isoforms, TrEMBL or Ensembl) were classified as highly confident TIS (indicated in green). D. Using experimental and metadata, we assigned the most likely origin of aTIS peptides and confirmed that alternative translation initiation within the same transcript (leaky scanning) contributes to the majority of alternative TIS identified.
Using ribosome profiling, of the 2,578 N-terminal proteoforms identified by means of N-terminomics, we were able to confirm translation initiation events for 89% of dbTIS (1,895 of 2,135) and 29% of aTIS (130 of 443 aTIS) in Jurkat cells (Fig 1B–C). Using a Mann-Whitney one sided test we further confirmed that aTIS detected by ribo-seq (130) had significantly lower translation initiation signal (R_LT-MCHX) at the start codon compared to dbTIS (1,895) (p-value = 0.011). The same was observed when directly comparing dbTIS to aTIS variants (31 pairs) detected by ribo-seq and originating from the same gene (Wilcoxon signed-rank test for paired samples; p-value of 0.00068). Further, from the data mapping to Ensembl and Swiss-Prot Isoforms databases, we found that 23% of the aTIS proteoforms (101 out of 443) could likely be explained by a database annotated alternative splicing event (Fig 1D). Additionally, only 19 aTIS N-termini were annotated in the proteolytic knowledgebase of protein N-termini, TopFIND 3.0, as potential protease cleavage products [39] (Fig 1C). However, 7 of these 19 could be disregarded due to Nt-acetylation evidence and/or ribo-seq confirmed TIS selection. Overall, given the N-terminal selection strategies applied and the metadata at hand, only very few of the here reported N-termini can be considered as potential proteolytic products. The majority of remaining N-terminal proteoforms (330) is thus likely generated upon leaky scanning or other translation initiation mechanism as previously shown by others and us [9, 38]. Leaky scanning frequently results in initiation at a downstream start codon in the immediate proximity of the first AUG codon (Fig 2A) [9], however, supported by ribo-seq data, we found examples of translation initiation by means of leaky scanning at more distant start codons (Fig 2B). Relying on the current Ensembl human transcriptome annotation and splice variants present in the Swiss-Prot Isoform database, we were additionally able to confirm the expression of N-terminal proteoforms from alternatively spliced transcripts (Fig 2C).

Interestingly, we found 12 aTIS proteoforms being exclusively expressed in the absence of dbTIS, judging from the lack of preceding ribo-seq and proteomics evidence [9] (Fig 2D). The two most likely scenarios leading to the creation of N-terminal proteoforms starting at two consecutive Met residues (for example at position 1 and 2), are leaky scanning or alternative iMet processing. Our proteomics data provides evidence of consecutive Met-starting N-termini for 11 genes and matching ribo-seq data indicate that translation initiation may occur at both Met residues (Fig 2E-F).

**Protein turnover measurement**

To measure turnover times of N-terminal proteoforms we used human Jurkat T-lymphocytes that were pre-labelled with either light (Arg⁰) or medium (Arg⁶) isotopes of arginine and harvested 0.5, 1.5, 4, 8, 12, 24 or 48 hours after refreshing light medium or swapping the Arg⁶ label to heavy arginine (Arg¹⁰) (see Fig 3A). Note that arginine was used as the SILAC label given that by nature of the N-terminal COFRADIC procedure, all N-terminal peptides end on arginine. After mixing equal amounts of light and medium/heavy labelled proteome
Assessment of N-terminal proteoform stability

samples taken at the 7 different time points, N-terminal COFRADIC analyses were performed.

Figure 2 - Examples of translation initiation events confirmed by ribo-seq in Jurkat cells. A-B. Alternative translation initiation due to leaky scanning. C. Alternative splicing. D. The preferential expression of an N-terminally truncated proteoform. E-F. Evidence for alternative translation initiation occurring at two consecutive AUG codons. Grey arrows mark the identified TIS. Transcripts were colored accordingly: coding sequence in green, 5’ leader in red and introns in grey. Jurkat LTM and Jurkat CHX tracks correspond to translation initiation and elongation signal captured by ribo-seq.

From the total of 2,578 N-terminal peptides identified (Appendix Table II.1A) we retained 1,972 peptides that were identified in at least 3 of the time points analysed (Appendix Table II.1B). Their SILAC ratio values were then used to calculate protein stabilities by monitoring protein degradation and synthesis based on changes in M/L and H/L isotope ratios, with the crossing point between the degradation and synthesis profiles enabling a direct readout of the 50% protein turnover (Fig 3B) [2]. Using a simple exponential model we calculated turnover times for 1,928 (98%) of the 1,972 proteoforms with minimal $R^2$ coefficients $\geq 0.8$. To evaluate the relationship between the number of observations and the predictive power of our model, we re-calculated turnover times of selected proteoforms.
based on a subset of available observations showing that turnover times can be precisely reproduced using as few as 3 randomly selected data points (Appendix Fig II.1). Our model very well described the variation in experimental data (as supported by a high median R² value of 0.98; see Appendix Fig. II.2), overall demonstrating that protein degradation in Jurkat cells follows mono-exponential kinetics. Only a minority of degradation and synthesis curves (2%) showed monotonic changes in time not following an exponential trend (R² < 0.8). In 13 such cases turnover times were determined using a linear interpolation of the crossing point, but otherwise such proteoforms were excluded from further analysis. In total, we could assign 50% turnover times to 1,941 N-terminal proteoforms (Appendix Table II.1B).

Figure 3 - Experimental setup using pSILAC and N-terminal COFRADIC to assess N-terminal proteoform stability. A. Jurkat cells were pre-labeled with light or medium L-Arg isotopes. A label swap was performed of cells growing in medium Arg, whereas light cells were cultured in the same medium. Cells were harvested at different time points and equal proteome amounts were mixed, followed by N-terminal COFRADIC fractionation and LC-MS/MS analysis. B. Ratios between three dynamic forms of N-terminal peptides reflect protein synthesis and degradation rates, allowing the calculation of 50% turnover times. C. The
Assessment of N-terminal proteoform stability

distribution of turnover times (N=1,894) calculated was not unimodal, with a sizable number of quickly turned over proteins and the median protein turnover rate being 21.6 h. The doubling time of Jurkat cells (24 h) is indicated in red.

To account for cell growth, our exponential model of protein degradation incorporated the doubling time measured for Jurkat cells (24 h, see Appendix Fig. II.3). This allowed for the determination of two additional parameters, namely protein degradation rate constants (k_{deg}) and protein half-life values [h]. Since 50% turnover values can be measured more directly and reflect both protein synthesis and degradation, we selected 50% turnover time as the desired parameter to perform our subsequent analyses.

Protein function and turnover: correlation with previous reports

Overall, protein turnover times in Jurkat cells varied from less than 1, to more than 48 hours with a median turnover time of 21.6 h (Fig 3C). Protein degradation constants determined for Jurkat cells (control sample) [40], protein turnover times measured in HeLa cells (whole cell lysates) [2] and protein half-lives derived from NIH3T3 mouse fibroblasts [7] were retrieved and compared to their corresponding values of canonical proteoforms quantified in our study. We detected a moderate correlation between our data and previous reports, with Spearman rank correlation coefficients reaching 0.49 for HeLa data [2], 0.48 for NIH3T3 data [7] and 0.44 for Jurkat data [40]. Moreover, a median protein half-life of 50.5 hours measured in our study was in good agreement with previously reported values in mouse NIH3T3 cells (47.8 h) [7] and human Jurkat cells (55.8 h) [40]. Further, we observed a highly similar median turnover time (21.6 h) and turnover distribution to estimates obtained in HeLa cells (20 h) [2].

Based on gene ontology (GO) biological process term enrichment of our data using the Gorilla software [41] and considering one proteoform per gene, we found that proteins with common functional annotations displayed comparable cellular stabilities. To reduce the complexity and redundancy of enriched categories we applied REViGIO [42] and visualised selected GO terms with false discovery rate (FDR) q-values lower than 0.05. Lower stability was detected for proteins involved in mitosis, chromosome segregation, cell cycle and apoptosis, next to regulation of RNA transcription and protein ubiquitination (Fig 4A). On the other hand, RNA splicing, protein translation, folding and transport as well as various metabolic processes of carbohydrates, nucleotides, aldehydes, ketones, glutathione, nitrogen and phosphorus were conducted by more stable proteins (Fig 4B). Along the same line, category enrichment analysis of KEGG pathway terms clearly pointed to increased stability of spliceosome and ribosome components next to RNA degradation, aminoacyl-tRNA biosynthesis, glycolysis and gluconeogenesis, and pentose phosphate pathways (all with FDR values below 9.0E-04) [2, 7]. Pfam data did not point to a significant influence of protein domain composition on protein turnover.
Overall, the results of our annotation enrichment analysis agree very well with previous studies in mouse [7] and human [2, 40] and thus, protein turnover rates seem to be conserved between different cell lines and organisms, especially at the level of protein groups and biological processes.

Figure 4 - GO term enrichment analysis for unstable (A) and stable (B) proteins. Horizontal bar chart representations are given for significantly enriched GO terms in the human proteome (FDR q-value ≤ 0.05).

Stability, abundance and structural disorder of alternative proteoforms

To facilitate the interpretation of the relationship between alternative translation and protein turnover, whenever an iMet-retaining and -processed proteoform was identified, we only selected the expected mature N-terminus per proteoform, according to the rules of
Assessment of N-terminal proteoform stability

iMet processing [15]. In total, we assigned protein turnover rates for 323 aTIS and 1,571 dbTIS N-terminal proteoforms (Fig 5A).

**Figure 5 - Characterisation of identified N-terminal proteoforms.** A. We assigned the turnover time of 1,941 proteoforms, including 1,894 proteoforms derived from 1,571 dbTIS and 323 aTIS next to 47 N-termini generated by alternative iMet processing. dbTIS and aTIS derived proteoforms are indicated in shades of blue and shades of red, respectively. Stability of 100 N-terminal proteoform pairs derived from the same gene (dark red). B. Distribution of iMet positions giving rise to 323 N-terminal proteoforms with assigned turnover times (to avoid redundancy a single iMet processed proteoform was considered). C. Distribution of turnover values shows differences in the overall stability of dbTIS versus aTIS proteoforms.

The majority of the identified aTIS resulted in relatively short N-terminal protein truncations (protein length reduced by less than 13%). The distribution of aTIS corresponding iMet positions centered within the first 200 amino acids with a median position of 52 (Fig 5B), an observation in line with the fact that leaky ribosome scanning is likely to be responsible for the majority of downstream translation events observed (see above and [9, 38]).

Interestingly, N-terminal proteoforms derived from aTIS usage frequently displayed distinct stabilities (Fig 5C). We were able to quantify turnover rates for 100 dbTIS/aTIS pairs and observed that some aTIS displayed altered stabilities as compared to their dbTIS counterparts. More specifically, we detected numerous examples of aTIS proteoforms with increased as well as decreased stability (Fig 6A-B). This effect appeared not to correlate with the distance between aTIS and dbTIS positions (Fig 6C). Since the differences in stabilities
between aTIS and dbTIS expressed forms of the same gene were not uniform, no statistical difference was detected between their turnover times (Fig 6C, Table 1). However, when we compared the entire populations of aTIS to dbTIS, alternative proteoforms were on average found to be significantly more stable (p-value=1.52e-08) (Fig 5C, Table 1).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Dataset</th>
<th>Test</th>
<th>p-value</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>turnover time</td>
<td>complete dataset</td>
<td>Mann-Whitney test</td>
<td>1.52e-08</td>
<td>aTIS &gt; dbTIS</td>
</tr>
<tr>
<td></td>
<td>dbTIS/aTIS pairs</td>
<td>Wilcoxon signed-rank test</td>
<td>0.5021</td>
<td>no difference</td>
</tr>
<tr>
<td>disorder</td>
<td>complete dataset</td>
<td>Mann-Whitney test</td>
<td>0.00157</td>
<td>aTIS &gt; dbTIS</td>
</tr>
<tr>
<td></td>
<td>dbTIS/aTIS pairs</td>
<td>Wilcoxon signed-rank test</td>
<td>0.01216</td>
<td>aTIS &gt; dbTIS</td>
</tr>
<tr>
<td>spectral counts</td>
<td>complete dataset</td>
<td>Mann-Whitney test</td>
<td>4.31e-09</td>
<td>aTIS &lt; dbTIS</td>
</tr>
<tr>
<td></td>
<td>dbTIS/aTIS pairs</td>
<td>Wilcoxon signed-rank test</td>
<td>3.44e-09</td>
<td>aTIS &lt; dbTIS</td>
</tr>
<tr>
<td>NSAF</td>
<td>complete dataset</td>
<td>Mann-Whitney test</td>
<td>1.15e-07</td>
<td>aTIS &lt; dbTIS</td>
</tr>
<tr>
<td></td>
<td>dbTIS/aTIS pairs</td>
<td>Wilcoxon signed-rank test</td>
<td>6.45e-09</td>
<td>aTIS &lt; dbTIS</td>
</tr>
</tbody>
</table>

Table 1 - Differences in turnover, disorder and abundance values between aTIS and dbTIS.

For the complete dataset of aTIS and dbTIS with reliable turnover measurements and considering a single iMet proteoform (N=1,894; 323 aTIS and 1,571 dbTIS), a comparison of turnover, disorder and abundance values was performed using Mann-Whitney one-sided test for independent samples. Differences in paired dbTIS-aTIS data were measured using a Wilcoxon signed-rank one-sided test for paired samples. All analyses were performed at p ≤ 0.05 significance level.

We further determined if aTIS indicative N-termini share features that are distinct from dbTIS. Given that translation is a one-directional process and downstream initiation due to leaky scanning only occurs with a limited probability, in a normal cellular context one might expect that many aTIS variants are less efficiently produced than dbTIS variants, leading to lower abundance of the former [9, 43]. To investigate this we used spectral counts as well as NSAF scores [44] as quantitative proteoform measures. Using a Mann-Whitney one sided test to compare the aTIS (323) and dbTIS (1,571) groups we confirmed that, on average, aTIS variants are less abundant with high significance values when using both spectral counts (p-value= 4.31e-09) as well as NSAF scores (p-value= 1.15e-07). The same was observed when directly comparing dbTIS to aTIS variants originating from the same gene using a Wilcoxon signed-rank test for paired samples (p-values of 3.44e-09 and 6.45e-09 respectively for spectral counts and NSAF, see Table 1). Finally, we observed a slight positive correlation between turnover time and protein abundance (Spearman rank correlation of 0.285). Further analysis indicated reduced levels of the top 10% of unstable proteins and higher levels of the 10% most stable proteins in relation to the mean protein abundance in our dataset (Benjamini–Hochberg FDR of 2.29e-11 and 7.44e-05, respectively).

Intrinsic disorder in protein structures is an important factor that determines protein degradation rates (reviewed in [45]). The extent of structural disorder can be assessed,
Assessment of N-terminal proteoform stability

amongst others, by studying the amino acid composition of proteins [46]. For every proteoform sequence, the percentage of intrinsically disordered regions was predicted using the RAPID server [47]. The average disorder of aTIS proteoforms was found to be higher than that of dbTIS proteoforms (Mann-Whitney one sided test, p=0.00157). These observations were confirmed in paired samples of aTIS and dbTIS products of the same gene (Wilcoxon signed-rank test, p=0.01216, see Fig 6D). In conclusion, aTIS proteoforms were characterized by lower expression levels and higher structural disorder than dbTIS proteoforms. Moreover, the aTIS population displayed an overall higher stability compared to dbTIS. Additionally, a

Figure 6 - N-terminal proteoforms may display different stabilities. A. Degradation profiles of AIMP2, MARE2 and AN32E dbTIS and aTIS proteoforms as representative examples of N-terminal proteoforms that differ in stability. B. MS spectra of the N-terminals of two proteoforms originating from the aminoacyl tRNA synthase complex-interacting multifunctional protein 2 (AIMP2): a dbTIS indicative N-terminus with processed iMet and starting at the second amino acid Pro, an aTIS indicative peptide starting at position 3 and retaining its iMet. As evident from the MS spectra shown, these N-terminal proteoforms displayed largely different turnover rates. C. Differences in turnover times were calculated for 100 dbTIS and aTIS proteoform pairs and presented in ascending order. The extent of N-terminal truncation in each pair was colour-coded to visualize the absence of correlation between the number of lacking amino acids and the direction of turnover shift. D. An increase in disorder content between aTIS and dbTIS proteoforms can be observed for a multitude of dbTIS/aTIS pairs.
low negative correlation was observed between protein turnover and the percentage of disordered sequence (Spearman rank correlation of -0.209), suggesting that unstable proteins tend to have more disordered regions, as reported in previous studies [5, 6]. In line with this, the 10% most stable proteins had significantly lower disorder content, contrary to the 10% least stable proteins with highly disordered structures (Benjamini–Hochberg FDR of 3.27e-11).

**Protein turnover and N-terminal amino acid composition**

We next investigated if the identified N-termini hold intrinsic information on the so-called N-degrons [27], and what the role is of iMet cleavage in this process.

To correlate the N-end rule with our pSILAC dataset we examined the relation between the turnover rate and the identity of the two utmost N-terminal residues. A Bartlett test indicated insufficient homogeneity of variance between the studied groups, leading to the selection of non-parametrical tests for further analysis. First, we performed a pairwise comparison of protein stability across N-termini grouped according to their N-terminal amino acid identity. A Kruskal-Wallis rank sum test indicated a significant impact of N-terminal residues on protein turnover, whereas a subsequent multiple comparison test allowed to indicate the most deviating pairs (with p<0.05). Met-Asn and Met-Asp were among the least stable N-termini which deviated significantly from stable N-termini such as Met-Thr, Met-Val, Met-Ala and Pro starting N-termini (Fig 7A).

To further dissect our data, we compared the stability of N-termini undergoing iMet processing, leading to N-terminal Ser, Ala, Thr, Cys, Val, Gly and Pro. Interestingly, a general lower stability was observed for proteoforms with their iMet residue removed compared to proteoforms that retained the iMet. A significant impact of iMet processing on turnover rates was confirmed using a Kruskal-Wallis rank sum test (p < 2.2e-16). To decouple the effect of Nt-acetylation from iMet processing we compared completely or partially Nt-acetylated (Ser, Ala, Thr, Cys, Val, Gly) to in vivo free (Pro) N-termini and for both groups we observed a lower stability of proteoforms with their iMet removed (see Fig 7B). Considering individual amino acids, only Thr and Val deviated significantly from their corresponding iMet retaining N-termini (see Fig 7B-C). However, a less strict analysis with pairwise Wilcoxon rank sum test (without correction for multiple testing) pointed to significant deviation between iMet-processed and -retaining Ala-, Ser- and Gly- starting N-termini. To conclude, turnover rates of N-termini with different susceptibility for Nt-acetylation [18] were generally affected by iMet processing.

These findings suggest that irrespective of (the degree of) Nt-acetylation, iMet processing, rather than Nt-acetylation, seems to be a main contributing factor to N-terminal proteoform stability. In particular and in line with previous findings [26], no correlation was found with the Nt-acetylation susceptibility of protein N-termini (Fig 7B). Secondly, the presence of a hydrophobic residue at the second position has no apparent influence on average turnover
Assessment of N-terminal proteoform stability

rates as proteins starting with N-termini such as Met-Leu, Met-Phe, Met-Tyr and Met-Ile have stabilities close to the median stability of all protein N-termini (Fig 7A). A more detailed analysis of grand average of hydropathy (GRAVY) of 2 and 10 most N-terminal residues revealed no correlation with protein turnover times (Spearman correlation coefficients of 0.013 and 0.0085 respectively).

Figure 7 - Relationship between N-terminal amino residue(s) identity and 50% turnover rates. A. All protein N-termini were grouped according to the identity of their N-terminal
residues and sorted in ascending order of stability. B. Methionine processing typically affects N-termini displaying amino acids with small gyration radii (S, A, T, C, V, G or P residues at the second position). Although the majority of these N-termini undergo methionine cleavage, some (partially) retain their first amino acid. Subsequently, iMet processed N-termini may be Nt-acetylated by NatA in the following order of efficiency (S, A, T, C, V and G) with the exception of P which always remains unmodified [18]. The stability of iMet processed N-termini is very similar for S, A, T, C and G and P starting N-termini, but their iMet retaining counterparts have overall higher stability, and this especially for MT and MV N-termini. In contrast to methionine processing, Nt-acetylation susceptibility does not appear to have an effect on the stability of proteoforms. *The extent of in vivo Nt-acetylation was described by Van Damme et al. [18]. C. Differences in turnover times where calculated for N-termini starting with S, A, T, C and G and P and their iMet retaining counterparts (47 proteoform pairs with valid turnover measurements). Decreased stability of processed N-termini could be observed, especially for T and V starting N-termini.

Ubiquitination and protein turnover

Post-translational modifications, such as ubiquitination, may affect protein stability and are thus worthwhile to consider in the context of N-terminal proteoforms. Ubiquitination has a known, direct impact on protein degradation and was previously studied at the proteome-wide level using Jurkat cells as a model system [48]. Our group reported on a compilation of more than 7,500 in vivo ubiquitinated peptides in more than 3,300 different proteins [48]. Using this dataset, we investigated if annotated N-terminal proteoforms holding possible (previously identified) ubiquitination sites displayed different stabilities than proteoforms not found to be ubiquitinated. We compared the stability of 348 dbTIS containing at least one detected ubiquitination site to 388 N-termini with one or more lysine residues but no experimental evidence of ubiquitination in Jurkat cells. A one-sided Mann-Whitney test revealed a significant higher median stability of ubiquitinated proteoforms compared to non-ubiquitinated proteoforms (p= 0.0005; see Appendix Fig. II.4), although in each case a very broad range of protein stabilities could be observed. To provide more insight, we decided to investigate particular cases of ubiquitinated N-termini. The existence of additional ubiquitination sites may contribute to the difference in turnover observed between longer and shorter proteoforms, as reported in the case of the mu opioid receptor (MOP) [12]. Interestingly, in our dataset, we found 31 dbTIS proteoforms holding unique ubiquitination sites not present in a shorter proteoform (aTIS). 11 such proteoforms displayed reduced stability (turnover time decreased by at least 2 hours) compared to their aTIS counterparts (Appendix Table II.2). For these proteoforms ubiquitination potentially contributes to the observed differences in stability. We also found 14 proteoform pairs where ubiquitination of the longer variant possible contributed to an increased stability and 5 cases where irrelevant differences in stability between proteoform pairs were observed (<2 h).
Assessment of N-terminal proteoform stability

Protein turnover in macromolecular complexes

Following up on a recent report on the precise regulation of protein synthesis rates being proportional to the stoichiometry of macromolecular protein complexes [49], we here investigated if protein synthesis and degradation are coordinated for members of known macromolecular protein complexes (Appendix Table II.1B). According to the strategy reported by [5], we retrieved the human core protein complexes from the CORUM database [50]. Next, we assigned turnover times to members of complexes using only the dbTIS proteoforms identified in our study. We calculated the absolute difference in stability between every protein and all other members of a given complex. Significant outliers with statistically larger differences to other proteins in a complex were identified using the Mann-Whitney test ($\alpha=0.05$, one-sided). In this way we estimated the dispersion of turnover rates for 333 proteins within 173 macromolecular complexes with 3 or more members identified, and came to the conclusion that a majority of turnover rates correlate well within a complex (127 or 73% of the complexes). Inconsistencies were observed in 46 complexes of which 74% involved a single outlier. In 28 complexes we detected a single unstable component, hinting to a rate-limiting factor for complex formation.

Next, we investigated if protein-protein interactions could be affected by the expression of alternative N-terminal proteoforms. Here, a comparative analysis of GO terms related to protein function in aTIS- versus dbTIS-derived proteoforms revealed that alternative proteoforms were clearly enriched in protein forming homomultimers (FDR q-value of 0.0096) hinting to the likely impact of alternative proteoform incorporation on the overall stability of protein complexes.

Verification of protein turnover times

Cycloheximide (CHX) pulse-chase experiments were performed to confirm the stability of selected proteoforms. Half-lives of 6 selected endogenous proteins (Fig 8A) were assayed by western blotting using lysates of Jurkat cells treated with CHX for 0 to 24 hours. Protein levels of short-lived securin, lamin B, $\beta$-catenin and GCIP-interacting protein p29 levels were monitored by western blotting and compared to those of stable proteins such as actin and GAPDH. Protein half-lives deduced from western blot analyses corresponded well to those calculated from pSILAC data. Overexpression of selected V5-tagged proteoforms in the human colorectal HCT116 carcinoma cell line allowed us to validate differential turnover of identified dbTIS and aTIS products from two different genes; MARE2 and AN32E (Fig 8B-C). Data from the latter experiments further suggest that common trends can be observed in the regulation of proteoform stability across different human cell lines, despite discrepancies in the exact values calculated using CHX-chase and pSILAC.
Figure 8 - Comparison of protein turnover measured by pSilAC and CHX-chase. A. Jurkat cells were treated with 100 µg/ml CHX for 0, 0.5, 1.5, 4, 8, 12 or 24 hours. Protein degradation was monitored by western blotting and stabilities of several short-lived endogenous proteins were confirmed using antibodies (including lamin B, securin, β-catenin, GCIP interacting protein p29) and compared to stable proteins (such as GAPDH and actin). Proteoform-specific bands were used to calculate protein half-lives which were in a good agreement with turnover times obtained from pSilAC B-C. Validation of the differential turnover time of the dbTIS and aTIS derived proteoforms of MARE2 and AN32E. Selective C-terminal V5-tagged proteoforms were overexpressed in HCT116 cells for 24 hours and CHX pulse-chase experiments were performed as described for Jurkat cells. Degradation of overexpressed proteoforms was monitored by an anti-V5 antibody and compared to the turnover of stable proteins such as actin and tubulin.
Assessment of N-terminal proteoform stability

Discussion

The selection of translation initiation sites attributes to the complexity of proteomes as N-terminally truncated and extended proteoforms may be differentially regulated, localized [51, 52], display altered stabilities [12] and functionalities [53-55] especially in the case of some large Nt-truncations. Proteoforms arisen from alternative translation initiation are generally well conserved among eukaryotic orthologous protein species [9]. Selection between alternative translation start sites is regulated under conditions of cellular stress [36] and may change the repertoire of N-terminal proteoforms expressed in a developmental and tissue-specific manner [53, 55].

pSILAC was already used in several proteome studies to determine the dynamic properties of thousands of proteins in a single experiment. Most studies made use of non-synchronised, proliferating cell lines [2, 6, 7, 56], while others employed growth arrest [1] or studied protein synthesis and degradation rates in dynamically changing systems (for example during differentiation [5] or following drug-mediated inhibition of Hsp90 [40]). We decided to employ a pSILAC approach similar to the one described by [2] which enabled the measurement of protein half-lives ($t_{1/2}$) as well as protein turnover, reflecting the overall protein synthesis and degradation rates [2]. To better reflect our experimental results, we simplified the mathematical model proposed by Boisvert and colleagues [2]. In their setup, levels of medium Arg were retained at 20% instead of asymptotically approaching zero as degradation of proteins proceeds, an observation which they attributed to amino acid recycling. In Jurkat cells we observed a complete isotope replacement, similar to another recent pSILAC study conducted in the same cell line [40].

Here we performed a first global, N-terminal proteoform-specific analysis of protein stability and identified numerous N-terminally truncated proteoforms with distinct turnover rates compared to their full-length counterparts. Among the compared dbTIS/aTIS pairs, the stability of alternative proteoforms was not correlated to the length of truncation. The regulation of protein turnover rather seemed proteoform-specific, pointing to many different factors involved in the synthesis and degradation of specific proteins. One such contributing factor may be post-translational modifications, as we found 11 proteoform pairs where ubiquitination of the longer variant possibly contributed to reduced proteoform stability.

To validate the difference in half-lives of selected proteoform pairs we turned to the use of a more classic method for measuring protein turnover, being CHX pulse-chase experiments followed by western blot analysis. We confirmed turnover rates of several endogenously expressed short- and long-lived proteins. Due to the limitation in molecular weight resolution and sensitivity we subsequently turned to over-expression studies for the validation of differential Nt-proteoform stabilities from two human genes. Nonetheless, since upon incubation of the cells with protein synthesis inhibitors, protein degradation may
be altered [57, 58] and cell toxicity can only be controlled for a relatively short period, CHX treatment may preclude a correct delineation/validation of long protein half-lives.

We further evaluated the origin of N-terminal proteoforms identified in our study using ribosome profiling and extensive mining of public databases (Appendix Table II.1C) and found that 74% of the aTIS were generated by alternative translation initiation, whereas 23% likely pointed to splicing events. Truncated proteoforms displayed a lower translation initiation signal at the start codon, in line with their reduced proteome abundance estimated by spectral counting. These finding are consistent with a lower frequency of downstream translation initiation assumed from the leaky scanning model of translation initiation and corroborated by previous reports (Lee et al, 2012). Next to their lower abundance, the limited overlap between aTIS identified by means of proteomics and ribo-seq might be additionally attributed to the TIS calling approach used, since typically a higher interference is observed at internal TIS typically flanked by signal at both 5’ and 3’ end. Moreover, some aTIS may actually be the sole or preferential TIS site selected due to alternative splicing and, in some cases, alternative translation initiation. In this context, previous studies demonstrated that their corresponding dbTIS sites are typically embedded in less optimal Kozak consensus sequences [9]. Indeed, we found 12 aTIS proteoforms exclusively expressed in the absence of dbTIS both at the translome and the proteome level. In some cases a truncated proteoform may also display superior stability over the canonical proteoform. Viewing their general lower abundance, stable aTIS proteoforms are more likely to be detected. The latter could explain why the aTIS population was slightly enriched for long-lived proteins.

As reported previously [1, 5], turnover rates were similar for proteins belonging to similar functional groups or macromolecular complexes. This is supported by the notion that some proteins get stabilised upon complexation with other proteins, lipids, sugars or nucleotides [1]. In line, other studies suggested that protein subunits turn over quickly where the complex is assembled, but are stabilised in cellular compartments where the complex is functional [2]. Some of the identified protein truncations are predicted to be benign for a protein’s function or binding but have large consequences for protein stability. For example, a single amino acid truncation of AIMP2 increases protein stability (see Fig 6A-B) but was not predicted to alter its sequence disorder content, localisation (TargetP [59], WoLF PSORT [60]) or any of its known interacting domains. Considering such cases, a differential regulation of N-terminal proteoform stability may potentially result in elimination or introduction of unstable components into a complex. It is also important to consider that the relative expression levels of proteoforms are responsive to cellular conditions [53, 61, 62]. In this context, depletion of one proteoform may increase the impact of another proteoform’s turnover on protein complex formation. As we observed that aTIS proteoforms were enriched in proteins forming homomultimers, generation of alternative proteoforms may be important in the context of homomultimerisation. Interestingly, in the case of the transcription factor GATA-1, proteoforms generated by alternative translation initiation
were already shown to produce dimers [55] with altered activities dependent on the relative ratios of the two proteoforms in the complex.

The extent of intrinsic disorder in protein structures was correlated with turnover times, indicating that a change in structural disorder of protein N-termini from truncated proteoforms might have an impact on the proteoforms’ cellular stabilities. In line with the observation that proteins displaying iMet cleavage are overrepresented in vivo, the general belief is that retained iMet reduces the stability of proteins [15, 63]. Unexpectedly, our stability data showed a reduced stability of iMet processed proteoforms in numerous cases. However, it is important to discriminate between the effect of iMet removal on protein abundance and the effect on protein turnover time, as these two terms should not be used interchangeably. Relying on previous quantitative data on protein Nt-acetylation [9], another common co-translational modification, Nt-acetylation seemed to have no overall impact on turnover rates. Of note however, due to the experimental methodology applied, we cannot distinguish Nt-acetylated from Nt-free proteoform counterparts. In fact, the role of Nt-acetylation in protein stability is debated [25, 26] and is likely protein context and NAT dependent [30, 31]. It was postulated that Nt-acetylation may stabilize otherwise flexible structures of N-termini and reduce their propensity for aggregation [24]. At the same time, acetylated protein N-termini are substrates of the N-end rule pathway [29] which may constitute a clearance mechanism for N-terminally misfolded or non-complexed proteins [21, 32, 33, 64]. Moreover, it is difficult to directly compare our results to the destabilizing properties of the corresponding N-terminal amino acids originally described by Varshavsky and colleagues, especially considering differences in the studied organisms and the techniques used for protein turnover determination (pSILAC vs. CHX-chase). We believe that the property of iMet retention to increase the stability of N-terminal proteoforms described in our study may exist as an additional mechanism next to the N-end rule pathways to ensure the fine tuning of protein stability in eukaryotic cells.

Recent large-scale studies on protein stability report a limited correlation to other high-throughput data and especially, differences in the methodologies applied appear to account for the observed discrepancies [1, 2, 6, 65, 66]. In this context, pSILAC approaches share general advantages, as they target endogenous proteins expressed at physiological levels in untreated cells. We therefore sought to cross-examine our present turnover data with previous studies performed using pSILAC in human (HeLa and Jurkat cells [2, 40]) and mouse (NIH3T3 cells [7]). We observed a positive correlation of our data with these other protein stability studies, with Spearman’s correlation coefficients of 0.48, 0.44 and 0.49 for [7, 40] and [2], respectively. Moreover, comparison of gene ontology enrichment analyses indicated that protein turnover rates seem to be conserved between different cell lines and organisms. Despite similar experimental procedures, there was still a lot of unexplained variation between the datasets, which may be attributed to differential regulation of protein stability in different cell lines or differential growth conditions in the case of the Jurkat cell line, while data processing and interpretation are other factors to consider. Importantly, we
clearly demonstrate the contribution of proteoform-specific turnover rates and that differences in stability between N-terminal protein variants may in principle only be resolved by focusing on their unique peptides, whereas shared peptides are a potential source of error in turnover measurements.

Protein turnover measurements were recently extended to protein isoforms and (modified) protein pools [2, 67]. The Lamond group provided evidence for differential stabilities and localisation of proteoforms. Their shotgun approach, however well suited to study splice variants, had limited application for N-terminal proteoforms. Our pSILAC/N-terminal COFRADIC approach can be further extended to study changes in stability and proteoform expression levels during dynamic processes or in different tissues [5, 40]. Since targeting to diverse micro-environments allows fine-tuning of local protein concentrations, adding a spatial context to protein turnover may increase our understanding of the functional diversity of proteomes in future studies.

**Materials and Methods**

**Cell culture**

For N-terminal COFRADIC analyses, human Jurkat T-lymphocytes (clone E6-1 ATCC TIB-152, ATCC), were subjected to SILAC labelling [68]. Cells were cultured in a 5% CO₂ gas-equilibrated, humidified incubator in Roswell Park Memorial Institute (RPMI) - 1640 medium without arginine and lysine (Silantes, GmbH, Munich, Germany) containing 10% fetal bovine serum dialyzed using a 10 kDa cut-off membrane (Gibco, Life Technologies, Paisley, Scotland, UK), 2 mM alanyl-L-glutamine dipeptide (GlutaMAX, Gibco), 50 units/ml of penicillin (Gibco), 50 µg/ml of streptomycin (Gibco) and 146 µg/ml light (¹²C₆) L-lysine (Cambridge Isotope Labs, Andover, MA, USA). Media were supplemented with either light (Arg⁰, ¹²C₆¹⁴N₄) or medium (Arg⁶, ¹³C₆¹⁴N₄) L-arginine (Cambridge Isotope Labs) at a final concentration of 48 µg/ml at which Arg to Pro conversion was not detected. To achieve a complete incorporation of the labelled amino acids, cells were maintained in culture for 7 population doublings (Appendix Fig. II.5).

The pulsed SILAC strategy for protein turnover measurements was previously described by Boisvert et al. [2]. Briefly, suspension cultures containing equal numbers of unlabelled (Arg⁰) and labelled (Arg⁶) cells were taken and the medium was removed by centrifugation for 5 min at 350 g. Cell pellets were washed with PBS and centrifuged for 5 min at 350 g. The unlabelled cells were then suspended in light RPMI medium, whereas the medium labelled (Arg⁶) cells were suspended in medium containing heavy Arg (Arg¹⁰, ¹³C₆¹⁵N₄) L-arginine. Following medium change, separate starting cultures of 10⁷ cells were cultured at a concentration of 2*10⁵ cells/ml for periods of 0.5, 1.5, 4, 8, 12, 24, or 48 h, after which cultures of 10⁷ cells were collected by centrifugation. Following two rounds of re-suspending in PBS and centrifugation at room temperature, cell pellets were immediately frozen on dry ice and stored at -80°C.
Proteins were isolated from cell pellets by 3 rounds of freezing and thawing in 1 ml of 50 mM sodium phosphate buffer, pH 7.5 and cell debris was removed by a 10 min centrifugation at 13,200 g at 4°C. The protein concentration in the supernatants was measured by the Bradford method and for every time point sample, an equal amount of medium/heavy labelled proteome sample was mixed with the corresponding equivalent of light control sample.

Cycloheximide pulse-chase experiments

100 µg/ml cycloheximide was added to Jurkat cells that were cultured at a density of 180,000 cells/ml. 3.46 x 10^6 cells were harvested after 0, 0.5, 1.5, 4, 8, 12 or 24 hours of treatment and pelleted for 3 minutes at 800g. Cells were lysed in 100 µl of RIPA buffer (50 mM Tris-HCl pH 8.0, 150 mM NaCl, 1% NP-40) with protease inhibitors added (Roche). All lysates were flash-frozen and stored at -80°C until further processing. Samples were thawed, centrifuged for 10 min at 13,200g, followed by measurement of protein content using the Bradford method. 40 µg of protein material from each sample was separated by SDS-PAGE (4-12% Criterion Bis-Tris gel, Bio-Rad) and protein levels were monitored by western blotting. The following antibodies were used: anti-lamin B (NA12, Calbiochem), anti-securin (ab3305, Abcam), anti-β-Catenin (C2206, Sigma), anti-SYF2 (ab3610, Abcam), anti-GAPDH (ab9484, Abcam), anti-actin (A2066, Sigma) and anti-α-tubulin (T5168, Sigma). The intensity of bands was assessed using the LICOR Odyssey software for western blot image processing.

To validate the proteomics-derived differential protein turnover of dbTIS and aTIS proteoforms of MARE2 and AN32E, complete cDNA clones were obtained from the ORFeome 7.1 collection (ID 4499) and I.M.A.G.E. (ID 3047915) respectively. Site-directed mutagenesis (Stratagene, Agilent Technologies) by mutating ATG (Met) to TTG (Leu) codons was performed to analyse the expression of selected proteoforms. (Mutagenized) constructs were subcloned into the eukaryotic pEF-DEST51 expression vector (Gateway, Life Technologies) and transfected in HCT116 cells using Fugene HD (Promega) at a 0.8/3/150,000 ratio of DNA [µg], Fugene HD [µl] and number of cells. Cycloheximide pulse-chase experiments were performed as described above for Jurkat cells, except that 150,000 HCT116 cells were harvested for every overexpressed construct and time point analysed. Cells were lysed in 80 µl RIPA buffer and about 10 µg of total protein material was separated by SDS-PAGE. Degradation of overexpressed proteoforms was monitored by western blotting using the anti-V5 antibody (R96025, Life Technologies) and compared to the turnover of stable proteins such as actin [7].

N-terminal COFRADIC

The combined proteome mixtures were analysed by N-terminal COFRADIC [18, 69]. In summary, proteins are first blocked at their N-ends (α-amines) and lysines (ε-amines) by acetylation. After trypsin digestion, terminal peptides are enriched for by means of SCX at low pH and by applying two consecutive RP-HPLC separations. In the latter, any remaining
internal peptides and C-terminal peptides are shifted by reaction of their free \( \alpha \)-amino-groups (new free \( \alpha \)-amino-groups introduced by the action of trypsin) with 2,4,6-trinitrobenzenesulphonic acid (TNBS) thereby allowing for their segregation from N-terminal peptides. Fractions enriched for N-terminal peptides are subsequently analysed by means of LC-MS/MS. All steps of the N-terminal COFRADIC analysis, including SCX enrichment, were performed as described [69]. More specifically, solid guanidinium hydrochloride (Gu.HCl) was added to the proteome mixtures to a final concentration of 4 M in order to denature all proteins. Next, proteins were reduced and alkylated using TCEP and iodoacetamide respectively. Upon desalting of the S-alkylated samples in 50 mM sodium phosphate (pH 8.0), primary free amines were N-acetylated (no distinction between can be made between \textit{in vivo} Nt-free (and thus \textit{in vitro} Nt-acetylated) and \textit{in vivo} Nt-acetylated N-termini by adding sulfo-N-hydroxysuccinimide (NHS)-acetate (Pierce). Twice the molar excess of glycine over the NHS ester was subsequently added to quench any non-reacted NHS-acetate. Possible O-acetylation of Ser, Thr or Tyr residues was reverted by adding hydroxylamine (Fluka) to the modified protein mixtures. A final desalting step was performed in protein digestion buffer (10 mM ammonium bicarbonate, pH 7.9) and the proteomes were digested overnight at 37 °C with sequencing-grade, modified trypsin (Promega, Madison, WI, USA) (enzyme/substrate of 1/100, w/w). Subsequent steps of the N-terminal COFRADIC analysis, including SCX enrichment, were performed as described [48].

Following label-swapping, one N-terminal COFRADIC analysis was performed and per time-point (7 N-terminal COFRADIC analysis in total). During the primary RP-HPLC separation of the actual N-terminal COFRADIC procedure, peptides were collected in 13 fractions of 4 min each. Next, 12 secondary fractions were collected per primary fraction and pooled as described in [69], overall resulting in 36 samples for LC-MS/MS analysis per N-terminal COFRADIC setup.

**LC-MS/MS analysis**

LC-MS/MS analysis was performed using an Ultimate 3000 RSLC nano HPLC (Dionex, Amsterdam, the Netherlands) in-line connected to an LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). The sample mixture was loaded on a trapping column (made in-house, 100 µm I.D. × 20 mm, 5 µm beads C18 Reprosil-HD, Dr. Maisch). After back flushing from the trapping column, the sample was loaded on a reverse-phase column (made in-house, 75 µm I.D. × 150 mm, 5 µm beads C18 Reprosil-HD, Dr. Maisch). Peptides were loaded in solvent A’ (0.1% trifluoroacetic acid, 2% acetonitrile (ACN)) and separated with a linear gradient from 2% solvent A’’ (0.1% formic acid) to 50% solvent B’ (0.1% formic acid and 80% ACN) at a flow rate of 300 nl/min followed by a wash reaching 100% solvent B’. The mass spectrometer was operated in data-dependent mode, automatically switching between MS and MS/MS acquisition for the ten most abundant peaks in a given MS spectrum. Full scan MS spectra were acquired in the Orbitrap at a target value of 1E6 with a resolution of 60,000. The ten most intense ions were then isolated for
Assessment of N-terminal proteoform stability

fragmentation in the linear ion trap, with a dynamic exclusion of 20 s. Peptides were fragmented after filling the ion trap at a target value of 1E4 ion counts. Mascot Generic Files were created from the MS/MS data in each LC run using the Mascot Distiller software (version 2.4.2.0, Matrix Science, www.matrixscience.com/Distiller.html). To generate these MS/MS peak lists, grouping of spectra was allowed with a maximum intermediate retention time of 30 s and a maximum intermediate scan count of 5. Grouping was done with a 0.005 Da precursor tolerance. A peak list was only generated when the MS/MS spectrum contained more than 10 peaks. There was no de-isotoping and the relative signal-to-noise limit was set at 2.

The generated MS/MS peak lists were searched with Mascot using the Mascot Daemon interface (version 2.3.01, Matrix Science). Searches were performed in the Swiss-Prot database with taxonomy set to human (UniProtKB/Swiss-Prot database version 2011_08, containing 20,244 human protein entries). The Mascot search parameters were set as follows: acetylation at lysine side-chains, carbamidomethylation of cysteine and methionine oxidation to methionine-sulfoxide were set as fixed modifications. Variable modifications were acetylation of N-termini and pyroglutamate formation of N-terminal glutamine (both at peptide level). Endoproteinase semi-Arg-C/P (semi Arg-C specificity with Arg-Pro cleavage allowed) was set as enzyme allowing for no missed cleavages. Mass tolerance was set to 10 ppm on the precursor ion and to 0.5 Da on fragment ions. Peptide charge was set to 1+, 2+, 3+ and instrument setting was put to ESI-TRAP. Only peptides that scored above the threshold score, set at 99% confidence, were withheld. The FDR was estimated by searching a decoy database (a shuffled version of the human Swiss-Prot database version 2011_08) and ranged from 0.32% to 2.27% for individual samples, with a global FDR of 1.18% on the spectrum level.

Ribosome profiling

For ribosome profiling, Jurkat cells were grown in RPMI medium (Gibco) supplemented with 10% fetal bovine serum, 2 mM alanyl-L-glutamine dipeptide (GlutaMAX, Gibco), 50 units/ml penicillin and 50 µg/ml streptomycin at 37 °C and 5% CO₂. Cultures were treated with 50 µM lactimidomycine (LTM) [70, 71] or 100 µg/ml cycloheximide (CHX) (Sigma, USA) for 30 min at 37 °C before cell harvest. Subsequently, cells were collected by centrifugation (5 min at 300g), rinsed with ice-cold PBS and recovered again by centrifugation in the presence of CHX. 10⁸ cells were re-suspended in 1 ml ice-cold lysis buffer (10 mM Tris-HCl, pH 7.4, 5 mM MgCl₂, 100 mM KCl, 1% Triton X-100, 2 mM dithiothreitol (DTT), 100 µg/ml CHX, 1 × complete and EDTA-free protease inhibitor cocktail (Roche)) [72]. Following 10 min of incubation on ice, samples were passed through QIAshredder spin columns (Qiagen). The flow-through was clarified by centrifugation for 10 min at 16,000 × g and 4 °C. 600 µl of the recovered supernatant was subjected to RNase I (Life Technologies) digestion using 1000 U of enzyme. Digestion of polysomes proceeded for 55 min at room temperature and was stopped with SUPERase.In RNase Inhibitor (Life Technologies). Next, monosomes were
recovered by ultracentrifugation at 75,000 RPM in a cooled TLA-120.2 rotor over a 1 M sucrose cushion in 20 mM HEPES-KOH pH 7.4, 5 mM MgCl₂, 100 mM KCl, 2 mM DTT, 100 µg/ml CHX and 100 U/ml Superase.In. Subsequent steps were performed as described by Ingolia et al. [73] with some minor adjustments. RNA was extracted from the samples using a heated acid phenol; chloroform; isooamyl alcohol (125:24:1) procedure. 17 µg of RNA was subjected to electrophoresis under denaturing conditions in 15% TBE-Urea polyacrylamide gel (Life Technologies). Ribosome protected fragments (RPFs) of 28-34 nucleotides were extracted from the gel in RNase-free water for 10 min at 70 °C and precipitated with 1/10 volume of 3 M sodium acetate pH 5.2, 1.5 µl GlycoBlue (Life Technologies) and 1 volume of isopropanol overnight at −80 °C. Following a dephosphorylation reaction (end-repair) with 10 U of T4 polynucleotide kinase (New England Biolabs), the RPFs were depleted of ribosomal RNA contaminants using the Ribo-Zero Magnetic Gold Kit (Illumina) according to the manufacturer’s protocol. Subsequently, fragments were ligated to a 1.5 µg of Universal cloning linker using 200 U of T4 RNA ligase II (both New England Biolabs). The ligated product was size-selected using denaturing gel electrophoresis and purified from gel as described above. Subsequently, reverse transcription (RT) was performed with 200 U of SuperScript III (Life Technologies) according to [73], followed by a size-selection of the RT product (described above). Next, samples were subjected to circular ligation using 100 U of CircLigase (Illumina) and amplified by PCR using Phusion polymerase (New England Biolabs) for 12 cycles of denaturation (10 s at 98 °C), annealing (10 s of 65 °C) and elongation (5 s at 72 °C) using primer pairs compatible with the Illumina sequencing platform [73]. The resulting ribosome-profiling libraries of LTM- and CHX- treated Jurkat cells were sequenced on a NextSeq 500 instrument (Illumina) to yield 75 bp single-end reads.

Measurement of proteoform turnover

Arg-ending N-terminal peptides were selected that started at position 1 or 2 (entries stored in the Swiss-Prot database) or derived from alternative translation initiation at positions >2) and which complied with rules of iMet processing and Nt-acetylation [9]. For every unique N-terminus (unique proteoform), SILAC ratios (M/L, H/L and H/M) calculated for the highest scoring peptide identification were included in the analysis. M/L and H/L ratios were normalized, according to equation 1:

\[
\text{if } M + H = L \text{ then } \frac{M}{L} + \frac{H}{L} = 1
\]

where L, M and H denote light, medium and heavy isotope-labelled peptide intensities.

M/L and H/L ratios were plotted in function of time as degradation and synthesis curves, respectively. The crossing over point between the two curves corresponds to 50% protein turnover, see Equation 2:

\[
\text{if } \frac{M}{L} + \frac{H}{L} = 1 \text{ (Eq. 1)} \text{ and } \frac{M}{L} = \frac{H}{L} \text{ then } \frac{M}{L} = \frac{H}{L} = \frac{1}{2}
\]
Assessment of N-terminal proteoform stability

50% protein turnover was calculated as the crossing point between protein degradation and synthesis curves for proteoforms quantified in at least 3 time points. We assumed that protein degradation follows a first-order process. Therefore, an exponential curve with one parameter (B) was fitted to M/L data using the R software (version 0.98.739) and the turnover time was calculated as follows:

Equation 3: \( \frac{M}{L} = e^{(-k_{\text{deg}} \frac{\ln(2)}{t_{cc}})t} \), simplified to \( \frac{M}{L} = e^{Bt} \), hence \( t_{\text{turn}} = \frac{\ln(\frac{1}{2})}{B} \)

where \( k_{\text{deg}} \) is the protein degradation constant; \( t_{cc} \), cell doubling time and \( t_{\text{turn}} \), turnover time.

Alternatively, the turnover time was approximated from linear equations between two M/L and two H/L points lying on opposite sides of the crossing point. We observed a very high correlation between the exponential and the linear method (Pearson coefficient of 0.98). Exponential curve fitting was used for protein turnover determination if the quality of fit determined by the coefficient of determination (R\(^2\)) was at least 0.8. In other cases, turnover time was calculated from the linear approximation, provided a descending trend of M/L changes in time was observed. Turnover measurements fulfilling criteria described above were regarded as valid, while other turnover rates remained unassigned and were excluded from analysis. Protein degradation constants were derived from the parameter B (Equation 3) of individual M/L profiles and protein half-lives \( (t_{1/2}) \) were calculated accordingly to

Equation 4: \( t_{1/2} = \frac{\ln(\frac{1}{2})}{-k_{\text{deg}}} = \frac{\ln(\frac{1}{2})}{(B + \frac{\ln(2)}{t_{cc}})} \)

Due to the constrains of Equations 3 and 4, negative values of \( k_{\text{deg}} \) were disregarded and protein half-lives estimated to be infinite due to continuous cell divisions: \( t_{1/2} \gg t_{cc} \rightarrow t_{1/2} \approx \infty \).

Quantification of proteoform abundance

Per unique proteoform, spectral counts were averaged among all time points whenever identified (using an arithmetical mean). Normalized spectral abundance factor (NSAF) values [44] were calculated based on amino acid lengths of individual proteoforms. All proteoforms identified in at least one time point were taken as the complete dataset (Appendix Table II.1A).

Analysis of ribosome profiling data

Ribosome profiling (ribo-seq) performed in Jurkat cells treated with CHX to study translation elongation and LTM to study translation initiation were analysed using the PROTEOFOMER pipeline [74]. PROTEOFOMER was used to perform the quality control, mapping onto a reference genome, identification of TIS and generation of a protein database.
The 3’ adapter sequences were clipped using the fastx_clipper and reads shorter than 26 or longer than 34 nucleotides were discarded for the ribo-seq based TIS calling. The reads were first mapped onto small nuclear RNA, tRNA and rRNA sequences to eliminate contaminating reads. The remaining reads were subsequently mapped onto the human GRCh37 reference genome (Ensembl annotation bundle 75) allowing for a maximum of 2 mismatches along the entire read. Reads mapping to a maximum of 16 locations on the genome were allowed and retained for every loci. All mapping steps were performed using STAR 2.4.0i. RPF alignments were reduced to a specific position (the P-site nucleotide) using an offset from the 5’ end of the alignment based on the read length (+12, +13 and +14 respectively for alignment of ≤ 30 bases long, 31–33 bases long, and ≥ 34 bases long). Translation start sites were identified by setting different parameter values for the different TIS categories in the tool. A TIS was called at the Ensembl annotated start site if i.) it had a minimum read count of 5; ii.) the position was a local maximum within a 7 nucleotides window (one codon upstream and one codon downstream); iii.) the difference between the LTM (foreground) and CHX (background) normalized ratio (RLTM-CHX) within the window reached at least 0.01. A TIS was called within the coding region of a transcript (downstream of annotated start sites) if it had a minimum read count of 15 and RLTM-CHX of 0.15. For TIS located at the 5’UTR, 3’UTR and other non-protein-coding transcripts, the minimum read count was set to 10 with a RLTM-CHX of 0.05.

We used the above settings to define translation initiation events in Jurkat cells. Taking into consideration every annotated transcript splice variant, all possible ORFs starting at the detected TIS sites were in silico translated. The N-terminal peptides identified in our pSILAC experiment (N= 2,578) were mapped onto the ribo-seq derived protein sequences allowing to associate peptides with ORFs and their corresponding transcripts.

Analysis of translation initiation and splicing

To study the contribution of splicing and (alternative) translation initiation to N-terminal proteoform expression, we mapped the N-terminal peptides identified in our proteomics experiment onto the human Swiss-Prot Isoform, UniProt TrEMBL, Ensembl human protein sequences next the ribo-seq data from Jurkat cells. Further, a comparison was made with ribo-seq data obtained from another human cell line (HCT116 colorectal cancer cells, unpublished data).

In case of ribo-seq matching evidence, a peptide corresponding to position 1 was assumed to be generated via translation initiation. A peptide mapping at position 2 was considered a TIS only if the first methionine was preceded by an alanine, cysteine, glycine, proline, threonine or valine and therefore likely co-translationally processed. We used the Jurkat data as reference i) to retrieve RLTM-CHX values for called TIS sites; ii) to associate transcripts with a normalized RPF read count (the normalized read count was calculated by summing all the reads that map across the transcript and dividing by its length) and iii) to determine their exon coverage. Since an N-terminal peptide might map to multiple transcripts, additional
Assessment of N-terminal proteoform stability

information regarding the expression was provided for every matching transcript to allow for a discrimination of transcripts poorly supported by the ribo-seq data.

We further investigated whether a downstream peptide is pointing to a possible annotated splicing or alternative translation event by checking if this peptide matched the N-terminal protein sequence previously reported in the protein databases; Ensembl, Swiss-Prot Isoform or UniProt TrEMBL protein sequences. A peptide was considered as a possible N-terminus of a splice variant if it mapped at position 1 or 2 to a given sequence in Ensembl or Swiss-Prot Isoform databases. Further, we qualified the Ensembl derived evidence of splicing as either poor (less likely) or good (likely) splicing events. A downstream peptide was considered to be supported by a likely splicing event if any of the Ensembl protein sequences associated to the peptide had a transcript support level less than or equal to three (with one being the highest rank). If none of the protein sequences associated to the peptide had a transcript support level less than three, it was considered as a poorly supported splicing event or not applicable if there was no transcript support level assigned to any of the valid transcripts. For N-terminal peptides matching an N-terminus of a Swiss-Prot Isoform sequence, we further retrieved the information if this protein isoform is generated via splicing or alternative translation initiation. The above mentioned data is presented in Appendix Table II.1C.

Analysis of proteoform turnover rates in macromolecular complexes

To investigate the dynamic properties of macromolecular complex substituents, protein turnover times were mapped to members of human protein complexes defined by the CORUM database (core set of 1,343 human complexes [75]). Only Swiss-Prot annotated proteoforms (dbTIS) with the most plausible iMet processing status ([18]) and high quality turnover measurement were considered [16, 18]. Absolute differences in turnover were calculated pairwise between all members of a complex. Subsequently, one-sided Mann-Whitney test at a 95% confidence interval (R, wilcoxon.test, stats package) was applied to detect proteins with significantly higher absolute differences than the remaining members of a complex.

Gene ontology enrichment analysis

Swiss-Prot accessions of the 1,894 proteoforms with high quality measurement of stability of one representative N-terminal proteoform were ranked based on their corresponding turnover values. Subsequently, an enrichment analysis of GO terms was performed using GOrilla (http://cbl-gorilla.cs.technion.ac.il/, [41]) using the “single ranked list of genes” option. Duplicate accessions were removed keeping the highest ranking instance per accession. The enrichment p-values were computed using mHG (minimal Hyper Geometric) statistics and corrected for multiple testing using the Benjamini and Hochberg method giving the FDR q-values. GO terms enriched in stable or unstable proteins with FDR q-values ≤ 0.05 were subsequently summarized using REViGO [42]. Redundant GO terms were removed and the remaining terms visualized in bar charts according to their significance (-log10(q-value)).
Analogously, we performed an enrichment analysis of GO terms in aTIS vs. dbTIS proteoforms using the GOrilla “target and background” option.

**Prediction of protein intrinsic disorder**

Canonical protein sequences identified in at least 3 time points (1,972 proteoforms) were retrieved and adjusted to start at the initiating residue of the N-terminal peptide identified. Of these proteoforms, the percentage of intrinsically disorder was predicted using RAPID (http://biomine-ws.ece.ualberta.ca/RAPID/index.php, [47]) which implements the IUPred algorithms [46, 76].

**Hydropathy of protein N-termini**

The grand average of hydropathy (GRAVY) was calculated for the 2 and 10 amino acid long N-terminal sequence of every proteoform (1,894 proteoforms) at http://www.gravity-calculator.de/. Increasing positive score indicated greater hydrophobicity. The calculation was based on the sum of hydropathy values of amino acids in Kyte-Doolittle scale [77], divided by the number of residues. Correlation between GRAVY values and turnover times was estimated using Pearson and Spearman correlation coefficients.

**Statistical analysis of turnover, abundance and disorder of aTIS versus dbTIS**

Differences in turnover, abundance and disorder values between aTIS and dbTIS groups were measured for paired dbTIS-aTIS data using a Wilcoxon signed-rank one-sided test for paired samples (R, stats package) and independently between the whole aTIS and dbTIS groups using a Kruskal-Wallis test for difference in distribution (R, stats package), multiple comparison test after Kruskal-Wallis (R, pgirmess package) and Mann-Whitney one-sided test for independent samples. All analyses were performed in R at p ≤ 0.05 significance.

**Category enrichment analysis in turnover, disorder and abundance data**

1D enrichment analysis of Pfam, KEGG, GO slim and other categories was performed using the Perseus software [78] at p ≤ 0.01 significance with a two-sided test with Benjamini-Hochberg FDR correction. Turnover, disorder and abundance values corresponding to each categorical term were tested for deviations from the complete distribution and a preference for larger or smaller values was assessed independently for every variable. Correlation of turnover, disorder and abundance values was assessed using Spearman rank correlation.

**Relation between protein turnover and N-terminal amino acid composition**

All proteoforms with measured turnover rates (including proteoforms with alternative iMet processing; 1,941 in total) were categorised according to the identity of their utmost N-terminal residues. Additionally, N-termini susceptible to iMet processing were further categorised as either Met retaining or Met cleaved. The general impact of N-terminal residues and iMet processing (categorical variables) on protein turnover was assessed by a
Assessment of N-terminal proteoform stability

Kruskal-Wallis rank sum test (R, stats package). A subsequent multiple comparison test for Kruskal-Wallis (R, pgirmess package) and pairwise Wilcoxon rank sum test (without correction for multiple testing; R, stats package) allowed to indicate the most deviating subgroups (with \( p<0.05 \)).

Data availability

MS/MS data was converted using PRIDE Converter [79] and is available through the PRIDE database [80] with the dataset identifier PXD002091 and DOI 10.6019/PXD002091.

Ribo-seq sequencing data has been deposited in NCBI’s Gene Expression Omnibus [81] and is accessible through GEO Series accession number GSE74279 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE74279).

Author Contributions

D.G. performed experiments, analyzed data, drafted and revised the manuscript. E.N. analyzed data. K.G. drafted and revised the manuscript. P.V.D. conceived the study, performed the proteomics experiments, analyzed data, drafted and revised the manuscript.
References


Assessment of N-terminal proteoform stability


RESULTS  Chapter III

Deep integration of OMICS data
Adapted from:


(* shared first authorship)
RESULTS  Chapter III

Deep integration of OMICS data

Abstract

Next-generation transcriptome sequencing is increasingly integrated with mass spectrometry to enhance MS-based protein and peptide identification. Recently, a breakthrough in transcriptome analysis was achieved with the development of ribosome profiling (ribo-seq). This technology is based on the deep sequencing of ribosome-protected mRNA fragments, thereby enabling the direct observation of in vivo protein synthesis at the transcript level. In order to explore the impact of a ribo-seq-derived protein sequence search space on MS/MS spectrum identification, we performed a comprehensive proteome study on a human cancer cell line, using both shotgun and N-terminal proteomics, next to ribosome profiling, which was used to delineate (alternative) translational reading-frames. By including protein-level evidence of sample-specific genetic variation and alternative translation, this strategy improved the identification score of 69 proteins and identified 22 new proteins in the shotgun experiment. Furthermore, we discovered 18 new alternative translation start sites in the N-terminal proteomics data and observed a correlation between the quantitative measures of ribo-seq and shotgun proteomics with a Pearson correlation coefficient ranging from 0.483 to 0.664. Overall, this study demonstrated the benefits of ribosome profiling for MS-based protein and peptide identification and we believe this approach could develop into a common practice for next-generation proteomics.
Introduction

A shotgun proteomics experiment typically involves the fractionation of a complex peptide mixture followed by LC-MS/MS analysis and the identification of peptides using one of several protein or peptide sequence database search tools [1-3]. N-terminal proteomics techniques such as N-terminal COFRADIC (combined fractional diagonal chromatography) expand on the results of a typical shotgun experiment by enriching for N-terminal peptides, thus revealing (alternative) translation start sites, while simultaneously measuring co-translational modifications of protein N-termini [4]. Protein reference databases only contain experimentally verified and/or predicted sequences and are therefore unlikely to contain a comprehensive representation of the actual protein content of a given sample. To resolve this shortcoming, recent efforts have been directed towards the combination of proteomics and next-generation transcriptome sequencing [5-8]. Proteogenomic approaches that delineate translation products based on mRNA sequencing data may improve protein identification in multiple ways. The transcriptome of a sample offers a more representative expression profile than could be obtained with a public database alone while at the same time reducing the search space through the elimination of unexpressed gene products [9]. The transcript data also contains useful information about sequence variations such as single nucleotide polymorphisms (SNP) or mutations and RNA splice and editing variants [9-11], which increases the chances of detecting new proteins or protein forms [12-14]. Despite the benefits of adding next-generation transcriptome sequencing to an MS-based proteomics experiment, there are still several improvements possible. Because of extensive translation regulation, the presence of a transcript does not necessarily imply the presence of the corresponding protein [15-17]. On top of that, several factors, including internal ribosome entry sites, the presence of multiple ORFs per transcript, non-AUG start codons and leaky scanning, ribosome frame shifting and stop codon readthrough hamper the prediction of the exact protein sequence(s) from a single transcript sequence [18-20]. Recently, a novel technique has been described that attempts to tackle these limitations: ribosome profiling [21].

Ribosome profiling, or ribo-seq, is based on the deep sequencing of ribosome-associated mRNA fragments, thus enabling the study of in vivo protein synthesis at the transcript level. In a ribo-seq experiment, eukaryotic translation is often halted using cycloheximide (CHX). The mRNA that is not protected by ribosomes after the translation halt is digested with nucleases and the monosome-mRNA complexes are isolated. Next, the protected mRNA sequences are separated from the ribosomes and converted into a DNA library, ready to be sequenced. The sequencing results in a genome-wide snapshot of the mRNA that enters the translation machinery. Additionally, (alternative) translation initiation sites can be studied with sub-codon to single-nucleotide precision through the use of antibiotics such as harringtonine (HARR) or lactimidomycin (LTM), which cause the ribosomes to halt at sites of translation initiation [22, 23]. When the exact translation start site is known, the ORF can be delineated, thus eliminating the need to translate the transcripts in three or six reading
frames. The measurement of mRNA at the translation level, combined with the knowledge of the exact translation start sites, makes ribosome profiling an excellent choice for the creation of a custom protein sequence search space for MS/MS-based peptide identification [24]. It has to be noted that ribo-seq does not generate direct evidence of mature proteins or protein stability and that some non-coding transcripts do not result in a protein product, despite being associated with ribosomes [25-27]. However, MS-assisted validation may help to resolve both issues. Apart from canonical translation products, ribosome profiling also aids in the identification of unannotated truncated and N-terminally extended protein variants and the validation of these variants can come from matching N-terminal COFRADIC data [24, 28].

In this study we created a custom protein sequence database based on LTM ORF delineation for the HCT116 cell line, a widely used human colon cancer cell model, to serve as the search space for MS/MS spectra obtained by means of shotgun proteomics and N-terminal COFRADIC (Fig. 1). Translation products derived from the ribosome profiling data of the HCT116 cells were combined with the public Swiss-Prot protein sequence database [29] to

**Figure 1 - Proteogenomic strategy for the identification of proteins and peptides using a Swiss-Prot/ribo-seq-derived database.** Ribo-seq was performed twice on the human colon cancer cell line HCT116, once with CHX to halt translation globally and once with LTM to stop translation specifically at translation initiation sites. After translation initiation site (TIS) prediction, the ribo-seq-derived ORFs were translated to create a custom protein sequence database. This database was then combined with the human Swiss-Prot protein sequence database. Proteome samples were prepared from the same HCT116 cells and analyzed using both shotgun proteomics and N-terminal COFRADIC. The proteins and peptides in these samples were then identified using the custom combined protein search space.
build an optimal protein search space for our proteomics data. The addition of ribo-seq data resulted in the identification of 22 new proteins, i.e. proteins that were not contained in the Swiss-Prot database, out of a total of 2,816 protein identifications in our shotgun proteomics experiment. On top of that, the inclusion of ribo-seq data improved the score of 69 proteins as a result of the discovery of proteins with a SNP, new isoforms and homologs and extended protein forms. Out of a total of 1,262 peptides, ribo-seq identified 18 extra N-termini in the COFRADIC experiment compared to Swiss-Prot alone, including 6 N-termini originating from extended protein forms with a near-cognate start site (i.e. the protein does not start with the canonical AUG codon). It needs to be noted that in the shotgun proteomics experiment 312 proteins were uniquely identified using the Swiss-Prot database, emphasizing the importance of proteomics techniques for the validation of next-generation transcriptome sequencing datasets. Finally, the correlation between the ribo-seq and shotgun proteomics data was calculated. Depending on the settings used, the Pearson correlation coefficient between the ribo-seq-derived normalized ribosome-protected fragments (RPF) counts and the normalized spectral counts of the shotgun experiment (i.e. emPAI [30] and NSAF [31] values) ranged from 0.483 to 0.664.

Results

A regular shotgun and an N-terminal COFRADIC proteomics experiment were performed on a HCT116 cell line to determine the effect of the addition of ribo-seq-derived translation products to the Swiss-Prot protein sequence database on MS/MS spectrum identification. The shotgun data were used for the overall assessment of protein expression, whereas the N-terminal COFRADIC data were specifically used for the validation of the ribo-seq-predicted translation initiation sites.

Shotgun proteomics

Using the combination of Swiss-Prot and the ribo-seq-derived database, we identified a total of 2,816 proteins in the HCT116 cells (Fig. 2A). The majority of these proteins (2,482 or 88.1%) were identified in both Swiss-Prot and the custom database. The addition of the ribo-seq data to the protein search space led to 22 extra identifications, which would not have been picked up with just the Swiss-Prot database. Besides 9 previously unannotated protein products, these new identifications included 13 proteins with a SNP and three alternatively spliced isoforms. The inclusion of ribo-seq data also improved protein identification and score significance for 69 proteins since higher peptide coverage was obtained (Appendix Fig. III.1 shows three examples). The proteins with an improved score coincided with SNPs (52 proteins), alternatively spliced isoforms (14 proteins) and three N-terminal extensions. The ribo-seq experiment also missed 312 proteins, but these were still picked up thanks to the inclusion of Swiss-Prot in the search space. All the identified proteins and their respective annotations can be found in Appendix Table III.1. An
**A** shotgun proteomics

- Total number of protein identifications: 2,816
  - Shared ribo-seq/SwissProt identifications: 2,482
  - Unique SwissProt protein identifications: 312
  - Unique ribo-seq protein identifications: 22
- 22 unique ribo-seq protein identifications
- 13 SNP variant proteins
- 6 new protein products
- 3 new isoforms/homologs

**B** N-terminal COFRADIC

- Total number of peptide identifications: 1,289
  - dbTIS: 1,071
  - dtIS: 208
- 9 5’-extensions
- 1 uORF
- 9 5’-extensions
- 7 near-cognate start sites
- 2 cognate start sites
- 208 dtIS
- 204 cognate start sites
- 2 near-cognate start sites (with ribo-seq validation)
- 2 near-cognate start sites (without ribo-seq validation)

**Figure 2** - Bar charts showing the number of protein and peptide identifications obtained from the shotgun proteomics and N-terminal COFRADIC experiments. **A.** Shotgun proteomics. The custom combined protein sequence database resulted in the identification of 2,816 proteins. Most of these proteins (2,482 or 88.1%) were picked up by both databases independently, while 312 and 22 proteins were uniquely identified in the Swiss-Prot and ribo-seq databases respectively. The 22 unique ribo-seq identifications contained six new proteins, 13 proteins with a SNP and 3 unannotated isoforms. The ribo-seq data also improved the protein identification and score of 69 proteins. **B.** N-terminal COFRADIC. Most of the 1,289 peptides that were found in the custom combined protein sequence database...
mapped to canonical, annotated N-termini (1,071 dbTIS peptides or 83.1%). Of the remaining N-termini, 208 started downstream of the canonical start site (beyond protein position 2), 9 mapped to a 5′-extension and one to an uORF. For both the up- and downstream start sites, we identified several near-cognate start sites.

approximate analysis of the turnover rate and half-lives of the 312 missed proteins using publically available datasets [32, 33] showed no significant difference between the missed and the other identified proteins (Wilcoxon rank-sum test, p > 0.05). A gene ontology enrichment analysis using the DAVID tool [34] revealed that several biological process ontologies involving protein transport and localization were significantly enriched in the 312 missed proteins, just as the corresponding cellular localization ontologies linked to the cytoskeleton, cytosol and non-membrane-bounded organelles (Appendix Table III.2).

N-terminal COFRADIC

In order to validate the TiSs identified by the ribo-seq experiment and thus the corresponding N-terminal protein isoforms, positional proteomics in the form of N-terminal COFRADIC was applied to the HCT116 cells. After LC-MS/MS analysis and the subsequent combined database search, we identified 1,289 N-terminal peptides (Fig. 2B). The greater part of these peptides mapped to canonical start sites (1,071 peptides or 83.1%), 208 peptides started downstream of the canonical start site (past protein position 2 in reference to Swiss-Prot), 9 peptides mapped to a 5′-extension and one to an uORF. Two examples of proteins with an N-terminal extension or truncation are given in Fig. 3. Ribo-seq uniquely identified 18 peptides, which would have been missed when only searching Swiss-Prot. Both the N-terminal COFRADIC and ribo-seq experiment provided evidence of translation initiation at near-cognate start sites, which was also reported in previous COFRADIC and ribo-seq studies [22, 23]. A complete list of all identified N-terminal peptides is provided as Appendix Table III.1.

We compared the list of identified protein extensions starting at non-AUG start sites with the previously published list of non-AUG derived N-terminal extensions predicted by Ivanov et al. (2011) [35] and found matching evidence for one N-terminally extended protein (Swiss-Prot entry name HDGF_HUMAN; extension of 50 amino acids starting at GTG) out of 9 identified in our proteomics study.
Deep integration of OMICS data
Figure 3 - Depiction of two different N-termini that were predicted by ribo-seq and identified using N-terminal COFRADIC. The figure shows a 5'-extension (A. Swiss-Prot entry name RBP2_HUMAN) and an N-terminal truncation (B. Swiss-Prot entry name HNRPL_HUMAN). The UCSC genome browser [57] was used to create the plots of the ribo-seq and N-terminal COFRADIC data and the different browser tracks are from top to bottom: CHX treatment data, LTM treatment data, N-terminal COFRADIC data, UCSC genes, RefSeq.
genes and human mRNA from GenBank. The different start sites (a: alternative start site, b: canonical start site) are clearly visible in the zoomed genome browser views, just as the three-nucleotide periodicity of the ribo-seq data, especially in the N-terminal truncation image. The MS/MS spectra and sequence fragmentations indicate the confidence and quality of the peptide identifications.

Correlation analysis

We calculated a Pearson correlation coefficient to investigate the relation between the ribo-seq coverage and MS protein abundance measurements. Only transcripts for which quantitative information was available from both the ribo-seq and shotgun proteomics experiments were used in all the plots and calculations. The Pearson correlation values for the different normalization and identification approaches are listed in Table 1 and Fig. 4a shows the correlation plots for the NSAF values, which were better correlated with the ribo-seq coverage than the emPAI values. The highest correlation ($r^2 = 0.664$) was obtained when using only validated dbTIS transcripts with a total RPF count $\geq 200$. The correlation coefficients were also calculated for the 312 protein identifications that were present in Swiss-Prot, but not in our ribo-seq-derived search space (Appendix Fig. III.2). These 312 identifications were missing from the ribo-seq data because no TISs were identified in the LTM-treated cells, but, as there was coverage in the CHX-treated cells, the correlation could still be calculated. The Pearson correlation coefficients ranged from 0.464 to 0.713, depending on the protein selection and normalization procedure, and were similar for the proteins identified in both the Swiss-Prot and ribo-seq database.

<table>
<thead>
<tr>
<th>1% FDR</th>
<th>i</th>
<th>ii</th>
<th>iii</th>
<th>iv</th>
</tr>
</thead>
<tbody>
<tr>
<td>emPAI</td>
<td>0.488</td>
<td>0.498</td>
<td>0.483</td>
<td>0.518</td>
</tr>
<tr>
<td>NSAF</td>
<td>0.608</td>
<td>0.642</td>
<td>0.634</td>
<td>0.664</td>
</tr>
</tbody>
</table>

Table 1 - Pearson correlation coefficients between MS protein abundance and ribo-seq coverage. MS protein identifications were performed with an FDR of 1% and protein abundances were calculated as emPAI and NSAF values. The correlation coefficients were computed for each of the following transcript filtering settings: i) all dbTIS transcripts without additional thresholds, ii) only transcripts with a validated MS identification (i.e. transcripts with a spectral count value $> 2$), iii) only dbTIS transcripts with a total RPF count $\geq 200$ and iv) only dbTIS transcripts with both a validated MS/MS-based identification and an RPF count $\geq 200$.

We also investigated the link between the correlation and the degree of protein stability. Fig. 4b shows the correlation plot for validated dbTIS transcripts with an RPF $\geq 200$ together with the instability indexes of the proteins. These indexes were obtained with the ExPASy ProtParam tool [36], where a protein with an instability index $< 40$ is predicted to be stable and a protein with an index $\geq 40$ is considered unstable. The majority of unstable proteins
were characterized by lower NSAF and RPF values than the stable proteins. As reported previously, protein stability is among the most significant factors governing the correlation between gene expression and protein abundance [11].

Figure 4 - A. Correlation plots of protein abundance estimates based on NSAF values and RPF counts. Top left: all dbTIS transcripts; top right: dbTIS transcripts with a validated MS/MS-based identification (i.e. transcripts with a spectral count value > 2); bottom left: dbTIS transcripts with an RPF count ≥ 200; bottom right: dbTIS transcripts with both a validated MS identification and an RPF count ≥ 200. The regression line is shown in green. For each plot, the number of data points used (i.e. the number of dbTIS transcripts) as well as the corresponding Pearson correlation coefficient ($r^2$) is shown. B. Correlation plot with
the inclusion of stability data. Only dbTIS transcripts with both a validated MS/MS-based identification and an RPF count ≥ 200 were used (bottom right plot in Figure 4A). Instability indexes were determined with the ProtParam tool [36]: proteins with an instability index < 40 were classified as stable and are shown in blue, whereas proteins with an instability index ≥ 40 were classified as unstable and are shown in orange.

Discussion

The successful identification of proteins and peptides from MS/MS spectra depends on a number of factors. A state-of-the-art mass spectrometer that provides high resolution and mass accuracy is a vital element of a proteomics experiment. Solid experimental design and a robust identification pipeline are two other important factors. As even small changes in database search algorithms can lead to different identification results, combining several search engines, such as X!Tandem [2] and OMSSA[3], helps to increase the number of PSMs [37]. A more recent approach to improve the number of PSMs is based on the custom tailoring of the search space through the use of next-generation transcriptome sequencing [7, 24]. The new and improved protein identifications based on our ribo-seq-derived search space were a first indication of the success of our proteogenomics strategy. Especially the identification of N-terminally extended proteins would not have been possible when using only Swiss-Prot. The positive correlation between protein abundance (measured as NSAF and emPAI values) and the ribo-seq footprint coverage (measured as RPF counts) also justifies the usage of the described proteogenomics approach. It has been described before how NSAF gives a more accurate estimate of protein abundance than emPAI as it uses more information (e.g. fragment ion intensities and protein length) [38, 39]. This could explain why the NSAF values correlated better with the ribo-seq data. Interesting to note is that proteins with a lower stability index displayed both lower protein abundances as well as lower RPF counts than their more stable counterparts (Fig. 4B). Several studies have reported correlation values between mRNA-seq coverage and protein abundance, ranging from 0.41-0.44 [40] to 0.51 [11] in mouse and between 0.42 and 0.43 in rat [14]. Nagaraj et al. (2011) published a Spearman’s correlation of 0.6 between FPKM-based transcript abundance and iBAQ-based protein abundance values for the human HeLa cell line [5]. The improved correlation observed in our study can be explained by the fact that, because it measures transcripts after they have entered the translation machinery, ribosome profiling is less affected by translation regulation. The ability of ribo-seq to take alternative translation events into account leads to a better delineation of ORFs, which could also improve the correlation. Another advantage of the ribo-seq-derived database was that it allowed us to identify translation initiation from non-AUG start sites at the protein level, for which only limited evidence is available so far [28, 41-43].

Without the addition of the Swiss-Prot database to our custom search space, a significant amount of proteins would have been missed (unique Swiss-Prot identifications in Fig. 2). These proteins were missing from the ribo-seq-derived search space because no detectable
LTM-signal could be observed. But since the CHX treatment resulted in coverage for these proteins, we could still calculate the correlation between protein abundance and RPF counts (Appendix Fig. III.2). The abundance values and RPF counts, together with their correlation values, ruled out low abundance or coverage as a reason for the missed identifications. A suboptimal LTM treatment and/or TIS calling could help explain the lack of TIS recognition and the resulting absence of the corresponding proteins from the ribo-seq-derived search space. These results demonstrate the importance of reference databases and MS for the identification and validation of next-generation sequencing-derived translation products.

The combination of N-terminal COFRADIC and ribo-seq data identified a number of alternative TISs. Translation via these start sites produces protein isoforms with a different N-terminus if the new start site maintains the reading frame (e.g. the 5’-UTR extension in Fig. 3). If the start site is not in the same reading frame, completely different proteins will be generated. The selection of upstream TISs can also lead to the creation of uORFs, which influence the downstream protein synthesis from the main ORF [44, 45]. Roughly half of all mammalian transcripts contain one or more upstream TISs, which are often associated with short ORFs [23]. In contrast to the previously reported frequent occurrence of uORFs in human and mouse ribosome profiling data [22, 23], we were able to identify only one N-terminal peptide of an upstream overlapping ORF in the PIDD gene (Appendix Table III.1). This limited evidence for uORF protein products could be attributed to several factors, such as a bias towards upstream (near-) cognate start site identification from ribosome profiling data [46] or the rapid degradation, small size and possibly low abundance of uORFs.

As sequencing techniques become more generally accessible, ribosome profiling has become [24, 27, 28, 41, 47] and will continue to be a valuable addition to MS-based protein and peptide identification, possibly taking over the role of mRNA sequencing for ORF delineation. The benefits of ribo-seq include the positive correlation between protein abundance and ribo-seq footprint coverage and the ability to predict TISs with single-nucleotide precision. Despite the advantages of ribo-seq, MS-based validation will remain indispensable, not only for the general identification of proteins (through shotgun proteomics), but also for the validation of ribo-seq-derived (alternative) TIS predictions (by means of N-terminomics techniques such as COFRADIC [4]). Furthermore, unlike ribo-seq or any other transcriptome sequencing technique, MS provides true *in vivo* evidence of proteins or peptides, while taking potential co- and post-translational modifications into account. We also found that both reference protein sequence databases and ribo-seq-derived search spaces can miss protein identifications and that the best results were obtained when these databases were combined. Overall, our results show the usefulness of a ribo-seq-based proteogenomics approach. The ultimate goal will now be the construction of an automated pipeline for the easy conversion of ribo-seq data into a custom protein sequence search space that incorporates both sequence variation information and TIS prediction, ready to be searched for protein identifications.
Materials and Methods

Cell culture for proteomics

The HCT116 cell line was kindly provided by the Johns Hopkins Sidney Kimmel Comprehensive Cancer Center (Baltimore, USA). Cells were cultivated in DMEM medium supplemented with 10% fetal bovine serum (HyClone, Thermo Fisher Scientific Inc.), 100 units/ml penicillin (Gibco, Life Technologies) and 100 µg/ml streptomycin (Gibco) in a humidified incubator at 37°C and 5% CO₂. Prior to the proteomics experiments, the HCT116 cells were subjected to SILAC labeling [48] as part of another experiment that compared the wild type HCT116 cells to a double knockout line, which was differently labeled (manuscript in preparation). For the N-terminal COFRADIC analysis, cells were transferred to media containing 140 µM heavy \( ^{13}C_6^{15}N_4 \) L-arginine (Cambridge Isotope Labs, Andover, MA, USA). For the shotgun proteome analysis, cells were cultured in medium supplemented with 140 µM medium heavy \( ^{13}C_6 \) L-arginine and 800 µM heavy \( ^{13}C_6 \) L-lysine. To achieve a complete incorporation of the labeled amino acids, cells were maintained in culture for at least 6 population doublings.

Cell culture and sample preparation for ribosome profiling

The HCT116 cells for the ribosome profiling experiments were cultivated in McCoy's 5A (Modified) Medium (Gibco) supplemented with 10% fetal bovine serum, 2 mM alanyl-L-glutamine dipeptide (GlutaMAX, Gibco), 50 units/ml penicillin and 50 µg/ml streptomycin at 37°C and 5% CO₂. Cultures at 80-90% confluence were treated with 50 µM LTM [49, 50] or 100 mg/ml CHX (Sigma, USA) at 37°C for 30 min. Subsequently, cells were washed with PBS, harvested by trypsin-EDTA, rinsed again with PBS and recovered by 5 min of centrifugation at 300 × g, all in the presence of CHX to maintain the polysomal state. Cell pellets were resuspended in ice-cold lysis buffer, formulated according to Guo et al. (2010) [51] (10 mM Tris-HCl, pH 7.4, 5 mM MgCl₂, 100 mM KCl, 1% Triton X-100, 2 mM dithiothreitol (DTT), 100 mg/ml CHX, 1 × complete and EDTA-free protease inhibitor cocktail (Roche)), at a concentration of 40 × 10⁶ cells/ml. After 10 min of incubation on ice with periodic agitation, lysed samples were passed across QIAshredder spin columns (Qiagen) to shear the DNA. Subsequently, the flow-throughs were centrifuged for 10 min at 16,000 × g and 4°C. The recovered supernatant was aliquoted, snap-frozen in liquid nitrogen and stored at -80°C for subsequent ribosome footprint recovery and cDNA library generation.

Shotgun proteome analysis

4.2×10⁶ cells were lysed in 20 mM NH₄HCO₃ pH 7.9 by three rounds of freeze-thawing. Total protein concentration in cell extracts was measured using Biorad’s Protein Assay (Biorad Laboratories, Munich, Germany) and 2 mg protein material was used for downstream processing. Digestion was performed overnight using trypsin (Promega, Madison, WI, USA; enzyme/substrate, 1/50) after adding 0.5 M guanidinium hydrochloride and 2% ACN to aid in
protein denaturation. Methionines were uniformly oxidized to methionine sulfoxides by adding 20 µl of 3% (w/v) H₂O₂ to 100 µl sample (equivalent to 500 µg proteins) for 30 min at 30°C. For chromatographic separation 100 µl peptide mixture was then immediately injected onto an RP-HPLC column (Zorbax® 300SB-C18 Narrow-bore, 2.1 mm internal diameter × 150 mm length, 5 µm particles, Agilent). Following 10 min of isocratic pumping with solvent A (10 mM ammonium acetate in water/ACN (98:2 v/v), pH 5.5), a gradient of 1% solvent B increase per minute (solvent B: 10 mM ammonium acetate in ACN/water (70:30 v/v), pH 5.5) was started. The column was then run at 100% solvent B for 5 min, switched to 100% solvent A and re-equilibrated for 20 min. The flow was kept constant at 80 µL/min using Agilent’s 1100 series capillary pump with the 100 µL/min flow controller. Fractions of 30 sec wide were collected from 20 to 80 min after sample injection. To reduce LC-MS/MS analysis time, fractions eluting 12 min apart were pooled, vacuum dried and re-dissolved in 20 µl 20 mM tris(2-carboxyethyl)phosphine (TCEP) in 2% acetonitrile.

**N-terminal COFRADIC analysis**

HCT116 cells were lysed in 50 mM HEPES pH 7.4, 100 mM NaCl and 0.8% CHAPS containing a cocktail of protease inhibitors (Roche) for 10 min on ice and centrifuged for 15 min at 16,000 g at 4°C. The protein sample was then subjected to N-terminal COFRADIC as described by Staes et al. (2011) [4].

**LC-MS/MS analysis**

The shotgun proteomics sample was subjected to LC-MS/MS analysis using an Ultimate 3000 RSLC nano HPLC (Dionex, Amsterdam, the Netherlands) in-line connected to an LTQ Orbitrap Velos (Thermo Fisher Scientific, Bremen, Germany). The sample mixture was loaded on a trapping column (made in-house, 100 µm id × 20 mm, 5 µm beads C18 Reprosil-HD, Dr. Maisch). After back flushing from the trapping column, the sample was loaded on a reverse-phase column (made in-house, 75 µm id × 150 mm, 5 µm beads C18 Reprosil-HD, Dr. Maisch). Peptides were loaded in solvent A’ (0.1% trifluoroacetic acid, 2% ACN) and separated with a linear gradient from 2% solvent A’ (0.1% formic acid) to 50% solvent B’ (0.1% formic acid and 80% ACN) at a flow rate of 300 nl/min followed by a wash reaching 100% solvent B’. The mass spectrometer was operated in data-dependent mode, automatically switching between MS and MS/MS acquisition for the ten most abundant peaks in a given MS spectrum. Mascot Generic Files were created from the MS/MS data in each LC run using the Distiller software (version 2.3.2.0).

The N-terminal COFRADIC sample was analyzed on the LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) which was operated in data-dependent mode, automatically switching between MS and MS/MS acquisition for the six most abundant peaks in a given MS spectrum.

All the MS data were converted using the PRIDE Converter [52] and are available through the PRIDE database [53] with the dataset identifier PXD000304 and DOI 10.6019/PXD000304.
**Deep integration of OMICS data**

**Peptide and protein identification and interpretation**

The protein and peptide searches were performed against our custom database using X! Tandem Sledgehammer (2013.09.01.1) and OMSSA 2.1.9 in combination with the SearchGui (1.16.4) tool [54]. For the shotgun proteomics experiment, pyroglutamate formation of N-terminal glutamine, acetylation of N-termini (both at peptide level) and methionine oxidation to methionine-sulfoxide were selected as variable modifications. Heavy labelled arginine (\(^{13}\)C\(_6\)) and lysine (\(^{13}\)C\(_6\)) were selected as fixed modifications. Mass tolerance was set to 10 ppm on precursor ions and to 0.5 Da on fragment ions. The peptide charge was set to 2+, 3+, 4+. Trypsin was selected as the enzyme setting, one missed cleavage was allowed and cleavage was also allowed when arginine or lysine was followed by proline.

For the N-terminomics experiment, the generated MS/MS peak lists were searched with Mascot (version 2.3) [55]. Mass tolerance on precursor ions was set to 10 ppm (with Mascot’s C13 option set to 1) and to 0.5 Da on fragment ions. The peptide charge was set to 1+, 2+, 3+ and the instrument setting to ESI-TRAP. Methionine oxidation to methionine-sulfoxide, \(^{13}\)C\(_2\)D\(_3\)-acetylation on lysines and carbamidomethylation of cysteine were set as fixed modifications. Variable modifications were \(^{13}\)C\(_2\)D\(_3\) acetylation of N-termini, acetylation of N-termini and pyroglutamate formation of N-terminal glutamine (all at peptide level). \(^{13}\)C\(_6\)\(^{15}\)N\(_4\) L-arg was set as fixed modification. Endoproteinase semi-Arg-C/P (Arg-C specificity with arginine-proline cleavage allowed) was set as enzyme allowing for no missed cleavages.

Protein and peptide identification and data interpretation were done using the PeptideShaker algorithm (http://code.google.com/p/peptide-shaker, version 0.26.2), setting the FDR to 1% at all levels (peptide-to-spectrum matching, peptide and protein).

**Ribosome profiling**

100 µl of the clarified HCT116 cell lysate (equivalent to 4x10\(^6\) cells) was used as input for ribosome footprinting. The A260 absorbance of the lysate was measured with Nanodrop (Thermo Scientific) and for each A260, 5 units of ARTseq Nuclease (Epicentre) were added to the samples. The nuclease digestion proceeded for 45 min at room temperature and was stopped by adding SUPERase.In Rnase Inhibitor (Life Technologies). Next, the ribosome protected fragments (RPFs) were isolated using Sephacryl S400 spin columns (GE Healthcare) according to the procedure described in ‘ARTseq Ribosome Profiling Kit, Mammalian’ (Epicentre). The RNA was extracted from the samples using acid 125 phenol : 24 chloroform : 1 isoamyl alcohol and precipitated overnight at -20°C by adding 2 µl glycogen, \(\frac{1}{10}\)th volume of 5 M ammonium acetate and 1.5 volumes of 100% isopropyl alcohol. After centrifugation at 18,840 x g and 4°C for 20 min, the purified RNA pellet was resuspended in 10 µl nuclease free water.
Library preparation and sequencing

Libraries were created according to the guidelines described in the ARTseq Ribosome profiling Kit, Mammalian protocol (Epicentre). The RPFs were initially rRNA depleted using the Ribo-Zero Magnetic Kit (Human/Mouse/Rat, Epicentre), omitting the 50°C incubation step. Cleanup of the rRNA depletion reactions was performed through Zymo RNA Clean & Concentrator-5 kit (Zymo Research) using 200 μl binding buffer and 450 μl absolute ethanol. The samples were separated on a 15% urea-polyacrylamide gel and footprints of 26 to 34 nucleotides long were excised. RNA was extracted from the gel and precipitated. The pellet was resuspended in 20 μl nuclease-free water. Next, RPFs were end polished, 3’ adaptor ligated, reverse transcribed and PAGE purified. Five μl of circularized template DNA was used in the PCR reaction and amplification proceeded for 11 cycles. The libraries were purified with AMPure XP beads (Beckman Coulter) and their quality was assessed on a High Sensitivity DNA assay chip (Agilent technologies). The concentration of the libraries was measured with qPCR and they were single end sequenced on a Hiseq (Illumina) for 50 cycles. The ribo-seq libraries have been deposited in NCBI’s Gene Expression Omnibus [56] and are accessible through the GEO series accession number GSE58207 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE58207).

Swiss-Prot/ribo-seq integrated database construction

The merged database was constructed using all human Swiss-Prot proteins (downloaded from http://www.uniprot.org, version 2014_03) and the translation products obtained from the ribosome profiling experiment (Fig. 1). The ribo-seq-derived translation products were created from both the predicted (alternative) TIS genomic locations based on the LTM ribosome profiling information (according to Lee et al., 2012 [23]) and the corresponding mRNA sequences obtained from Ensembl (version 70) that displayed overall CHX ribosome protected fragment (RPF) coverage. After reconstructing the amino acid sequences, the Ensembl identifiers were mapped to Swiss-Prot identifiers (to safeguard uniformity) using the pBlast algorithm.

In order to remove redundancy introduced by the combination of the ribo-seq-derived translation products and the Swiss-Prot protein sequences, duplicated sequences were removed, retaining the custom sequence. Moreover, only the longest form of a series of gene translation products (N-terminal extended or canonical) was withheld in the combined database. The custom database contained 68,961 sequences as compared to the 20,264 proteins in UniProtKB-SwissProt version 2014_03. Extra information on the custom DB creation can be found in Menschaert et al. (2013) [24].

Correlation analysis

Only the transcripts identified in both Swiss-Prot and the ribo-seq-derived translation products were selected for the correlation analysis. Ribo-seq measurements were expressed as the number of ribosomal footprints per CDS (RPF count), hereby correcting for a possible
Deep integration of OMICS data

3'UTR and 5'UTR bias as suggested by Ingolia et al. (2011) [22]. Two quantitative measures for protein abundance based on spectral counts (emPAI [30] and NSAF [31]) were calculated using the shotgun data. While the first method uses the number of peptides per protein normalized by the theoretical number of peptides, the so-called protein abundance index (PAI), the NSAF method takes both the protein length and the total number of identified MS/MS spectra in an experiment into account. For each dbTIS transcript for which quantitative ribo-seq and shotgun proteomics information was available a Pearson correlation coefficient was calculated between its normalized RPF count and its normalized spectral count. When more than one ribo-seq-derived transcript corresponded with a particular Swiss-Prot protein sequence, the one with the highest normalized RPF count was used. The different normalization and identification approaches were combined with the following additional transcript filtering settings: i) no extra cutoffs, ii) only dbTIS transcripts with a validated MS/MS-based identification (meaning that the spectral count value was higher than 2), iii) only dbTIS transcripts with a total RPF count ≥ 200, and iv) only dbTIS transcripts with both a validated MS identification and an RPF count ≥ 200. All correlation coefficients were computed using log-transformed RPF and emPAI/NSAF measures.

Author contributions

References

Deep integration of OMICS data


RESULTS  Chapter IV

Novel insights into translational control
Adapted from:

**Fijałkowska D, Verbruggen S, Ndah E, Jonckheere V, Menschaert G and Van Damme P.** eIF1 modulates the recognition of suboptimal translation initiation sites and steers gene expression via uORFs. (manuscript in preparation)
RESULTS  Chapter IV

Novel insights into translational control

eIF1 modulates the recognition of suboptimal translation initiation sites and steers gene expression via uORFs.

Abstract

Alternative translation initiation mechanisms such as leaky scanning and reinitiation potentiate the polycistronic nature of human transcripts. By allowing for reprogrammed translation, these mechanisms can mediate biological responses to stimuli. We combined proteomics with ribosome profiling and mRNA sequencing to identify the biological targets of translation control triggered by the eukaryotic translation initiation factor 1 (eIF1), a protein implicated in the stringency of start codon selection. We quantified expression changes of over 4,000 proteins and 10,000 actively translated transcripts, leading to the identification of 245 transcripts undergoing translational control mediated by upstream open reading frames (uORFs) upon eIF1 deprivation. Here, the stringency of start codon selection and preference for an optimal nucleotide context were largely diminished leading to translational upregulation of uORFs with suboptimal start. Affected genes were implicated in energy production and sensing of metabolic stress. Interestingly, knockdown of eIF1 elicited a synergic response from eIF5 and eIF1B.
Introduction

Qualitative and quantitative characterization of gene expression is indispensable to understand dynamic phenotypes of eukaryotic cells. Through technological advances in high-throughput sequencing and proteomics, it is now possible to follow gene expression from transcription to protein turnover [1-3]. The remaining fundamental challenges in modern biology include the unraveling of the full diversity of proteoforms (i.e., the different molecular forms of proteins) [4, 5] expressed from single genes. An increasing line of evidence suggests that mRNA translation may both be a rapid means of gene expression control [6-8] as well as a major source of proteoforms [9-12]. However, genes undergoing translational control [6, 13] and regulation of proteoform expression [14-16] remain poorly investigated.

Alternative translation initiation mechanisms allow to select between multiple start codons and open reading frames (ORFs) within a single mRNA molecule. Here, the scanning ribosomes may omit less efficient upstream start codons (e.g. non-AUG start codons and start codons embedded in a suboptimal nucleotide context) to initiate translation downstream in a process referred to as leaky scanning [6, 17]. Besides, reinitiation [6, 17, 18] of translation may occur when post-termination ribosomes are retained on the mRNA molecule and recycled after completing translation of a typically closely spaced upstream ORF (uORF) to support translation of a downstream ORF. A particular role in alternative translation was postulated for short ORFs situated in the mRNA 5’ leaders or upstream and partially overlapping the main protein-coding sequence (CDS). Due to the directionality of ribosomal scanning, these short uORFs may regulate protein translation [19, 20] or even impact on the selection of alternative translation sites giving rise to alternative protein N-termini and thus N-terminal proteoforms [14-16]. The importance of uORFs was supported by sequencing of ribosome associated mRNA regions (ribosome profiling, or ribo-seq) [21, 22] which provided evidence for the ubiquitous translation from non-AUG start sites situated outside annotated protein-coding regions. Such uORFs were characterized in a variety of organisms and conditions [7, 8, 23-25], and their impact on the translation efficiency of proteins was found to be conserved amongst orthologous genes [23, 24]. Considering the direction of scanning, ribosome profiling experiments revealed that ribosomes distribute asymmetrically across ORFs, as they readily accumulate at translation initiation and termination sites [21], an effect which may be enlarged due to pretreatment with translation elongation inhibitors [21, 26], overall warranting caution when interpreting uORF expression levels. Importantly however, further studies revealed that ribosome footprints of 5’leaders generally resemble those of coding sequences, suggesting genuine translation in these regions [22].

Translation initiation is a determining control step in translation [27]. In consequence, translational control is mainly facilitated by eukaryotic translation initiation factors (eIFs) which may readily respond to (extra)cellular conditions by changing the global rates of
proteins at the ribosome. To reduce the high energy cost of protein production, translational control can be triggered by nutrient deprivation and accumulation of unfolded proteins [13]. Here, eukaryotic translation initiation factor 1 (eIF1) was shown to directly regulate the translation initiation rate at suboptimal translation initiation start sites [28, 29]. Besides, eIF1 protein levels and its phosphorylation were linked to reprogrammed translation of uORFs [30, 31] and responses to stress stimuli [31]. Although this protein plays a central role in translation initiation [32], a genome-wide assessment of its role in translational regulation is lacking. By combining tailored proteomic strategies with ribosome profiling and mRNA sequencing we here identified the biological targets of the translation control exerted by eIF1.

Results

Integrative OMICS to map the translational landscape

To obtain a comprehensive view of the cellular response upon reduced eIF1 levels, we performed detailed analyses of transcriptional, translational and protein level regulation at a genome-wide scale in the near-diploid colon cancer cell line HCT116. mRNA-seq, ribo-seq, and label-free steady-state proteomics experiments were performed upon siRNA mediated knockdown of eIF1 (si-eIF1) and compared to control (si-Ctrl) conditions (Fig. 1). In ribo-seq experiments we made use of lactimidomycin (LTM) [33] and cycloheximide (CHX) as translation inhibitors, enabling the study of translation initiation and elongation respectively. Subsequently, the PROTEOFORMER [34] pipeline was applied to map ribo-seq reads to the human genome, identify translation initiation sites (TIS) and assess translation efficiencies of specific ORFs across the genomic sequence.

To gain additional insights into the biological role of eIF1 and its paralog gene eIF1B, we complemented the studies performed upon eIF1 knockdown with label-free shotgun proteomics data from eIF1 knockout (eIF1KO) and eIF1B knockout (eIF1BKO) HAP1 cell lines.
Overall, using ribo-seq in HCT116 cells we identified potential TIS at 201,934 unique genomic positions (Fig. 2A) of which only 37% represented AUG codons (Fig. 2B), an observation in line with previous reports [21]. Given the plethora of potential translation start sites and annotated splice variants, an integrative OMICS analysis was facilitated by rationalized filtering of transcripts without any evidence of transcript-specific translation or poorly expressed transcripts [35]. A single representative transcript per translated gene was selected using experimental data and meta-data (see Materials and Methods section Ribosome profiling data analysis). Next, minimal read count thresholds were applied at the transcript (mRNA-seq), CDS (ribo-seq) and uORF (ribo-seq) level. Furthermore, to complement transcript-level data, expression proteomics in HCT116 cells was performed by searching against a protein sequence database containing all Ensembl annotated CDS sequences (“aTIS database”). Consequently, the “aTIS database” was also used in the subsequent protein expression profiling in HAP1 cells, allowing to investigate the relationship between protein expression changes upon knockdown and knockout conditions. Finally, to facilitate the identification of novel peptides expressed in HCT116 cells that correspond to non-synonymous mutations and alternative translation events detected by the PROTEOFORMER pipeline [34], we searched the HCT116 proteomics data using a protein sequence database containing ribo-seq predicted ORFs to which all missing Ensembl annotated CDS sequences (“custom database”) were added [34].
Novel insights into translational control

Figure 2 - Overview of the identified TIS and TIS-delineated ORFs. A. Initiation sites were called at 201,934 unique genomic positions corresponding to aTIS, CDS, 5'UTR, 3'UTR regions or non-translated transcripts (ntr). B. Percentage of TIS codon usage in the entire genome compared to upstream (uTIS) codon usage in 10433 selected transcripts. Frequencies of at least 10% were indicated. C. Transcripts with (at least one) upstream (uORF) and/or upstream-overlapping ORFs (u-oORF) in 5' leader sequences. D. Comparison of observed u-oORF and CDS length distribution. Nucleotide conservation around the AUG start codon was analyzed for several ORF categories using phastCons (E) and phyloP (F) scores. Each category comprises at least 500 AUG-initiated ORFs.

Using an integrative OMICS approach in HCT116 cells, we quantified the expression changes of 4,197 proteins and 10,433 transcripts with actively translated CDS. Additionally, in 7,083 (68%) of these transcripts, we detected translational activity in the 5' leader sequence, pointing to a total of 15,894 uORFs and 8,554 upstream-overlapping ORFs (u-oORFs) (Fig. 2C). As expected, ORFs located in the 5' leader were generally 10 to 20-fold shorter, with a median length of 84 nucleotides (nt) for the u-oORFs compared to 1,440 nt for CDS (Fig. 2D). Next, we compared evolutionary conservation patterns of annotated, upstream and intronic AUG-initiated ORFs [36], which revealed that u-oORFs have overall intermediate nucleotide conservation scores, nonetheless with clearly elevated scores around the start
codon [36] (phastCons and phyloP analysis, Fig. 2E-F) and -3 nucleotide position (phyloP analysis, Fig. 2F), corresponding to the Kozak consensus sequence hallmarks [37]. Despite the increased nucleotide conservation around upstream start sites (uTIS), u(-o)ORF sequence conservation was much lower compared to CDS, underscoring their potential regulatory rather than peptide- or protein-coding roles [36].

We then compared the steady-state levels of translation, mRNA and protein (Fig 3A). In line with previous reports in human and mouse [1, 38, 39], we detected moderate to good correlation of protein and mRNA levels (Pearson coefficient r = 0.57-0.58). Protein synthesis rate was believed to largely explain the remaining variability [1, 3, 38], especially in the case of non-perturbed systems, and indeed, we have observed a slightly improved correlation of ribo-seq readout of translation to protein levels (r = 0.62), in line with previous reports [1, 10]. Our results further highlight the importance of downstream processes such as protein turnover in establishing a proteome level equilibrium. Moreover, inherent limitations of the applied technologies, namely the fact that ribo-seq captures a “snap-shot” of translation while proteomics captures steady-state protein abundance, increase the difficulty in accounting for the potential delay in the manifestation of the translational response at the protein level [40], as may be the case in our study, since translatome and proteome samples were harvested at the same time-point after knockdown. Translation of uORFs however, was only weakly correlated (r=0.36) to protein changes (Fig. 3A), in line with the postulated regulatory role of translationally active 5’ leaders.
Figure 3 - Correlation and reproducibility at different OMICS levels. A. Values represent average Pearson correlation coefficients for quantitative OMICS experiments performed in the si-Ctrl setup of HCT116 cells. LFQ - label free quantification of protein expression. B. Identification of differentially expressed genes (p=0.01) based on adjusted significance threshold in relation to data distribution obtained by Z-score transformation (Z-score cutoffs values of -2.58 and 2.58 corresponding (p=0.01) are indicated in red). C. Significantly regulated proteins in the HCT116 shotgun proteomics experiment were clustered alongside their corresponding mRNA, CDS and TE CDS Z-score fold changes. Additionally, HAP1 LFQ data - not used for clustering - was visualised. Columns with protein data represent log2 fold change between average LFQ intensities in knockdown/knockout compared to control conditions. The scale ranges from red (upregulation in knockdown/knockout) to green (downregulation in knockdown/knockout) Clusters depict different modes of protein regulation: cluster 1 – transcriptional downregulation; cluster 2 – transcriptional downregulation not reflected by CDS translation; clusters 3 and 4 - downregulated and upregulated proteins deviating from the expected transcriptional and translational regulation (31 proteins), respectively; cluster 5 – transcriptional upregulation; cluster 6 – transcriptional upregulation not reflected by CDS translation; cluster 7 – translational downregulation; cluster 8 – mixed (transcriptional and translational) downregulation; cluster
An OMICS perspective on eIF1 translational control

To study expression changes in response to eIF1 deprivation at the level of transcription, and CDS and uORF translation, we calculated ratios (fold changes) of normalized read counts between the eIF1 knockdown (si-eIF1) and control (si-Ctrl) conditions. Translation efficiency (TE) of the CDS or uORF was estimated by dividing normalized ribo-seq reads with normalized mRNA read count data of the corresponding transcript. Fold changes in TE were further calculated between both conditions. Finally, to determine the regulatory effect of uORFs on their downstream CDS, a ratio between uORF TE fold change and CDS TE fold change was calculated. In order to identify ORFs and transcripts significantly affected by differential eIF1 expression, we applied a Z-scoring strategy with adjustment for expression, as described by Andreev et al. [7]. To identify significantly regulated ORFs and transcripts, data points were grouped in bins according to read counts, and Z-score normalization was performed per bin. This approach allowed to assure sufficient stringency of the analysis in case of lowly expressed genes displaying a broader distribution of expression changes, while remaining sensitive to smaller fold changes in the case of highly expressed genes. Using a threshold p-value of 0.01 we detected significant deviations in response to eIF1 knockdown for 159 mRNAs, 125 CDSs and 291 u(-o)ORFs (Fig. 3B, Appendix Table IV.1). Additionally, 121 CDSs and 313 u(-o)ORFs (i.e. 81 u-oORFs and 232 uORFs) were affected at the TE level. Finally, differential regulation of translation efficiency was confirmed for 330 u(-o)ORF/CDS pairs (i.e. u(-o)ORF and CDS ORF are located on the same transcript), comprising 68 u-oORFs/ and 262 uORFs/CDS pairs, originating from 245 unique affected transcripts (Fig. 4A). These genes represented a broad spectrum of expression levels (Fig. 4A) and different modes of regulation (Fig. 5, 6 and Appendix Table IV.1), including u(-o)ORF initiated at non-AUG uTIS with a possible inhibitory effect on CDS expression and AUG-initiated u(-o)ORF potentially enhancing CDS expression, amongst others. Protein expression data revealed 238 significantly regulated proteins identified by an ANOVA test with a p-value of 0.05 (Appendix Table IV.1). A systematic comparison of affected genes highlighted that the majority of protein changes resulted from either transcriptional (41%), translational (28%) or concomitant transcriptional and translational regulation (18%) of gene expression, next to 31 proteins (13%) displaying significant deviations in expression likely caused by perturbed turnover and/or posttranslational modifications amongst others (Fig.3C).
Novel insights into translational control

Figure 4 - Translational response to elf1 deficiency is in part mediated by uORFs and depends on identity of the uTIS codon. A. 330 differentially regulated uORF/CDS pairs (from 245 transcripts) were identified at a 0.01 confidence level. Gene names were indicated for selected examples, also visualized in Fig. 5, Fig. 6 or Appendix Fig. IV.1. B. Lower elf1 levels allow for translational upregulation of non-AUG uORFs and downregulation of AUG uORFs. Only significantly regulated uORF/CDS pairs were presented. C. The relationship between uORF start codons and the expected directionality of translational regulation is maintained for both upstream and upstream-overlapping ORFs. D. The effect of u(o-)ORF start site on downstream CDS translation.

uORF features associated with elf1 mediated translational control

Our analysis confirmed that uORFs are implicated in translational control directed by elf1, thereby contributing to regulating protein expression. Therefore, we decided to elucidate if certain intrinsic features of uORFs may determine their potential to enhance or repress the expression of downstream CDS. Studies reported by the Atkins group demonstrate that elf1 levels may orchestrate the stringency of start codon selection [28, 30]. More specifically, overexpression of elf1 resulted in preference for AUG initiation. In light of these findings, we hypothesised that decreased elf1 levels should lead to more flexibility in ribosomes initiating
Figure 5 - Genes under uORF regulation. We detected a differential translational regulation of CDS and at least one uORF for 245 unique transcripts (genes) (Z-score fold change TE uORF/TE CDS). Clustering of their corresponding expression values revealed 5 modes of regulation with predominant effect of translation, transcription or a combined effect of both processes. Selected genes containing regulatory uORF(s), previously shown in Fig. 4A, were highlighted in red on the orange background of the entire cluster profile.

Translationally upregulated

Translationally downregulated

Transcriptionally upregulated

Mixed downregulation

Transcriptionally downregulated

translation at non-AUG codons. Whereas the vast majority (99.6%) of annotated CDS detected in our study have AUG initiation sites, u(-)ORFs display a broad spectrum of TIS codons (Fig.2B). In consequence, eIF1 knockdown, by increasing initiation (rates) at near-cognate start sites, is expected to impact the rate of leaky scanning and reinitiation, leading to altered incidences of both uTIS and aTIS initiation. Indeed, we observed a highly significant dependence between uTIS codon identity and the regulation of uORF/CDS pairs (Kruskal Wallis test p=4.1e-14). This relationship was further confirmed on a subset of transcripts with a single uORF (p=0.0014). uORFs with non-AUG start codons had frequently
upregulated uORF/CDS TE ratio (Z-score TE uORF/CDS ≥ 2.58), whereas AUG-initiated uORFs were typically downregulated compared to their CDS (Z-score TE uORF/CDS ≤ -2.58; Fig.4B; Chi2 test p=0.00035). Increased expression of non-AUG uORFs and decreased expression of

![Graphs showing Ribosome density profiles of transcripts with actively translated u(-o)ORFs]

Figure 6 - Ribosome density profiles of transcripts with actively translated u(-o)ORFs. Full length ribosome CHX reads of si-Ctrl and si-eIF1 samples were visualized across mRNA sequences as positive (blue) and negative (red) values, respectively. Only the 5’ proximal part of the 3’UTR sequence was included. Both uniquely and non-uniquely mapped reads were included. CDS and u(-o)ORFs were marked using grey or green/red/blue/yellow bars, respectively. Direction of expression changes (si-eIF1 vs. si-Ctrl) and start codon of the most 5’ proximal u(-o)ORF were indicated (see figure). Potential regulatory u(-o)ORFs were visualized for PHLDA2, ALDH1A3 FZD3 and AIM2. These u(-o)ORFs contributed to significant Z-score TE uORF/TE CDS changes.
AUG uORFs was also apparent when considering uORFS and u-oORFs separately (Fig. 4C). When considering the impact of the uTIS codon on TE of the downstream CDS (Fig. 4D), AUG uTIS codons more often associated with downregulation of uORF expression and enhanced expression of the corresponding CDS, while non-AUG-initiated uORFs displayed increased expression, thereby acting as CDS repressors. These results demonstrate that by relying on the principle of leaky scanning, eIF1 steers the stringency of start codon selection, thereby exerting translational control on protein-coding genes at a genome-wide scale. The direction (and likely also the degree) of eIF1-induced regulation is dependent on the cellular availability of the translation factor and on the nature of upstream and downstream start codons.

Higher CDS translation efficiency also coincided with increased distance between the uORF stop codon and the aTIS compared to a lower distance to uORF observed when TE CDS was repressed (Wilcoxon test p=0.00091, Fig. 7A). These results confirm that the relationship between repressiveness of uORFs and their distance to CDS [23] are relevant parameters upon eIF1 deprivation and suggest that more distant uORFs may act as enhancers of CDS translation. Sufficient spacing between the uORF and CDS is necessary to allow the occurrence of translation reinitiation in eukaryotes [41]. Kozak et al. [41] illustrated how 79 nt spacing stimulated efficient reinitiation in mammalian cells by enabling the 40S ribosome subunit to reload with translation initiation factors and initiator methionine tRNA, while resuming the scanning of mRNA. On the contrary, and similar to previous reports [24], we did not observe any relationship between the u-oORF TIS/aTIS distance or the overlap with CDS and the significance of the fold change in TE of the CDS (Fig. 7B, Appendix Fig. IV.2A).

Next we investigated if the number of u(-o)ORFs is relevant for CDS expression levels. Higher ribo-seq coverage of transcripts and, by extension, higher translational signals, coincided with an increasing number of detected u(-o)ORFs (Kruskal Wallis test shows significant relationship between log2 CDS ribo-seq read counts and u(-o)ORF count, p = 6.34e-07; Wilcoxon test shows decreased CDS read counts in transcripts with no u(-o)ORFs, p = 7.73e-09). Despite this potential bias due to sequencing coverage, translation efficiency was dependent on the number of u(-o)ORFs (Kruskal Wallis p < 2.2e-16), and TE rates at steady-state were generally higher for CDS without u(-o)ORFs (Wilcoxon p < 2.2e-16). Interestingly, CDSs with one or more u(-o)ORFs were found to be significantly more repressed upon eIF1 knockdown (lower Z-score of TE CDS fold change; Wilcoxon p = 0.0007; Fig. 7C).

Next, we sought to determine whether the primary nucleotide context surrounding uTIS codons was implicated in steering eIF1 regulation of u(-o)ORFs. Therefore, we first retrieved the sequence context of all aTIS called by our PROTEOFORMER pipeline. As described by Grzegorski et al. [42], we assigned the entire nucleotide context with a single numeric score based on the preferred nucleotide frequency at particular positions. The best scoring context in the aTIS set (GCCACCCXXXG, score=345.8) corresponded to the Kozak consensus sequence (GCC(A/G)CCXXXG, [37]). Of note however, this was not the most frequent motif observed in
Novel insights into translational control

our aTIS dataset, instead GGGAAGXXXGC (score= 303.7) was most frequently observed (0.19 %). Overall, and in line with the original report [42], higher aTIS scores were associated with an overall higher incidence (Fig. 7D) as well as higher levels of expression and increased translation efficiency of the corresponding CDS (Fig. 7E; Wilcoxon p ≤ 8.917e-10). Using the scoring system established for aTIS, we assigned scores to uTIS context sequences. Although uTIS identified in our dataset were clearly enriched in non-AUG codons, the nucleotide context score of these near-cognate start sites was generally much higher compared to AUG uTIS (Fig. 7F, Wilcoxon p < 2.2e-16) [43, 44]. This stronger nucleotide context around non-AUG uTIS was further confirmed for significantly regulated u(-o)ORF/CDS pairs (Wilcoxon p = 0.028). Interestingly, the preference for an optimal consensus sequence was perturbed upon eIF1 knockdown, resulting in decreased uTIS context scores in the group of regulated compared to non-regulated u(-o)ORF/CDS pairs (Fig. 7G; Wilcoxon p = 0.115 and p = 0.0034 for Z-score TE uORF/CDS threshold 0.01 and 0.05, respectively). The same was observed in transcripts with one u(-o)ORF (Wilcoxon p = 0.014 for Z-score TE uORF/CDS threshold 0.01). In contrast, aTIS nucleotide context scores were indistinguishable when comparing regulated versus non-regulated u(-o)ORF/CDS pairs (Fig. 7H). When considering transcripts with a single u(-o)ORF, a decreased quality of the uTIS context sequence in relation to aTIS was especially apparent for regulated AUG uTIS compared to the non-regulated AUG uTIS group (Fig. 7I, Wilcoxon p = 0.0751 and p = 3.61e-05 for Z-score TE uORF/CDS threshold 0.01 and 0.05, respectively). To further assess the relationship between uTIS context and the direction of u(-o)ORF regulation we turned to a more simplified metric. The quality of context sequence was assumed to be “strong” (indicated by 3 purine and 4 guanine relative to uTIS) or otherwise “weak” [19]. Using this metric we observed that upregulated AUG uTIS had significantly weaker context compared to downregulated AUG uTIS (Fisher’s exact test p = 0.081 and p = 0.026 for Z-score TE uORF/CDS threshold 0.01 and 0.05, respectively). In contrast, such relationship was not detected for non-AUG uTIS, results corroborated by sequence logo analysis (http://weblogo.berkeley.edu, [45]) presented in Appendix Fig. IV.3). These results allow us to conclude that eIF1 knockdown perturbed translation initiation rates at u(-o)ORFs when their start codon was embedded in a suboptimal nucleotide context sequence. Although translation initiation rates at uTIS with a poor context was affected upon eIF1 knockdown, uTIS start codon identity (AUG vs. non-AUG) seemed to be a stronger determinant of eIF1-driven start site selection.
Figure 7 - uORF and uTIS features relevant for translational regulation. A. Spacing between uORF stop codon and aTIS has impact on CDS translation irrespective of the uTIS codon (AUG vs. non-AUG). B. Distance between u-oORF start site and aTIS has no clear impact on CDS translation. C. Translation efficiency of CDS is decreased in the presence of uORF(s) (left panel) despite the fact that highly expressed CDS tend to have more uORFs (right panel). D. Correlation between aTIS context score and its frequency of occurrence in the human genome. E. Higher aTIS score can be expected for well expressed (left panel) and efficiently translated (right panel) CDS at the steady-state level (si-Ctrl conditions). mRNA abundance is also associated with aTIS score, although to a lesser extent. F. Cumulative frequency plot demonstrating that non-AUG uTIS are embedded in more optimal nucleotide context compared to AUG uTIS. G. Relaxed stringency of start codon selection by eIF1 allowed for significant regulation of uORFs with suboptimal uTIS context. H. Nucleotide context of the downstream aTIS had no apparent impact on the regulation. I. The impact of differences in uTIS and aTIS context scores on regulation in transcripts with 1 detected u(-o)ORF. * Wilcoxon test p-value < 0.05.

Impact of eIF1 knockdown on cell metabolism and energy status

To investigate the consequences of eIF1 deficiency at the cellular level, we analysed changes in protein expression upon si-eIF1 treatment in HCT116 cells. Additionally, we validated our findings with label-free shotgun proteomics data from two independent eIF1 knockout
Novel insights into translational control

(eIF1KO) HAP1 cell lines, which have shown overall good agreement with si-eIF1 results (Fig. 3C, clusters 5 and 8). Combined annotation enrichment analysis of ribo-seq and proteomics data revealed a decreased translation efficiency and, concomitantly, lowered expression of genes involved in glycolysis/gluconeogenesis (TE CDS FDR = 0.00027, LFQ si-eIF1 FDR = 0.00044, LFQ eIF1KO FDR = 0.00066) and in the TCA cycle (TE CDS FDR = 0.00021, LFQ eIF1KO FDR = 0.0021). In light of these findings, we decided to measure cellular ATP levels upon eIF1 knockdown (Fig. 8A). We confirmed that reduced expression of glycolytic genes was accompanied by decreased cellular ATP levels, hinting to a more general impairment of energy metabolism induced by eIF1 deficiency. Further analysis of gene subsets affected by translationally active 5’ leaders (significant Z-score TE uORF/CDS) against a background of all quantified genes was performed using GOrilla (Gene Ontology enrichment analysis and visualization tool) [46]. This analysis revealed translational downregulation of mitochondrial outer membrane translocase complex components (p = 0.000027). To test mitochondrial activity upon si-eIF1 treatment, we stained HCT116 cells with JC-10, a dye useful for determining mitochondrial membrane potential by flow cytometry. CCCP pre-treatment was performed as a positive control of absolute membrane depolarization. Overall, our results indicated that si-eIF1 treated cells may suffer from decreased mitochondrial activity (Fig. 8B).

Similar GO-based analysis of the up-regulated genes pointed to the increased expression of ribosomal proteins and thus increased ribosomal biogenesis, enhanced aminoacyl-tRNA synthesis and amino acid transport throughout all OMICS levels. More specifically, ribosome biogenesis (CDS FDR = 0.0065, TE CDS FDR = 0.0019), ribosomal protein (CDS FDR = 0.011, mRNA FDR = 0.00006, LFQ si-eIF1 FDR = 5.08e-11), aminoacyl tRNA biosynthetic process (mRNA FDR = 0.0083, CDS FDR = 0.0035, LFQ si-eIF1 FDR = 0.0025, LFQ eIF1KO FDR = 0.009) and amino acid transport (mRNA FDR = 0.017, CDS FDR = 0.00077, LFQ eIF1KO FDR = 0.0073) were significantly upregulated terms. Ingenuity Pathway Analysis (IPA) of ribo-seq data (Z-score fold change CDS) further identified a causative relationship between the transcription factor ATF4 and the upregulation of processes related to amino acid metabolism (Fig. 9A). In line, our data point to the increased expression of ATF4, which was linked to the translational downregulation of AUG uORFs in the 5’ leader of ATF4 upon eIF1 knockdown (Z-score TE uORF/CDS ≤ -2.58).
Figure 8 - Energy status upon eIF1 deprivation. A. Chemiluminescent readout of cellular ATP levels (lower panel). Glucose deprivation was additionally performed as a positive control for ATP level reduction. Knockdown efficiency was monitored by Western blot (upper panel) using anti-eIF1/1B primary antibody sc-49187 (Santa Cruz). Beta-actin was probed as a loading control using A2066 primary antibody (Sigma). *- unspecific band. B. Measurement of mitochondrial membrane potential by FACS. **- t test significance p<0.01.

Equilibrium of eIFs

Interestingly, knockdown of eIF1 elicited a synergic response from other eIFs. Our proteomics data suggest a significant regulation of four other eukaryotic initiation factors (or subunits thereof) upon eIF1 knockdown, including EIF1B, EIF2S1 (encoding eIF2α), EIF5 and EIF4B (Fig. 9B). More specifically, we observed a significant downregulation of eIF5 and eIF2α, while eIF4B and eIF1B were upregulated. In support of these findings, a similar expression profile was detected in the HAP1 eIF1KO cell lines, with the sole exception of eIF2α that was found significantly upregulated. Interestingly, eIF5 and eIF1 were previously reported to have opposing effects regarding the stringency of start codon selection [30]. eIF5 expression is regulated by uORFs with poor context AUG TIS and therefore expected to decrease when translational initiation equilibrium is shifted towards more flexible start codon selection [30]. In consequence, and in line with our results, one would predict that eIF5 expression is decreased upon eIF1 knockdown. Low eIF5 levels should counteract eIF1 knockdown to reconstitute the fidelity of translation initiation, however restoration
Figure 9 - OMICS analysis highlights the effect of eIF1 deficiency on the ATF4 interaction network and suggests a widespread remodelling of the translational machinery. A. Upregulation of ATF4 expression coincided with increased expression of its downstream targets involved in amino acid synthesis (ASNS Asparagine Synthetase (Glutamine-Hydrolyzing), PHGDH Phosphoglycerate Dehydrogenase involved in the early steps of L-serine synthesis and PSAT1 Phosphoserine Aminotransferase 1 that participates in glycine, serine and threonine metabolism) next to genes involved in amino acid transport (SLC3A2, SLC6A9 and SLC7A5 Solute Carrier Family members that transport L-type amino acids).
Additionally, the expression pattern of \textit{ATF4} upstream regulators, such as \textit{EIF2S1}, \textit{EIF2AK1} and \textit{GCN1L1}, were visualized. IPA analysis indicated a common node (\textit{FASN} Fatty Acid Synthase) between \textit{ATF4} and \textit{ERBB2} interaction network, which pointed to a possible impairment of adherent junction formation (due to downregulated \textit{KRT7} Keratin 7, \textit{SMAD} Family Member 3 and \textit{ERBB2} Receptor Tyrosine Kinase 2). In contrast to the \textit{ATF4} network, observation concerning adherent junctions could not be confirmed at the protein level. B. elf proteins significantly regulated upon elf1 knockdown were also found to be significantly regulated in elf1KO cells. Differential expression values were subjected to hierarchical clustering using normalized protein expression fold changes next to TE CDS, CDS and mRNA Z-score fold changes as input.

of the equilibrium is likely not possible when elf1 expression is persistently inhibited. elf1B, on the other hand, is an elf1 paralog gene whose role in translation initiation remains to be determined. elf1 and elf1B share 97% sequence similarity. Viewing the fact that both elf1 and elf1B were proven to be non-essential genes in HAP1 cells (both produced viable knockout cell lines), certain functional redundancy may be expected between these proteins. In consequence, elf1B upregulation may (in part) counteract elf1 deficiency.

\section*{Discussion}

By combining tailored proteomic strategies with next generation sequencing \cite{9, 10, 47} we aimed at identifying the biological targets of translation control exerted by elf1. Integrative OMICS studies face the problem of limited correlation observed between protein and mRNA levels. We improved on this aspect by relying on the ribosome-profiling readout for translation. We however detected 31 genes with significant changes at the protein level, independent of mRNA expression and translation efficiency. Further, in numerous cases, proteins levels only moderately corresponded to their synthesis rates (Fig. 3C, clusters 2 and 6). There may be many reasons as to why such discrepancies are observed. First of all, protein synthesis is delayed compared to transcriptional regulation, which reduces the overall correlation between mRNA and protein fluctuations \cite{40}, especially when sampling RNA and protein at the same time point after knockdown. Additionally, ribo-seq provides a snapshot of translational engagement of ribosomes, which may correspond to a transient state, while the magnitude of expression changes might be insufficient to impact on steady-state protein levels. Finally, compensatory effects (such as the possible elf1B activity) might be at play.

Using complementary high-throughput technologies, we here confirmed that elf1 levels determine the stringency of start codon selection at the genome wide scale and thereby orchestrate the rates of leaky ribosomal scanning and uORF translation. More specifically, low elf1 levels promote translation initiation at near-cognate codons and start sites embedded in a suboptimal nucleotide context (Fig. 10). The initiation context was previously also found to determine aTIS versus downstream TIS (dTIS) selection \cite{9}. Although our data
Novel insights into translational control suggests, that the direction (and likely also the degree) of eIF1-induced regulation is dependent on the cellular availability of the translation factor and on the nature of upstream and downstream start codons, eIF1 modification status [31], not monitored in this study, is another potential factor involved.

Figure 10 – Translation initiation upon eIF1 deficiency. Our data suggest that upon eIF1 knockdown the stringency of start codon selection and preference for an optimal nucleotide context were largely diminished. Overall, this leads to translational upregulation of uORFs with suboptimal start and decreased translation initiation rates at upstream AUG codons embedded in more optimal nucleotide context.

Although uORFs are omnipresent throughout protein-coding transcripts, uORF mediated regulation significantly affects only a subset of genes and tends to have an overall moderate effect on absolute expression levels (average fold change in CDS expression of about +/-25%), a range similar to miRNA effects [24]. Therefore, we may consider that the predominant role of uORFs involves the fine-tuning of proteome homeostasis, buffering the effects of stress conditions for most genes while providing the capacity for stress response to particular effectors [7]. Although the expression and functionality of uORF-derived peptides is debated with only a limited number of active peptides identified and characterized so far [48], proteogenomics strategies hold promise in expanding our knowledge in this field. We searched our shotgun proteomics data against a “custom database” enriched for ribo-seq delineated reading frames, leading to the identification of peptides derived from one uORF in the CYP4F11 gene and six u-oORFs (in MFGEB, POLR2M, SAMD1, PSMG4, SLC39A13 and RSU1 genes) (see Appendix Fig. IV.1 and Appendix Table IV.2 for these and other examples of novel proteoforms and non-synonymous variants identified in our study). Additionally, our
ribo-seq data has predicted the expression of 14 short ORF-encoded polypeptides previously identified by Slavoff et al. [49] and 60 alternative ORFs with peptide evidence reported by Vanderperre et al. [50]. Of note however, 10 peptides in the Slavoff et al. dataset and 118 peptides from the Vanderperre et al. study that were attributed to alternative proteoforms, belonged to either Ensembl or SwissProt annotated proteins. These discrepancies were likely due to updated database versions used in our study or differences between Ensembl/SwissProt and NCBI (RefSeq) annotation used as reference by the other studies.

Reliable quantification of uORF expression by ribo-seq may be challenging due to their short length and the bias in ribosome signal at 5’ leader introduced by the antibiotic treatment [26]. Nevertheless, potential biases are similar across all samples analysed, thus unlikely affect differential expression analysis, corroborated by the fact that some of the regulatory uORFs identified in our study were previously reported in studies that specifically avoided antibiotic pretreatment [7, 8] (see Appendix Fig. IV.1). Our data analysis pipeline also minimized the impact of ribosome accumulation at start and stop sites by adjusting the region used for CHX readout of translation (see Materials and Methods). Analysis of overlapping ORFs may be additionally difficult. For example, 68 of differentially regulated u-oORF/CDS pairs identified in our study were characterised by shorter than average overlaps (Appendix Fig. IV.2B), a likely consequence of including the region shared with the CDS for calculating u-oORF expression. Although our approach may underestimate the number of regulated u-oORFs, namely u-oORFs extensively overlapping with CDS, currently precise expression measurement of highly overlapping ORFs remains unfeasible using ribo-seq.

Our findings suggest the role of eIF1 and its cellular levels as a key mediator for translational regulation, but also underline the high interconnectivity of the translational machinery [1, 8, 40]. The non-essential nature of eIF1 gene in HAP1 cells was not originally anticipated [51] and the nearly 3-fold upregulation of eIF1B in eIF1KO cells supports the possible functional overlap between the paralog genes. On the other hand, no clear translational impairment, especially no apparent change in eIF1 expression, and an overall mildly perturbed to unperturbed proteome expression profile of eIF1BKO cells prohibit from drawing any conclusion regarding eIF1B activity without additional in vitro studies.

Overall, the effects of eIF1 deprivation comprise a widespread remodelling of the eIF network. Energy production and amino acid demand also seemed to be significantly perturbed. In particular, lower cellular ATP levels and decreased mitochondrial respiration were detected, however the severity of this phenotype needs further examination. Interestingly, changes in gene expression were frequently achieved by engaging uORFs. Particularly strong upregulation of gene expression mediated by uORFs was detected for ATF4, which coincided with enhanced expression of its downstream targets (Fig. 9A). Although our study indicates that ATF4 regulation was elicited by eIF1, previous reports indicate another likely reason for enhanced ATF4 translation. It involves the downregulation (inactivating phosphorylation) of eIF2α by upstream kinases in response to stress stimuli [6,
To set our findings in a broader context, we investigated the expression of *EIF2S1* (encoding eIF2α subunit) and *EIF2AK4*, encoding a kinase that phosphorylates eIF2α in response to amino-acid deprivation [13]. Although eIF1 knockdown coincided with decreased *EIF2S1* transcription, translation and protein expression, we have observed no significant expression change in case of *EIF2AK4*, while *GCN1L1* encoding a positive activator of the EIF2AK4 protein kinase activity, was clearly downregulated (Fig. 9A). Additionally, and despite the overall good agreement between the knockdown and knockout experiments, both eIF1KO cell lines showed increased eIF2α expression (Fig. 9B), suggesting that eIF1 deficiency is potentially sufficient to elicit the observed response and eIF2α is not a necessary mediator of the process.
Materials and Methods

Cell culture

The human colon cancer cell line HCT116 was kindly provided by the Johns Hopkins Sidney Kimmel Comprehensive Cancer Center (Baltimore, USA). The HAP1 wild type and CRISPR/Cas9 engineered knockout cell lines were obtained from Horizon Genomics GmbH, Vienna. In particular, a single elf1B knockout clone and two elf1 knockout clones were acquired (i.e. elf1-14bp deletion knock out (elf1KO cl. 1) and elf1-265bp insertion knock out (elf1KO cl. 2)). HCT116 cells were cultivated in McCoy’s medium (Gibco, Thermo Fisher Scientific Inc.) whereas HAP1 cells were grown in Iscove’s Modified Dulbecco’s Medium (IMDM, Gibco, Thermo Fisher Scientific Inc.). All media were supplemented with 10% fetal bovine serum (FBS, HyClone, Thermo Fisher Scientific Inc.), 2 mM alanyl-L-glutamine dipeptide (GlutaMAX, Gibco, Thermo Fisher Scientific Inc.), 50 units/ml penicillin (Gibco, Thermo Fisher Scientific Inc.) and 50 µg/ml streptomycin (Gibco, Thermo Fisher Scientific Inc.). HCT116 and HAP1 cells were maintained in a humidified incubator at 37°C and 5% CO₂ or 8% CO₂, respectively.

Knockdown experiments in HCT116 cells

For knockdown experiments, HCT116 cells were transferred to serum-free McCoy’s medium and seeded at a density of 2.4*10⁶ cells per 10 cm² plate. Cells were transfected with either 10 nM control si-RNA (si-Ctrl, ON-TARGETplus Non-targeting Control siRNAs: D-001810-01-05, Dharmacon, GE Healthcare Life Sciences) or 10nM si-RNAs targeting elf1 (si-elf1, SMARTpool: M-015804-01-0005, Dharmacon, GE Healthcare Life Sciences) using 63ul of HiPerFect (QIAGEN) per 10 cm² plate. 5 hours post-transfection, cells were supplemented with FBS at a final concentration (f.c.) of 10%. 1 plate per condition was harvested after 72 hours for label-free proteomics and qPCR experiments using Trypsin/EDTA solution (Gibco, Thermo Fisher Scientific Inc.). The remaining 4 plates per condition were used for the ribosome profiling and mRNA-sequencing experiments performed.

Label-free shotgun proteomics

For label-free shotgun proteomics experiments in HAP1 cells, 3 biological replicates of WT cells, elf1B knockout and both elf1 knockout clones were prepared. For the label-free shotgun proteomics experiment in HCT116 cells, 2 biological replicates of WT cells and si-elf1 were prepared in parallel with the ribo-seq experiment. 10 million cells per replicate were harvested and lysed by 3 rounds of freeze-thaw lysis in 300µL Gu.HCl lysis buffer (4M Gu.HCl, 50 mM NH₄HCO₃ pH 7.9). The lysates were sonicated using a Branson probe sonifier (output 20, 0.5 second pulses, 3 x 30 seconds, after each 30 seconds cooled on ice) and centrifuged for 30 minutes at 3,500g (4 °C). Supernatants were transferred to a fresh protein-low-bind tube (Eppendorf) and the protein concentration was determined according to the Bradford method. An aliquot equivalent of 200 µg was transferred to a clean 2 mL-
tube and diluted to 1 mg/mL with 4M Gu.HCl + 50 mM NH₄HCO₃. The protein mixture was further diluted with an equal volume of HPLC grade water followed by a precipitation with 4x volumes of -20 °C acetone for 2 hours at -20°C. Precipitated proteins were collected by centrifugation at 4,000g for 15 minutes (4 °C). Pellets were washed twice with 1 mL of ice-cold 80% acetone by vortexing and centrifugation at 16,000g for 5 minutes. Pellets were air-dried upside down for approximately 10 minutes at RT until no residual acetone odor remained. Protein pellets were then resuspended in 200µL TFE (2,2,2-trifluoroethanol) digestion buffer (11% TFE, 100mM NH₄HCO₃) and the pellets dissolved by sonication using a Branson probe sonifier (output 10-15, 0.5 second pulses, cool samples on ice if necessary) until a homogenous suspension was formed. Samples were digested overnight at 37°C using mass spectrometry grade trypsin/Lys-C mix (enzyme/substrate of 1:50 w/w, 4µg) while mixing at 550 rpm. The samples were acidified on the next day with trifluoroacetic acid (TFA) added to a final concentration of 0.5% and centrifuged 10 minutes at 16,000g. The supernatants were transferred to a fresh protein-low-bind tube and H₂O₂ was added to each sample at a f.c. of 0.5% and incubated for 30 minutes at 37 °C. Solid phase extraction of peptides was performed using C18 reversed phase sorbent containing 100µL pipette tips (Piez C18 tips – Thermo Scientific) according to the manufacturer’s instructions. The pipette tip was conditioned by aspirating the maximum pipette tip volume of water-ACN (50:50, v/v) and the solvent discarded. After equilibration of the tip by washing 3 times with the maximum pipette tip volume in 0.1% TFA in water (separate tubes and solvent prepared for each sample), 100 µL of the acidified samples (approximately 100 µg of proteins) were dispensed and aspirated for 10 cycles to achieve maximum binding efficiency. The tip was washed 3 times with the maximum pipette volume of 0.1% TFA in water-ACN (92:2, v/v) and the bound peptides eluted in LC-MS/MS vials with the maximum pipette volume of 0.1% TFA in water-ACN (30:70, v/v) by aspirating and dispensing the buffer twice. The samples were then vacuum-dried in a SpeedVac concentrator and re-dissolved in 20µL 2 mM TCEP (tris(2-carboxyethyl)phosphine) in water/ACN (98:2, v/v) for LC-MS/MS analysis.

LC-MS/MS analysis of label-free proteomes

Samples were analyzed by LC-MS/MS using an UltiMate 3000 RSLC nano HPLC (Dionex) inline connected to a Q-Exactive HF mass spectrometer (Thermo Fisher Scientific Inc.). Samples were separated on a 40 cm column packed in the needle (produced in-house, 75 µm I.D. × 400 mm, 1.9 µm beads C18 Reprosil-HD, Dr. Maisch) using a non-linear 150 min gradient of 2-56% solvent B’ (0.1% formic acid (FA) in water/ACN, 20/80 (v/v)) at a flow rate of 250 nL/min. This was followed by a 10 min wash reaching 99% solvent B’ and re-equilibration with solvent A (0.1% FA in water). Column temperature was kept constant at 50°C (CoControl 3.3.05, Sonation). The mass spectrometer was operated in data-dependent, positive ionization mode, automatically switching between MS and MS/MS acquisition for the 16 most abundant peaks in a given MS spectrum. The source voltage was set to 3.5 kV and the capillary temperature was 250°C. One MS1 scan (m/z 375-1500, AGC target 3E6 ions, maximum ion injection time of 45 ms) acquired at a resolution of 60,000 (at 200 m/z)
was followed by up to 16 tandem MS scans (resolution 15,000 at 200 m/z) of the most intense ions fulfilling predefined selection criteria (AGC target 1E5 ions, maximum ion injection time of 60 ms, isolation window of 1.5 m/z, fixed first mass of 145 m/z, spectrum data type: centroid, under fill ratio 2%, intensity threshold 1.3E4, exclusion of unassigned, singly charged precursors, peptide match preferred, exclude isotopes on, dynamic exclusion time of 12 s). The HCD collision energy was set to 32% Normalized Collision Energy and the polydimethylcyclosiloxane background ion at 445.12002 Da was used for internal calibration (lock mass).

Spectra identification was performed with MaxQuant (version 1.5.3.30) using the Andromeda search engine [52] with FDR set at 1% on peptide and protein level. Spectra were searched against the “aTIS database” (in-silico translated ORFs of protein-coding transcripts annotated in Ensembl 82 Sep 2015) or “custom database” (in-silico translated ORFs both annotated in Ensembl 82 Sep 2015 and predicted based on HCT116 ribo-seq data presented in this article). The mass tolerance for precursor and fragment ions was set to 4.5 and 20 ppm, respectively, during the main search. Methionine oxidation to methionine-sulfoxide was set as a fixed modification. Acetylation of protein N-termini was set as a variable modification. Trypsin/P was set as enzyme allowing for 2 missed cleavages. The match between runs function was enabled and proteins were quantified by the MaxLFQ algorithm integrated in the MaxQuant software [53]. Minimum of 2 ratio counts and only unique peptides were considered for protein quantification.

Data analysis of label-free shotgun proteomics

Further data analysis was performed with the Perseus software [54] (version 1.5.3.0) after uploading the protein groups file from MaxQuant. Proteins only identified by site and reverse database hits were removed as well as potential contaminants. To be able to perform GO (Gene Ontology), KEGG, Pfam, Corum and Keyword term enrichment, available annotations were uploaded from the Homo sapiens database (release date 2015-06-20). The replicate samples were grouped and the LFQ-intensities were log(2) transformed. Proteins with less than three (for HAP1 experiments) or two (for HCT116 experiment) valid values in at least one group were removed and missing values were imputed from a normal distribution around the detection limit (with 0.3 spread and 1.8 down-shift). A multiple-sample ANOVA test was applied with S0 parameter (corresponding to the minimal fold-change considered to be significant) set to 0.1 and p-value threshold of 0.01 (for HAP1 experiment) or 0.05 (for HCT116 experiment) enabling for an analysis of proteins with significantly different expression intensities between groups. Differentially expressed proteins were selected for subsequent Z-score normalization and k-means clustering. Median log2 LFQ values per condition were calculated for individual genes and 2D enrichment analysis [55] of annotation terms was performed using the Perseus software package with corrected p-value threshold of 0.02. The results of the “custom database” search of the HCT116 label-free shotgun proteomics data was visualized on the human
GRCh38 reference genome (Ensembl annotation bundle 82) as a BED track (Appendix File IV.1).

**Ribosome profiling**

Knockdown experiments were performed to obtain 4 x 10 cm² plates of si-Ctrl and 4 x 10 cm² plates of si-elf1 treated HCT116 cells. For ribosome profiling, 2 plates per condition were incubated with either 50 µM lactimidomycin (LTM) [56, 57] or 100 µg/ml cycloheximide (CHX) (Sigma, USA) in a total volume of 5 ml per plate for 30 min at 37°C. Subsequently, cells were harvested by trypsinization (Trypsin/EDTA solution, Thermo Fisher Scientific Inc.), collected by centrifugation (5 min at 300g), washed with ice-cold PBS with 100 µg/ml CHX added and recovered again by centrifugation. Cells were re-suspended in 1 ml ice-cold lysis buffer (10 mM Tris-HCl, pH 7.4, 5 mM MgCl₂, 100 mM KCl, 1% Triton X-100, 2 mM dithiothreitol (DTT), 100 µg/ml CHX, 1 × complete and EDTA-free protease inhibitor cocktail (Roche)) [58], lysed by incubation on ice for 10 min and subsequently passed through QIAshredder spin columns (Qiagen). The flow-through was clarified by centrifugation for 10 min at 16,000 × g and 4°C. The supernatant was recovered and protein concentration was measured. 600 µl of each lysate was subjected to RNase I (Thermo Fisher Scientific Inc.) digestion using 3000 U of enzyme for every 7 mg protein. The remaining lysate was snap-frozen and kept at -80°C for mRNA-sequencing. Digestion of polysomes proceeded for 45 min at room temperature. Subsequent steps were performed as described previously [59] with certain adjustments. Ribosome protected fragments (RPFs) of 26-34 nt were extracted from 22 µg of RNA. Ribosomal RNA derived contaminants were depleted from reverse transcribed samples using a custom set of 20 biotynylated deoxy-oligonucleotides. 100 pmol of mixed probes with 60ul MyOne Streptavidin C1 DynaBeads (Thermo Fisher Scientific Inc.) were used each time. The relative amount of probes was experimentally adjusted to the prevalence of the corresponding contaminants. Samples were amplified by PCR using compatible primers (Appendix Table IV.3). The resulting ribosome-profiling libraries were duplexed (samples treated with the same inhibitor, LTM or CHX, were joined) and sequenced on a NextSeq 500 instrument (Illumina) to yield 75 bp single-end reads.

**Ribosome profiling data analysis**

CHX and LTM data were analysed in parallel using the PROTEOFORMER pipeline [34]. The 3’ adapter sequences (Appendix Table IV.3) were removed using the fastx_clipper and reads between 26 and 34 nt in length were retained. Reads were initially mapped onto small nuclear RNA, tRNA and rRNA sequences to remove contaminant sequences. The remaining reads were then mapped onto the human GRCh38 reference genome (Ensembl annotation bundle 82) using STAR 2.4.0i allowing only unique mapping with a maximum of 2 mismatches. RPF alignments were assigned to a specific P-site nucleotide, as previously described [5, 34]. Separate BedGraph files (normalized for the total number of genome-mapped reads) were generated for the sense and antisense strand with CHX and LTM read alignments reduced to single-nucleotide genomic position (P-site). Additional Bedgraph files
were created with full-length CHX reads. The TIS calling algorithm was applied with default PROTEOFORMER settings, followed by ORF delineation, as described previously [5, 34]. Protein sequence database of in-silico translated Ensembl annotated CDS sequences (“aTIS database”) was generated from PROTEOFORMER output disregarding any detected single nucleotide polymorphisms (SNPs). CHX reads were counted in the annotated CDS and u(-o)ORF regions of protein-coding transcripts. To account for the accumulation of RPFs at start and stop codons ([21, 26, 60]), when counting reads we excluded the first and the last 15 nt in ORFs longer than 300 nt. For shorter ORFs though, we restricted the region relatively (5% of nucleotides were excluded at both sides). To account for differences in sequencing depth amongst CHX samples, read counts were normalized according to the total number of genome-mapped reads. The most probable protein-coding transcript per gene was selected based on the similarity between gene and transcript positional read coverage using a zero-inflated negative binomial model for count data (“pscl” R package). When experimental data was inconclusive and the algorithm failed (2% of genes), the preferred transcript was chosen according to the following order of importance: i) the longest CDS, ii) the most upstream aTIS, iii) the longest transcript. Subsequently, differential expression analysis was performed on the unique set of transcripts provided that the quantified ORFs met the minimum required number of CHX reads in both (si-eIF1 and si-Ctrl) conditions (at least 32 reads in CDS and 8 reads in u(-o)ORF, respectively). To generate the custom protein library (“custom database”) that contains both ribo-seq predicted and Ensembl annotated ORFs, ribo-seq data was mapped to the genome as described above, except that multiple mapping reads (up to 16 genomic loci) were allowed and SNP detection was enabled.

mRNA sequencing and data analysis

RNA was isolated from ribo-seq cell lysate with TRIzol reagent (Invitrogen, Thermo Fisher Scientific Inc.) according to manufacturer’s instructions. RNA quality was assessed by Agilent Bioanalyzer RNA 600 Nano Kit and RIN values above 9 were accepted. Random fragmentation, cDNA synthesis and library generation were performed according to TruSeq Stranded Total RNA Sample Preparation protocol (Illumina). Libraries were subjected to sequencing on a NextSeq 500 instrument (Illumina) to yield 75 bp single-end reads. The 3’ adapter sequences (Appendix Table IV.3) were trimmed using the fastx_clipper. Reads were pre-mapped onto small nuclear RNA, tRNA and rRNA. The remaining unmapped reads were then mapped onto the human GRCh38 reference genome (Ensembl annotation bundle 82) using STAR 2.4.0i allowing only unique mapping with a maximum of 2 mismatches. Reads were counted across annotated protein-coding transcripts and normalized for differences in sequencing depth between samples as described for ribo-seq data. Separate normalized BedGraph files were generated for the sense and antisense strand. From the set of selected transcripts (see ribosome-profiling data analysis section), mRNAs with less than 32 reads in all samples were excluded from further analysis.
Differential expression analysis at the transcript level

Differential expression analysis was performed according to Andreev et al. [7]. Only selected transcripts exceeding the minimal read count thresholds at both mRNA and CDS level were included in the analysis. In brief, normalized read counts across mRNA, CDS and uORF regions were used to calculate expression fold change between si-eIF1 and si-Ctrl samples. Subsequently, these ratios (fold changes) were log2 transformed and grouped in bins of 300 transcripts/ORFs according to the minimal read count value across samples. Each bin was individually Z-transformed and evaluated for differential expression. mRNAs, CDSs and uORFs found to be significantly regulated at a 1% significance level (absolute Z-score value ≥ 2.58) were selected. Subsequently, translation efficiencies were calculated separately for si-eIF1 and si-Ctrl samples by dividing the CDS or uORF read counts by their corresponding mRNA read counts. Next, a ratio of TE values was obtained, representing the fold change in TE CDS or TE uORF between si-eIF1 and si-Ctrl conditions. Finally, differential TE regulation between uORF and CDS of the same transcript was assessed by first calculating the ratio TE uORF/TE CDS in each condition and then obtaining the ratio between conditions. TE CDS, TE uORF and TE uORF/TE CDS fold changes were subjected to log2 and Z-score transformation as described above. To be able to perform GO (Gene Ontology), KEGG, Pfam, Corum and Keyword term enrichment, annotations were retrieved for all 10433 transcripts analyzed. 1D enrichment analysis [55] of annotation terms was performed using the Perseus software package with corrected p-value threshold of 0.02.

Calculation of context score

Sequences between -10 and +10 bp around aTIS called by the PROTEOFORMER pipeline were used to establish a scoring system enabling us to grade the entire nucleotide context with a single numeric score based on the preferred nucleotide frequency at particular positions, as described by Grzegorski et al. [42]. As such, we observed that the frequency of nucleotide occurrence from positions +6 to +5 deviates significantly from the neutral nucleotide frequency (1:1:1:1 A:T:G:C) [61] (Chi2 test, p=<0.001). Once the region of interest was specified, nucleotides were assigned a percentage frequency per position and the total score was calculated as a sum of these frequencies.

Conservation analysis

Metagenic plots were generated based on phastCons and phyloP scores over a 25-nt windows around the start site of the aTIS, uORF and uoORF groups. Averaged scores were also calculated from a control set of intergenic ORFs derived from the same pool of transcripts. Conservation information of the multiple alignments of 99 vertebrate genomes to the human genome was downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg38 subfolders phastCons100way and phyloP100way. We ensured that only 25-nt regions were included that completely reside in exonic regions of mRNA.
Isolation of total RNA and RT-qPCR

RNA isolation and quality assessment was performed as described above. cDNA synthesis was performed using 0.5 µg RNA, 0.5 µl PrimeScript RT Enzyme Mix I, 2.5 µM Oligo dT Primer and 5 µM random hexamers by 15 min incubation at 37°C followed by 5 sec at 85°C. qPCR primers (Appendix Table IV.3) were designed using QuantPrime software [62]. PCR efficiency over a series of cDNA dilutions and product specificity was assessed by qbase+ software (Biogazelle) and melting curve analysis (LightCycler 480 Software, Roche), respectively. mRNA expression levels of EIF1 and EIF1B were quantified in triplicate in relation to 3 reference genes (HPRT1, GAPDH and ACTB).

ATP measurement

HCT116 cells were seeded in DMEM medium supplemented with 10% FBS, 2 mM GlutaMAX, 50 units/ml penicillin and 50 µg/ml streptomycin (all Thermo Fisher Scientific Inc.) and cultivated in a humidified incubator at 37°C and 5% CO2. 5,000 HCT116 cells per well in a 96-well plate were transfected with 10nM si-RNA (si-eIF1 or si-Ctrl) and 0.75 µl HiPerFect (QIAGEN) using the fast forward procedure according to manufacturer’s instructions. Following 72 hours of knockdown, medium was replaced with either complete DMEM medium with 10% FBS or glucose-free DMEM with 10% FBS dialyzed using a 10 kDa cut-off membrane. Treatment continued for 6 hours and cellular ATP levels were measured using CellTiter-Glo® Luminescent Cell Viability Assay (Promega), according to manufacturer’s instructions.

Mitochondrial membrane potential measurement

si-RNA mediated knockdown was performed as described in the “Knockdown experiments in HCT116 cells” section of the Materials and Methods. After 72 hours post-transfection, cells were harvested by trypsinization (trypsin-EDTA solution, Gibco, Thermo Fisher Scientific), counted and dissolved in fresh medium at a concentration of 1 million cells/ml. 1 ml of cells was treated with either 0.5 µl of DMSO or 0.5 µl of 100mM CCCP in DMSO (mitochondrial membrane depolarization agent, Sigma). Cells were incubated for 15 min at 37°C and centrifuged for 3 min at 300xg. The JC-10 reagent from the Mitochondria Membrane Potential Kit (Sigma) was dissolved in the appropriate buffer at a final concentration of 0,5 % (v/v). Cells were suspended in 1 ml of JC-10 solution and incubated for 30 min in darkness at 37 °C, collected by centrifugation and suspended in 1 ml pre-warmed PBS. Cells were subjected to flow cytometry analysis. 20,000 events (cells) were measured in 488 and 633 nm channels with compensation for emission cross-over. Events were gated to separate cells with active mitochondria (red and green JC-10 emission) from cells with depolarized mitochondria (only green JC-10 emission).
Data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [63] partner repository with the dataset identifier PXD004980 (reviewer account username: reviewer62516@ebi.ac.uk and password: sVaRaznu).


Besides, a BED file of peptides identified in the “custom database” search of the HCT116 label-free shotgun proteomics data and mapped onto the human GRCh38 reference genome (Ensembl annotation bundle 82) was provided in Appendix File IV.1.

Acknowledgements

The authors thank Prof. Kris Gevaert for conceptual assistance, manuscript revision and financial support of this research (Research Foundation - Flanders (FWO-Vlaanderen), project number G.0440.10). D.F. is supported by a Ph.D. grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen). P.V.D. acknowledges support from the Research Foundation - Flanders (FWO-Vlaanderen), project number G.0269.13N. The authors thank José Van Der Heyden and Ali Adiby for assistance in experiments.

Author contributions

D.F. performed experiments, analyzed data, drafted and revised the manuscript. S.V. and E.N. analyzed data and revised the manuscript. V.J. performed experiments. G.M. drafted and revised the manuscript. P.V.D conceived the study, analyzed data, drafted and revised the manuscript.
References


GENERAL DISCUSSION
GENERAL DISCUSSION

Although the complete human reference genome sequence is known for over a decade, functional annotation of genomic DNA regions remains challenging, especially in the case of protein-coding genes which may generate multiple protein versions (proteoforms) from a single locus [1]. Until recently, one of the most widely acknowledged sources of protein diversification in mammalian genomes was alternative splicing [2, 3]. However, positional proteomics has marked a new chapter in the proteogenomics era by providing novel insights into translation initiation. The unexpected discovery that numerous alternative protein N-termini are formed co-translationally [4-6] was supported by evidence that such novel proteoforms undergo N-terminal modifications linked to the initiation phase of protein synthesis occurring at the ribosome. These observations highlighted the incompleteness of current protein start site annotation and sparked our interest in exploring the real N-terminal proteome complexity. Therefore, we initiated N-terminomics studies in 3 mouse and 7 human cell models (including cell lines and primary cells) to systematically assess the occurrence of alternative N-terminal proteoforms in these organisms. Based on positional proteomics data, we estimated that approximately 20% of unique protein N-termini result from the expression of truncated or extended proteoforms, likely as a consequence of alternative translation initiation or in some cases also due to alternative splicing [7]. Overall, we reported on 1,254 human and 484 mouse alternative N-termini (i.e., N-termini previously not annotated in databases), with the grand majority pointing to N-terminal truncations (98% and 96% in human and mouse datasets, respectively). Of note is that these datasets confirmed several previously reported truncated proteins resulting from downstream translation initiation events [8-12]. Metadata for the genuine translational origin of the alternative delineated TIS landscape was provided by including various sources of proteome-matching data on transcript annotation, ribosome profiling, start codon context, PhastCons sequence conservation analysis and subcellular localization prediction (see below).

Our discoveries of N-terminal proteoforms (Fig. 1) coincided with major developments in the genome-wide mapping of translation (initiation) event by means of ribosome profiling or ribo-seq. By deep sequencing of ribosome-protected mRNA fragments, ribo-seq allowed to study translation initiation with sub-codon resolution [13, 14]. This presented a unique opportunity to explore the relationship between translation events captured at the mRNA and protein levels. Overall, ribo-seq data obtained in a variety of mammalian cells revealed that only half of the TIS captured genome-wide made use of AUG as the translation initiation codon, although more recent strategies for ORF delineation pointed to 92% of AUG start codons in human [15]. Regardless of the TIS calling algorithms, distinct codon usage was shown for TIS residing in 5’ mRNA leader sequences (upstream TIS or uTIS) as opposed to the annotated CDS. uTIS utilized mostly non-AUG codons, while downstream TIS or dTIS predominantly made use of AUG codons [14]. Most dTIS occurred in-frame, presumably leading to protein N-terminal truncations [14]. In fact, increased frequency of downstream
in-frame AUG codons in the 5’ proximal part of annotated coding sequences was initially reported by Kochetov et al. [16] purely based on the in silico analysis of human mRNA sequences. Our N-terminomics findings were in line with these observations, as the vast majority of truncated proteoforms we identified utilized AUG start sites, while from 37 N-termini pointing to near-cognate translation initiation sites, nearly all gave rise to N-terminal protein extensions [7].

Conformity between ribo-seq and N-terminomics extended further as both techniques pointed to a major contribution of leaky scanning in selecting alternative TIS. Annotated TIS falling in strong Kozak consensus sequences were frequent in genes with no detectable dTIS, while weak nucleotide context of annotated TIS frequently coincided with the occurrence of downstream initiation [7, 14], overall suggesting that leakiness of the first AUG codon determines the translation initiation rate at the downstream AUGs.

As previously shown to be the case for selected proteoforms [17-19], N-terminal extension and truncations may lead to the gain or the loss of subcellular targeting signals, putting forward translation initiation as an attractive mechanism of regulating protein localization. By predicting the localization of full-length and truncated N-terminal proteoforms using TargetP [20], truncated proteoforms were found less likely to contain N-terminal presequences such as a mitochondrial targeting or a secretory pathway signal sequence. Nevertheless, we discourage from drawing general conclusions, as trafficking of proteoforms to different cellular compartments ought to be individually investigated. For example, the p43 component of the mammalian multi-synthetase complex loses its mitochondrial targeting sequence when translation initiates at a downstream AUG [21]. Alternatively however, opposed to the full length protein version, a truncated proteoform of Flap endonuclease 1 exposes a cryptic mitochondrial targeting signal [22]. Our predictions as well as other lines of evidence [23] indicate that alternative translation initiation can give rise to iso-functional, though localization-specific N-terminal protein variants.

Besides steering protein localization, TIS selection can regulate protein stability and protein expression levels. Such intriguing properties were reported for several truncated proteoforms identified in our study [24-26]. Downstream initiation may produce degradation-resistant proteoforms, as is the case for truncated IkBα lacking phosphorylation sites essential for targeted proteasomal degradation. In consequence, this overly-stable proteoform may act as a dominant-negative regulator of NF-κB activity. Interestingly, a low level of truncated proteoform synthesis may also have undesired effects. A nonsense mutation in IKKγ/NEMO found in patients suffering from anhydrotic ectodermal dysplasia prevents full-length protein synthesis, but still permits the expression of a truncated and functional variant. Unfortunately, this protein variant is produced in insufficient amounts for the development of protective immune responses in these patients. According to our data, alternative N-terminal proteoforms tend to display decreased spectral counts compared to their full-length counterparts, generally suggesting either lower translation efficiency or...
Figure 1 – A graphical representation summarizing the major findings described in this thesis studying alternative translation initiation mechanisms and their genome-wide effects on N-terminal (Nt-) proteoform expression, stability, localisation and regulation of translation initiation.
possibly increased degradation ultimately leading to lower concentration of protein products expressed from such alternative initiation codons. Yet again, this trend is not universal, as a more balanced expression can be observed from closely spaced TIS that give rise to N-terminal proteoforms differing less than 6 amino acids. With no easily discernible difference in molecular weight, such protein products can easily be overlooked by conventional detection methods such as Western blotting, underlining the unique capabilities of N-terminal proteomics approaches.

Nevertheless, there are certain inherent limitations to N-terminomics or more specifically N-terminal COFRADIC and mass spectrometry in general, overall restricting the range of (theoretically) detectable N-termini. First, the coverage of N-termini with cleavage sites of a single protease is known to be incomplete, while combining results obtained with different proteome-digestive enzymes assures a more comprehensive detection of N-termini [27]. Second, physical and chemical properties of N-terminal peptides may restrict their efficient recovery during extraction, separation or ionization. N-terminal peptides are frequently the only unique peptides that allow differentiation between proteoforms of the same gene, which makes them per definition “one-hit wonders”. Additionally, quantification of N-terminal proteoform expression using only spectral counts as abundance estimates is subject to considerable errors, inherent to MS2 data acquisition [28, 29], and may be less reliable especially for low abundant proteins. Although our dataset is enriched in proteoforms likely expressed at low(er) levels and thus characterized by low spectral count numbers, we never directly compared abundances of individual proteins, but rather of protein groups. In this case, the uncertainty of measurement is diminished by the number of observations. Additionally, proteoform quantification is generally reproducible across samples and cell lines. Despite these drawbacks, N-terminal proteomics remains an excellent identification tool for proteoforms, while associated spectral counts, treated with careful consideration, provide quantitative estimates that allow to draw conclusions regarding proteoform abundance.

Beside (N-terminal) proteomics approaches, proteoform synthesis rates can be measured using ribosome profiling. Relying on lactimidomycin to detect translation initiation events in human cells, we observed that dTIS had significantly lower ribosomal signal at the start codon compared to annotated TIS of the same gene [30], in line with abundance estimates gathered at the N-terminal peptide level. Reduced ribosome footprint counts at downstream start sites are perfectly explained by the leaky scanning model. According to the model, the efficiency of upstream initiation regulates the number of leaky ribosomes that reach the downstream TIS. In consequence, the number of ribosomes detected on a particular TIS is not a direct measure of TIS strength, but rather greatly depends on the presence of preceding start sites [31]. Therefore, we may expect to observe large accumulation of ribosomes at weak TIS (such as the non-AUG codons) upstream of strong TIS, but it’s less probable to detect such weak start sites downstream. The quantitative value of translation initiation profiling with lactimidomycin and harringtonine was explored by Michel et al. [31],
who proposed an elegant model which explained why non-AUG TIS are commonly found in mRNA 5’leaders but, despite high ribosome occupancy, are generally inefficiently recognized as start sites, likely resulting in low synthesis rates of thereby encoded proteoforms, supported by limited mass spectrometry evidence for N-terminally extended proteoforms and uORF-encoded peptides [7]. To further improve on the quantitative aspect of translation initiation profiling, Gao et al. proposed a new experimental strategy, called QTI-seq [32]. By administering lactimidomycin shortly after cell lysis, they preserved the original ribosomal occupancy at the start codon and prevented new rounds of translation initiation during prolonged pre-incubation with the drug (especially in the 5’ part of mRNA). This method provided hitherto unavailable, experimental translation initiation rate estimates upon dynamically changing (extra)cellular conditions. Such new strategies for ribosome profiling experimental design and data analysis are expected to provide improved approximations of proteoform synthesis rates, especially with regard to truncated proteoforms frequently captured by mass spectrometry-driven approaches, yet difficult to quantify by ribosome profiling due to the overlap with other ORFs.

With a substantial inventory of N-terminally truncated proteoforms at hand [7], we set out to characterize the stability of these novel proteome components (Fig. 1). Cellular protein stability has the potential to reflect a protein’s individual properties and biological functions [33, 34]. In eukaryotic cells, a wide variety of proteins with roles in cell cycle progression, signal transduction and metabolic regulation is quickly degraded by the proteasome [35], whereas in contrast, some constituents of abundant protein complexes (i.e., the ribosome, spliceosome) can be extremely stable [33, 34]. Large-scale studies of proteome dynamics were made possible by applying stable isotope labeling in combination with mass spectrometry. In previous studies, researchers either made use of non-synchronised, proliferating cells [34, 36, 37], while others employed growth-arrest conditions [33]. Cells were then harvested at different time points after the growth medium swap, ranging from only two [33] to up to eight time points [34]. Since the frequency of measurements increases the coverage of the dynamic range of the protein levels studied, we decided to employ the approach of Boisvert et al. [34], which is considered to provide more experimental data in proportion to computational approximations. Our experimental setup enabled the measurement of proteoform half-lives as well as protein turnover, which reflects the overall protein synthesis and degradation rates [34].

At the time that our project was initiated, only limited evidence indicated that alternative translation initiation could indeed produce protein variants that differ in stability from their database-annotated counterparts [24, 38]. Alternative translation initiation of the opioid receptor previously showed that an N-terminal extension rich in lysine residues and prone to ubiquitination, contributes to the high degradation rate of this proteoform [38]. As suggested, the main reason for generating this N-terminally extended variant was to decrease translation efficiency of the canonical protein. Such rapidly degraded N-terminal proteoforms may be prematurely dismissed as potential translation “by-products”. Instead,
we challenged the common perception by considering that N-terminal isoforms may possibly have equal or superior stability as compared to database-annotated isoform.

To address these hypotheses, we performed the first large-scale study on protein stability that discriminates between protein products of a single gene that differ in their N-terminal parts. We combined our proteome analysis with ribosome profiling to confirm that N-terminal proteoforms are translated from alternative translation start sites within the same gene. We concluded that N-terminal proteoforms raised by alternative translation initiation are typically less abundant than canonical proteoforms, but may display altered stabilities. Interestingly, truncated proteoforms often displayed a lower translation initiation signal at the start codon and reduced proteome abundance. These finding were consistent with a lower frequency of downstream translation initiation assumed from the leaky scanning model of translation initiation also corroborated by previous reports [14]. Occasionally, truncated proteoforms were exclusively expressed in the absence of their full-length counterparts both at the translatome and the proteome level.

N-termini were previously reported to hold intrinsic information on the so-called N-degrons [39]. N-end rule mediated degradation was shown to affect unmodified iMet-starting N-termini followed by a hydrophobic amino acid [39] and neo-N-termini generated upon proteolytic cleavage [39, 40]. Additionally, a specific branch of this pathway was implicated in targeted degradation of proteins with Nt-acetylated residues [41]. Overall, we did not find a direct link between our data and the N-end rule pathway which proposes a relation between the nature of protein N-terminal residues and protein degradation rates. More specifically, we discovered that the removal of the initiator methionine by methionine aminopeptidases generally reduced the stability of processed proteoforms, while N-terminal acetylation susceptibility did not seem to influence protein turnover rates. Discrepancies between mass spectrometric studies of protein turnover and the N-end rule were previously reported by Doherty et al. [36]. However, while well described in yeast, the N-end rule is incompletely studied in human cells, and additional efforts are needed to characterize some of the executioner enzymes and their targets in higher eukaryotes in general. Since the N-end rule pathway holds an exceptionally broad range of potential substrates, additional mechanisms have been postulated to repress undesirable protein degradation. For example, N-degrons may be shielded by proper folding of the mature protein or by its integration into macromolecular complexes [42]. As there are many ways for proteins to evade degradation mediated by N-degrons, this pathway may constitute a clearance mechanism for N-terminally misfolded or non-complexed proteins [43, 44]. Given that protein N-termini are often more flexible and disordered [45], the conditional silencing of N-degrons may involve other types of regulation besides a physical protection of the (acetylated) N-terminal residues.

In contrast to other proteome studies that did not distinguish between N-terminal proteoforms, we combined SILAC pulse-labeling [34] with N-terminal proteomics [46] to
study the stability and turnover rates of proteoforms with heterogeneous N-termini. Another advantage of our approach was the fact that we relied on the complementary use of positional proteomics and in-house generated ribosome profiling data to aid in the characterization of the human N-terminome. So far we often relied on existing ribo-seq datasets [6, 7]. However, to fully explore the possibilities coming from coordinated OMICS analyses, we optimized ribosome profiling experiments in our laboratory. Moreover, we used our knowledge on co-translational modifications occurring on protein N-termini to unambiguously assign an identified N-termini as a proxy of translation initiation.

We and others have shown that bioinformatics-assisted integration of ribosome profiling data to the protein sequence search space facilitates the discovery of novel proteoforms, enabling a more comprehensive (re-)annotation of the translation initiation landscape [6, 7, 47, 48]. Here, we applied ribosome profiling next to positional and whole proteome analysis in a colorectal cancer cell line (i.e., HCT116 cells) to identify features of human proteins that cannot simply be identified using standard analysis pipelines or annotations (Fig. 1). We have built upon the successful marriage of the two techniques introduced in our lab [6]. Here, we have extended our proteogenomic strategy by taking into account sample-specific SNP variants, next to alternative translation products, overall leading to improved identification scores of 69 proteins next to 22 novel proteoforms identified in the shotgun proteomics experiment [49]. Abundance estimates derived from ribo-seq and shotgun proteomics data correlated well, especially for sufficiently covered ORFs. Interestingly, poorly expressed proteins and ORFs were often predicted to be unstable (according to the instability index of ExPASy ProtParam tool [50]). Furthermore, 18 new N-terminal proteoforms (including 5’ extensions, N-terminal truncations and a uORF) were identified uniquely in the ribo-seq enriched custom database by N-terminal COFRADIC. Nonetheless, the majority of N-terminal peptides linked to alternative translation initiation pointed to downstream AUG usage (204 proteoforms), N-termini which in principle can be identified in conventional protein databases when making use of semi-specific enzyme settings, or by using non-redundant peptide search space supplemented with predicted truncated proteoform-specific peptides [51].

Proteomics evidence of upstream (near-cognate) start site usage appeared underrepresented, compared to the expected frequency as estimated by ribosome profiling, meaning that besides N-terminal extensions the products of uORFs and out-of-frame translation products remained largely undetected. Many of these hypothetical cases are expected to result in the production of small peptides and proteins that have routinely been excluded from database sequence annotations. In contrast to the frequent occurrence of short ORFs in ribosome profiling data [13, 14], we were able to identify only one N-terminal peptide of an upstream overlapping ORF in the PIDD gene. This limited evidence for uORF protein products could be attributed to several factors, such as a bias towards upstream (near-) cognate start site identification from ribosome profiling data [52] (see also Results
Chapter 4) or the rapid degradation, small size and possibly low abundance of uORF translation products.

Although the sensitivity of ribosome profiling to detect TIS sites and delineate reading frames remains unprecedented, the bioinformatics approaches used for TIS assignment may be ineffective in some cases. In fact, custom databases do not saturate the expressed proteome and we supplement them with Swiss-Prot entries to comprehensively capture the proteome complexity. This can be partially explained by the incompleteness of Ensembl genome annotation used as basis for ribosome footprint mapping or incorrect ORF delineation and in-silico translation, especially in the case of mitochondrial genetic code which utilizes non-standard translation start and stop codons. Moreover, ribo-seq provides an instant snapshot of translation, while certain stable proteins may not undergo synthesis at the time of sampling. This, as well as the recent finding that ribosome occupancy may not always imply effective translation initiation [31] or active elongation [53], necessitates the demand for proteomics evidence to identify products of translation.

We have compared the list of protein extensions and truncations identified using a customized ribo-seq derived protein database with previous studies that mapped translation initiation in cultured human cells (lactimidomycin treated HEK293 cells [14]; harringtonine and lactimidomycin treated human foreskin fibroblast cell cultures (HFFs) [47] and puromycin treated THP-1 cells [54]). We used the GWIPS-viz browser [55] to look for evidence of translation initiation at these alternative start sites confirmed by our proteomics and ribosome profiling experiments. Despite certain shortcomings (different translation inhibitors and TIS calling algorithms used), we observed a complete overlap of TIS selection between the 4 human cell lines under study for 11 out of 29 proteoforms. However in many instances, cell type specific TIS selection could be observed. In line with this finding, cell type or cellular context dependence of TIS selection and thus translation regulation was previously observed for FGF2 [56], K2P2.1 [57] and BNIP-2 [58].

Experience gained in the course of this project facilitated the development of PROTEOFORMER [59], an automated bioinformatics pipeline for conversion of ribo-seq data into a custom protein sequence search space that incorporates both sequence variation information and TIS prediction.

Translation initiation site recognition depends on the fidelity of codon-anticodon base pairing. Initiation is considered to be the most important rate limiting step of protein synthesis at the ribosome. Unsurprisingly perhaps, translational control is exerted mainly during the initiation process and often engages uORFs situated in translationally active mRNA 5’ leaders [60]. eIF1 and 1A are key regulators that modulate ribosome conformational changes needed for successful translation initiation. eIF1 and 1A both impose an open conformation of the 43S scanning complex [61]. Upon establishing a stable interaction between the initiation site and the Met-tRNA, the ribosome undergoes spatial remodeling, which is antagonized by eIF1. This factor probably prevents incorrect initiation.
by discriminating against non-AUG codons and a suboptimal nucleotide initiation context, and can dissociate aberrantly assembled complexes [62]. elf1 mutant studies in yeast revealed that premature elf1 release near the P-site decreased the stringency of start codon selection [63]. Dissociation of elf1 probably leads to partial closure of the ribosomal conformation followed by elf5-mediated hydrolysis of GTP bound to elf2. Genetic suppressor studies in yeast linked both elf2 and elf5 to the accuracy of initiation site selection, as several mutations in these genes significantly promoted initiation at near-cognate, non-AUG codons [64]. Other studies reported a role of the elf3-elf1 interaction interface in AUG recognition in S. cerevisiae [65]. Further, similar endeavors were undertaken to prove a role of elf4G in non-AUG discrimination, mediated by its interactions with elf1 and 5 [66].

This convincing line of evidence situates elf1 at the forefront of translation initiation factors involved in steering start codon selection. We therefore decided to extend our knowledge in mapping translational events at steady state levels to investigate dynamic systems. For this purpose, we used ribosome profiling, mRNA sequencing and label-free shotgun proteomics to assess the impact of elf1 on start codon selection and uORF usage in the already well-characterized cell line HCT116 (Fig. 1).

While non-AUG translation initiation has been experimentally verified at the single transcript level, its genome-wide prevalence, the impact on steady-state protein levels and the in vivo regulation governing the recognition of such near-cognate translation initiation codons remained unclear. Our results indicate that elf1 depletion significantly reduces the incidence for translation initiation at AUG start sites and affects the stringency of nucleotide context recognition. In consequence, the expression of 245 genes was significantly affected by altered start codon selection. Translation efficiency of uORFs seemed to be especially perturbed, affecting the synthesis rate of their downstream canonical proteins by mechanisms such as leaky scanning or reinitiation. Our findings are well in line with previous reports where elf1 over-expression reduced initiation at TIS residing in a suboptimal KOZAK context as well as initiation at near-cognate codons [67].

Protein synthesis, as one of the most energy-consuming processes in living cells, is frequently perturbed in response to various stress conditions [68-70]. Perhaps unsurprisingly, translational control programme elicited by elf1 resulted in decreased translation efficiency and lower expression of genes involved in glycolysis/gluconeogenesis and TCA cycle, accompanied by decreased cellular ATP levels and decreased mitochondrial activity. These observations hinted to a more general impairment of energy metabolism induced by elf1 deficiency. Translation machinery itself was also placed under regulatory pressure, as indicated by the upregulation of genes implicated in ribosomal biogenesis, enhanced aminoacyl tRNA synthesis and amino acid transport observed at protein and transcript levels. Additionally, we observed concomitant change in expression of other elfs, which seemed to be affected mostly at the transcription level, suggesting that secondary
effects of eIF1 knockdown attempt to restore the stringency of start codon selection, for example by downregulating eIF5 [71]. Upregulation of eIF1B, a paralog gene sharing high sequence similarity to eIF1, suggests that further compensatory effects might take place, potentially accounting for some of the differences observed between the knockdown and the knockout models studied.

Next to several uORFs and 5’ extensions identified with the use of customized sequence library, we captured many peptides containing non-synonymous mutations. In most cases these peptides were associated with canonical proteins, however some of these peptide identifications were likely miss assigned to pseudogene expression. Pseudogenes generally share high sequence similarity to certain protein-coding genes, though they are exempted from purifying/negative selection. During the mapping of ribosomal footprints containing SNPs, search engines tend to preferably map reads to pseudogenes (with a 100% identity) instead of protein-coding genes, where the same footprint may map with a mismatch. This truly complicates protein inference and underlines that peptide identifications pointing to novel proteoforms, perhaps especially in the case of pseudogene expression, need thorough inspection. Sample-specific detection of genetic variation is still not a standard task of ribo-seq analysis pipelines, as such pseudogene expression [48] might easily be overestimated.

With broad accessibility and decreasing costs of high-throughput sequencing, ribosome profiling, (full-length) mRNA sequencing and SNP genotyping may become widely used to complement MS-based protein and peptide identification.
References:


BROAD RELEVANCE AND FUTURE PERSPECTIVES
Our laboratory has made a substantial contribution to the annotation of N-terminal proteoforms in yeast, fruit fly, mouse and human [1-4]. By the extensive mapping of the translation initiation landscape, the increased complexity of proteomes versus genomes and transcriptomes could in part be explained by the process of translation initiation, adding another layer of complexity. We anticipate that our results may be extrapolated to other model organisms as proteoforms arisen from alternative translation initiation are generally well conserved among eukaryotic orthologous genes [3, 5, 6].

With our research we hope to augment the current understanding of the proteome complexity and improve the human genome annotation by considering alternative translation start sites. We and others have demonstrated that ribosome profiling readout of translation enables the exploration of protein-coding potential outside of annotated open reading frames in the genome. However, with the constant progress in ribosome profiling data analysis, we were made aware that the sole association of mRNA with the ribosome is not necessarily indicative of translation [7]. Moreover, translation of certain non-canonical ORFs might play a regulatory role with no direct (functional) implications of the resulting peptides. Therefore, our future efforts will be dedicated to increase the stringency of translated ORF prediction from ribosome profiling data. Recently, new approaches were proposed to distinguish merely ribosome-associated from actively translated mRNA regions. Confidence of short ORF identification can be increased by exploring their sequence signatures (conservation, homology, mutation rates) and close inspection of ribosome profiles (ribosome footprint lengths and triplet periodicity) [8-11]. In this aspect, proteomics is increasingly recognized as a validation tool for reading frames delineated from the translatome [4, 11]. Despite the omnipresence of ribosome footprints in so-called "untranslated" (UTR) mRNA regions, long non-coding RNAs and pseudogenes, few of such events are supported by matching peptides. Optimized peptide and short protein enrichment will be required to boost the identification of short ORF translation products or, alternatively, disproof their coding potential.

The association of functional features to largely unexplored genomic regions may be especially important to tackle the problem of missing heritability [12, 13]. A large proportion of disease or trait associated variants (SNPs) maps to "non-coding" regions of the human genome [13]. Lacking the appropriate annotation, the relevance of such variants for the phenotype is difficult to assess. However, an increasing line of evidence demonstrates that mutations in mRNA 5' leaders are prevalent and may result in perturbed translational control over gene expression, deciphering disease etiology and heritability [14, 15]. It seems that polymorphisms found in uORFs may be both disease causing mutations as well as epistatic modulators that affect the predisposition to disease or severity of the phenotype.
N-terminal proteoforms are generally overlooked, but may explain common observations in biomedical research such as the recognition of numerous proteoforms by antibodies. In consequence, rather than looking at one specific gene or transcript product, many antibodies and, by extension, ELISA tests and immunofluorescence studies, may profile several proteoforms or alternatively miss out on specific protein variants. Our innovative project can serve as a base for the development of new protein expression approaches and selective antibody generation. So far, the Human Protein Atlas consortium has dedicated significant efforts to systematically generate splice-isoform specific antibodies. We have shown that, next to alternative splicing, alternative translation initiation further augments proteoform diversity, the study of which would benefit from the development of selective antisera. Our work might also have important implications for the rapidly developing field of genome editing. A gene knockout may be achieved by introducing a point mutation in its DNA sequence. However, this genetic manipulation may induce undesired outcome, such as the formation of a truncated proteoform with an altered stability. In addition, knocking out one gene might result in the disappearance of multiple gene products with possibly different localizations and functions, as is the case for certain N-terminal proteoforms, warranting caution. As such, our data may aid in the rational design of genome editing efforts minimizing the risk of such complications, next to providing novel routes for investigating previously unannotated genomic elements. In collaboration with Horizon Discovery, we have elaborated on this topic in a webinar https://www.horizondiscovery.com/resources/webinars/alternative-proteoforms-a-new-layer-of-proteome-complexity.

N-terminally truncated and extended proteoforms remain poorly characterized. Results covered by this PhD thesis should be of general relevance to the whole scientific community and can contribute to the common awareness of N-terminal proteoform expression by implementation to publicly available databases. Our research is also the first attempt to globally characterize N-terminal proteoform stability. Indeed, we demonstrated that N-terminal proteoform expression may have important functional implications for the proteome, by generating functionally related proteins with different cellular turnover rates [16]. However, many other biological aspects of N-terminal proteoform expression are still insufficiently studied. The scarce available literature indicates, that such proteoforms may be differentially regulated, localized [17, 18] and display altered functionalities [19-21]. Moreover, selection between alternative translation start sites is regulated under conditions of cellular stress [22] and may change the repertoire of N-terminal proteoforms expressed in a developmental and tissue-specific manner [19, 21]. With the avenue of automated high-throughput screening methods in for example transgenic zebrafish [23], genome-wide assessment of (N-terminal) proteoform expression throughout organs and developmental stages is now within reach. Such studies will surely shed light on proteoform functionalities and allow for prioritization of most interesting examples for further, proteoform-oriented assays.
Studies of the regulatory mechanisms underlying translation may reveal novel patterns in TIS recognition, leading to improved start site prediction algorithms [24, 25] that are much needed for interpreting genomic data. This is especially important for the proper prediction of N-terminal protein extensions and (regulatory) uORFs that are frequently initiated at non-AUG start sites [3-5, 26-28]. Translational control driven by eIF1 introduced significant alterations in start codon (context) usage, leading to the discovery of numerous uORFs that potentially regulate the expression of their downstream canonical protein-coding sequences. Further, our results highlighted the interconnectivity of the translation machinery and pointed to potential, but yet unexplored roles for eIF1B (an eIF1 paralog gene). Our future efforts will be dedicated to increase the understanding of translation regulation and proteoform expression. Translation may be perturbed in response to a variety of stress signals to which the cellular response is orchestrated by eIFs [29-32]. There are however substantial gaps in our knowledge regarding the functional potential of numerous eIFs and their paralog genes, for example eIF1B. Another interesting case is eIF4E2, a paralog of the conventional cap-binding eIF4E which was reported to potentially steer translation under hypoxic conditions observed deep within tumors [33]. Further elucidation of eIF4E2 target genes holds promise to indicate novel targets of therapeutic intervention that may reduce angiogenesis and cancer cell survival.

Finally, we aim at characterizing the (differential) localization of proteoforms across subcellular compartments. These N-terminally different proteins may show a loss of targeting signal, as well as acquire novel functions. Indeed, several proteins were already reported to exist in differential subcellular localizations due to alternative translation initiation [17, 34, 35]. Besides displaying iso-functional, though compartment-specific enzymatic activity [34], several of such protein forms showed novel and unique properties, such as transcription factor activity [35]. Our future research will aim at putting alternative TIS selection and its resulting proteoforms into a spatial cellular context, which is necessary to understand their biological role.

Efforts undertaken in the framework of this doctoral thesis lay grounds for improved characterization of functional translation products. As such, we foresee that deep integration of complementary OMICS approaches will ultimately result in a comprehensive catalogue of translation initiation events across various genomes.
BROAD RELEVANCE AND FUTURE PERSPECTIVES

References:


SUMMARY

Proteins are responsible for most biological activities of living cells and thus are considered the functional entities of genetic information. Protein biosynthesis is carried out by ribosomes which decode mRNA to amino acid sequences in a process called translation. Despite its fundamental importance - acknowledged by the Nobel Prize in Chemistry awarded in 2009 to Venkatraman Ramakrishnan, Thomas A. Steitz and Ada E. Yonath for structural and functional studies of ribosomes - the mechanism of translation is still not entirely understood. The initiation of protein synthesis is presumed to occur at the 5' proximal AUG codon encountered by the scanning ribosome, albeit significant deviations from this canonical model were described. Alternative translation initiation mechanisms such as leaky scanning and re-initiation allow for the selection between different start sites and/or reading frames within a single mRNA transcript, ultimately leading to the expression of multiple protein products or proteoforms. By allowing for reprogrammed translation, these mechanisms contribute to the overall proteome diversity and allow gene regulation. So far, individual N-terminal proteoforms harbouring different N-terminal extensions or truncations were described to display altered localisation, stability or functionalities, demonstrating the biological importance of alternative translation initiation. However, the exact scope and regulatory role of this process remains to be investigated on a genome-wide scale.

In this respect, technological advances in next generation sequencing and state-of-the-art mass spectrometry became indispensable to match the complexity of transcriptomes, translatomes and proteomes. Recently, ribosome profiling, based on deep sequencing of ribosome protected mRNA fragments, revealed a highly underestimated occurrence of alternative translation initiation sites (TIS) and widespread selection of non-AUG start codons throughout various genomes. Additionally, N-terminal COFRADIC, a technology developed in our laboratory, provided unprecedented precision in the assignment of protein N-termini, showing that in higher eukaryotes around 20% of all identified N-terminal peptides point to alternative translation initiation sites including non-AUG start codons, to incorrect annotation of translation start site or to alternative splicing.

With these technologies at hand, we decided to elucidate the true landscape of N-terminal proteoforms in mouse and human. Here, in a pioneering study we reported on more than 1,700 unique alternative protein N-termini identified at the proteome level in human and murine cellular proteomes. Customized databases, created using the translation initiation maps obtained from ribosome profiling data, additionally demonstrated the use of near-cognate codons as translation start sites decoded to initiator methionine, and the existence of N-terminal extended protein variants at the proteome level. Various newly identified TIS were confirmed by mutagenesis, and meta-analyses demonstrated that downstream TIS (dTIS) reside in strong Kozak-like motifs and are conserved among eukaryotes, hinting to a possible biological impact. Finally, we predicted that the usage of dTIS may result in N-
terminal truncated proteoforms with altered subcellular localization patterns, providing an additional regulatory mechanism for functional diversification of the proteome.

Next, we performed a proteome-wide study on protein stability using positional proteomics and ribosome profiling to distinguish between N-terminal proteoforms of individual genes. By combining pulsed SILAC with N-terminal COFRADIC we monitored the stability of 1,941 human N-terminal proteoforms, including 147 N-terminal proteoform pairs that originate from alternative translation initiation, alternative splicing or incomplete processing of the initiator methionine. N terminally truncated proteoforms were typically less abundant than canonical proteoforms and often displayed altered stabilities, likely attributed to individual protein characteristics, including intrinsic disorder, but independent of N-terminal amino acid identity or truncation length. We discovered that the removal of initiator methionine by methionine aminopeptidases reduced the stability of processed proteoforms, while susceptibility for N-terminal acetylation did not seem to influence protein turnover rates. Taken together, our findings revealed differences in protein stability between N-terminal proteoforms and pointed to a role for alternative translation initiation and co-translational initiator methionine removal in the overall regulation of proteome homeostasis.

To further explore the advantages of deep OMICS data integration for peptide identification, we performed an integrative proteomics and ribosome profiling study using a human cancer cell line as a model system. Here, ribosome profiling was used to delineate (alternative) translational reading-frames and map single nucleotide polymorphisms in a sample-specific manner. By including evidence of genetic variation and alternative translation in a customized database, this strategy improved the identification score of 69 proteins and allowed for the identification of 22 new proteins in the shotgun proteomics experiment. Additionally, we observed an improved correlation between the quantitative measures of ribo-seq and shotgun proteomics as compared to many previously reported relationships between mRNA and protein abundance. Furthermore, we discovered 208 truncated proteoforms, 9 N-terminal protein extensions and obtained N-terminal peptide evidence of the expression of one upstream overlapping ORF.

Despite the scarce peptide evidence for upstream ORF (uORF) expression, ribosome profiling data suggested that uORFs frequently populate 5’ leader sequences of protein coding genes and are more widespread than initially anticipated. By repressing or activating canonical protein translation, these uORFs may likely serve their biological role regardless of the corresponding peptides produced. Eukaryotic translation initiation factors seem to be key regulators of uORF activity that link cellular metabolism and external stimuli to elicit specific translational control programmes. In our work, we combined proteomics with ribosome profiling and mRNA sequencing to identify the biological targets of translation control triggered by the eukaryotic translation initiation factor 1 (eIF1), a protein implicated in the stringency of start codon selection. We quantified expression changes of over 4,000 proteins and 10,000 actively translated transcripts, leading to the identification of 245 genes
undergoing translational control mediated by upstream open reading frames (uORFs) upon eIF1 deprivation. The stringency of start codon selection and preference for optimal nucleotide context were largely diminished leading to translational regulation of uORFs with suboptimal start sites (non-AUG, poor context). Affected genes were implicated in energy production and sensing of metabolic stress. Interestingly, knockdown of eIF1 elicited a synergic response from eIF5 and eIF1B, highlighting the interconnectivity of the translational machinery.

Our results demonstrate that ribosome profiling and (positional) proteomics provide highly complementary data that give valuable novel insights to alternative translation initiation at a genome-wide scale. We believe that comprehensive discovery of alternative proteoforms, studies on their abundance, stability and regulation are essential to improve our understanding of gene expression and proteome diversity, and may in the long-term perspective facilitate the assignment of biological functions to protein N-terminal variants.
SAMENVATTING


Daarnaast is de technologische vooruitgang rond next-generation sequentieanalyses en massaspectrometrie onontbeerlijk gebleken om de complexiteit van transcriptomen, translatomen en proteomen in kaart te brengen. Meer specifiek toont ribosoomprofiling of Ribo-seq - een recent ontwikkelde technologie die gebaseerd is op de sequentieanalyse van mRNA fragmenten die door ribosomen beschermd worden - het wijdverspreid gebruik aan van alternatieve translatie-initiatieplaatsen (TIS), inclusief de selectie van niet-AUG startcodons, en dit in verschillende genomen. Daarenboven konden we in ons labo via positionele proteoanalyse (zijnde N-terminale COFRADIC, een technologie oorspronkelijk ontwikkeld in ons laboratorium) de diversiteit van het N-terminoom heel nauwkeurig in kaart brengen. Uit een gedetailleerde analyse van eiwit N-termini bleek dat in hogere eukaryoten ongeveer 20% van alle geïdentificeerde N-terminale peptiden afkomstig zijn van alternatieve translatie-initiatie, alternatieve splicing van prematuur mRNA of de incorrecte annotatie van translatiestartplaatsen in referentie eiwitdatabanken.

Aangezien deze technologieën in het labo beschikbaar waren, werd besloten om het landschap van N-terminale proteovormen in kaart te brengen. In een eerste studie rapporteerden we over de identificatie van meer dan 1.700 unieke, alternatieve eiwit N-termini (zijnde alternatieve N-terminale peptiden) afkomstig van diverse cellulaire muis of mens proteomen. Aangepaste eiwitdatabanken, gebaseerd op het translatie-initiatieprofiel afgeleid uit Ribo-seq data toonden tevens het gebruik aan van niet-AUG startcodons aan die gedecodeerd werden tot initiatormethionines, naast de aanwezigheid van N-terminaal

Vervolgens werd de cellulaire eiwitstabiliteit onderzocht met behulp van positionele proteoomanalyses en ribosoomb profilering. Door het combineren van ‘pulsed SILAC’ (pSILAC) met N-terminale proteomanalyses konden we de stabilité van ongeveer 2.000 N-terminale proteovormen in kaart brengen. Meer bepaald werden 147 N-terminale proteovormparen afkomstig van alternatieve translatie-initiatie, alternatieve splicing of partiële verwijdering van het initiatormethionine geïdentificeerd. N-terminaal verkorte proteovormen waren doorgaans minder abundant in vergelijking tot hun databank geannoteerde variant en vertoonden vaak een gewijzigde stabilité. Deze wijziging in stabilité kon toegeschreven worden aan individuele eiwitkenmerken zoals eiwitopvouwing, maar bleken onafhankelijk van bijvoorbeeld de identiteit van het N-terminale aminozuur of de wijziging in eiwitlengte veroorzaakt door het gebruik van een alternatieve TIS. We ontdekten dat het verwijderen van het initiatormethionine door methionine aminopeptidasen de stabilité van deze proteovormen verminderde, terwijl de gevoeligheid voor N-terminale acetylatie - een andere cotranslationele modificatie - de stabilité van eiwitten nagenoeg niet leek te beïnvloeden. Samengevat wijzen onze bevindingen op verschillen in eiwitstabiliteit tussen N-terminale proteovormen en een rol voor alternatieve translatie-initiatie en co-translationele verwijdering van het initiatormethionine in de globale regeling van de proteoomhomeostase.

Om de voordelen van de integratie van diverse OMICS-data voor de identificatie van MS/MS-spectra te onderzoeken, voerden we een integratieve proteoom en ribosoomb profilering studie uit op een humane cellijn als modelsysteem. Ribosoomb profilering werd gebruikt om (alternatieve) open leesramen en genetische variatie (nucleotide polymorfismen) in kaart te brengen. Door de incorporatie van dergelijke informatie in een aangepaste eiwitdatabank, werd een verbetering in identificatiescore van 69 eiwitten vastgesteld en konden 22 nieuwe proteovormen geïdentificeerd werden op basis van proteoomdata. We vonden voorts een verbeterde correlatie tussen kwantitatieve ribo-seq (translatoom) data en proteoomdata, dit in vergelijking met de reeds eerder gerapporteerde, veelal slechte correlatie tussen mRNA en eiwit expressie. Via N-terminale proteoomanalyses, werden ook 208 N-terminaal verkorte en 9 N-terminaal verlengde eiwit proteovormen geïdentificeerd, naast het eiwitproduct van één stroomopwaarts overlappend open leesraam (upstream ORF of uORF).
In tegenstelling tot de schaarse evidentie van peptidendata afkomstig van de expressie van uORFs, suggereert ribosoomprofilering een wijdverspreide translationele activiteit van uORFs in de 5' leader sequenties van eiwitcoderende genen. Meer bepaald wordt aangenomen dat uORFs hun rol voornamelijk uitoefenen door het onderdrukken of activeren van standaard eiwittranslatie, en dit ongeacht de overeenkomstige peptiden die mogelijks gevormd worden. Eukaryotische translatie-initiatiefactoren (eIFs) lijken belangrijke regulatoren van uORF activiteit door het koppelen van het cellulaire metabolisme en welbepaalde (externe)stimuli aan wijzigingen in translatie. In ons meest recente werk werden verschillende OMICS-strategieën (proteomics, ribosoomprofiling en mRNA sequentie-analyse) aangewend om de biologische doelwitten van de eukaryote translatie-initiatiefactor 1 (eIF1) in kaart te brengen. eIF1 is betrokken bij de striktheid van startcodon selectie. Het monitoren van wijzigingen in expressie van meer dan 4.000 eiwitten en 10.000 actief vertaalde transcripten resulteerde in de identificatie van 245 genen onderhevig aan eIF1 translationele controle gemedieerd door uORFs. De striktheid van startcodon selectie en de voorkeur voor een optimale nucleotidecontext was grotendeels verminderd, wat resulteerde in een gewijzigde expressie van uORFs met suboptimale TIS (niet AUG, slechte context) en bijgevolg de translationele regulerend van de corresponderende stroomafwaarts gelegen open leesramen. Voornamelijk genen betrokken bij energieproductie en de detectie van metabole stress bleken onderhevig te zijn aan deze vorm van eIF1 gereguleerde translationele controle. Interessant is het ook op te merken is dat neerregulatie van eIF1 een synergetisch effect van eIF5 en eIF1B teweegbrengt, een bevinding die de interconnectiviteit van verschillende translatiefactoren benadrukt.

Onze resultaten tonen dus aan dat ribosoomprofiling en (positionele) proteoomanalyses zeer complementaire data leveren die waardevolle nieuwe inzichten geven omtrent alternatieve translatie-initiatie op een genoomwijde schaal. Wij zijn ervan overtuigd dat de identificatie en karakterisering van N-terminale proteovormen onder meer door functionele studies (proteovorm stabilititeit, lokalisatie, abundantie en regulatie) essentieel is om genregulatie beter te begrijpen en proteoomdiversiteit beter in kaart te brengen. Op termijn kan dit het toewijzen van biologische functies aan N-terminale eiwitvarianten vereenvoudigen.
CURRICULUM VITAE
CURRICULUM VITAE

Personal information

Name: Daria Fijałkowska
Maiden name: Gawron
Address: Meulestedekaai 6, 9000 Gent, Belgium
Telephone: +32 486 52 53 59
E-mail: daria.gawron87@gmail.com
Date of birth: 13.10.1987
Place of birth: Zabrze, Poland

Professional information

Address: VIB Medical Biotechnology Center,
UGent Department of Biochemistry,
Faculty of Medicine and Health Sciences
Ghent University
A. Baertsoenkaai 3
B-9000 Ghent, Belgium
Tel. +32 9 264 93 60
Fax +32 9 264 94 96
E-mail: daria.gawron@ugent.be
ResearcherID: F-9130-2016

Education

2009 – 2011 Master of Biotechnology diploma, specialty Biotechnology of Peptides and Proteins, received from the University of Wroclaw, Faculty of Biotechnology, Poland. Master thesis entitled:

“Optimization and application of real-time PCR method in gene expression studies in selected biological models. Relation between mRNA and protein level.”

2006 – 2009 Bachelor of Biotechnology diploma received from the University of Wroclaw, Faculty of Biotechnology, Poland. Bachelor thesis entitled:

“Activation of sphingosine pathway to apoptosis in eutopic endometrial cells in women with endometriosis.”
Experience and academic record

10.2011 – present  PhD student at Ghent University, Belgium

01.2013 – present  personal PhD scholarship awarded by the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT)

10.2011 – 12.2012  Ghent University scholarship recipient

02.2010 – 06.2010  exchange student at the University of Antwerp, Belgium

Research project on genetics of hearing loss performed under supervision of Prof. Guy Van Camp entitled: “Sequencing and analysis of variations in BMP2 gene in a large Belgian population”.

08.2010 – 09.2010  student internship at the Cardiac Surgery Development Foundation in Zabrze, Poland

Work at the laboratories of histopathology, biotechnology and biocybernetics.

10.2010 - 12.2010  laboratory technician at the University of Wroclaw, Poland

Project commissioned by the EIT+ Foundation (http://www.eitplus.pl/), entitled: “Evaluation of VDR and CD14 gene expression in HL-60 cells exposed to vitamin D and its analogues, using real-time PCR method”.

Additional scientific training

“Ingenuity Pathway Analysis (IPA) training”, BioInformatics Training and Services (BITS), Flanders Institute for Biotechnology (VIB), Ghent, Belgium

“RNA-Seq analysis for differential expression”, BioInformatics Training and Services (BITS), Flanders Institute for Biotechnology (VIB), Ghent, Belgium

“Introduction to the analysis of NGS data”, BioInformatics Training and Services (BITS), Flanders Institute for Biotechnology (VIB), Leuven, Belgium
Publications in scientific journals


Conference contributions


APPENDIX

**Part I: RESULTS Chapter I**
- Appendix Figure I.1
- Appendix Figure I.2
- Appendix Table I.1
- Appendix Table I.2
- Appendix File I.1

**Part II: RESULTS Chapter II**
- Appendix Figure II.1
- Appendix Figure II.2
- Appendix Figure II.3
- Appendix Figure II.4
- Appendix Figure II.5
- Appendix Table II.1
- Appendix Table II.2

**Part III: RESULTS Chapter III**
- Appendix Figure III.1
- Appendix Figure III.2
- Appendix Table III.1
- Appendix Table III.2

**Part IV: RESULTS Chapter IV**
- Appendix Figure IV.1
- Appendix Figure IV.2
- Appendix Figure IV.3
- Appendix Table IV.1
- Appendix Table IV.2
- Appendix Table IV.3
- Appendix File IV.1
Appendix Part I

Appendix Figure I.1

Appendix Figure I.1 - WebLogos created for the flanking sequences (12 bases upstream, 9 bases downstream) of the dbTIS sites of 5,000 randomly chosen CCDS proteins for both *Homo sapiens* and *Mus musculus*. Both the probability and bits values are plotted. The CCDS set is built by consensus between the European Bioinformatics Institute (EBI), the National Center for Biotechnology Information (NCBI), the Wellcome Trust Sanger Institute (WTSI), and the University of California at Santa Cruz (UCSC).

Appendix Figure I.2

Appendix Figure I.2 - Conservation plots for the flanking (12 bp upstream, 9 bp downstream) exonic regions of human and mouse dbTIS sites of 5,000 randomly chosen CCDS proteins. The conservation measure averages the phastCons score at every position of all flanking sequences after alignment based on the translation start site. For all flanking positions, the mean is provided together with its 95% confidence interval. The CCDS set is built by consensus between the European Bioinformatics Institute (EBI), the National Center for Biotechnology Information (NCBI), the Wellcome Trust Sanger Institute (WTSI), and the University of California at Santa Cruz (UCSC).
Appendix Table I.1

Appendix Table I.1 can be accessed via the URL:

https://goo.gl/hoz76i

Appendix Table I.1 - List of identified N-termini matching Swiss-Prot annotated and non-annotated translation initiation sites in mice and human. A. List of 2,879 N-termini (corresponding to 2723 unique protein entries) indicative of dbTIS in human. B. List of 1,231 N-termini (corresponding to 1030 unique protein entries) indicative of dTIS in human. C. List of 1,771 N-termini (corresponding to 1708 unique protein entries) indicative of dbTIS in mice. D. List of 465 N-termini (corresponding to 418 unique protein entries) indicative of dTIS in mice. Lists of 19 and 23 identified non-canonical mouse (E) and human (F) N-terminal peptides matching N-terminal extensions, uORF, and overlapping uORF as determined by ribosome profiling [1, 2]. In each case the start and end positions, the first two amino acid residues (P1’ and P2’) and the amino acid residue preceding the identified N-terminus (if any) (P1), identified peptide sequence, Ensembl gene and transcript ID, chromosomal and DNA strand location, chromosomal location and exon number of TIS start, UniProt database primary accession number, whenever the iMet-retaining and iMet processed -termini were identified (-/+ iMet) and a dTIS or dbTIS in the same protein accession could be identified, Uniprot name, protein description, matching protein isoforms (if any), the overall degree of Nt-acetylation (if determined), N-terminal modification status/states confirmed by MS/MS, spectral count, maximum Mascot ion score, delta threshold (= maximum score - minimum threshold), minimum threshold score and number of experiments and cell lines in which the N-terminus was identified are indicated. In the dTIS lists, additional information concerning the corresponding mouse peptide and identification thereof in the mouse proteome background (retrieved by globally aligning the surrounding sequences of aTIS sites in all known homologs or orthologs (found in the HomoloGene database) [3]), whether the dTIS matches a dbTIS of (a) Swiss-Prot or TrEMBL ent(r)y(ies), the flanking nucleotide context, the length of the canonical and truncated CDS and gene name info are additionally given. In the lists of non-canonical mouse (E) and human (F) N-terminal peptides matching N-terminal extensions, uORF, and overlapping uORF as determined by ribosome profiling, additional information on the start codon (type), resulting type of protein product, the start position relative to the dbTIS and transcript region of the TIS start are given. N-termini are ranked alphabetically according to their protein description (or gene name) and start position.
# Appendix Table I.2

<table>
<thead>
<tr>
<th>Gene name</th>
<th>TIS aa pos.</th>
<th>Sequence of forward mutation primer (mutation: ATG&gt;CTG)</th>
<th>Sequence of reverse mutation primer (mutation: ATG&gt;CTG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIMP2</td>
<td>1</td>
<td>TGCTTTGTTCTTCGCCCTGGCAGTTACAGCA</td>
<td>CTTGGTACATCGCGAGGAGCAGACAAAGCA</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>GGTCTCTGCCATGCCGGCTCTGTTCAAGTTGCGC</td>
<td>GCTTACCTGAGTCAAGCGATGGGAACCC</td>
</tr>
<tr>
<td></td>
<td>1+3</td>
<td>TGCTTTGTTCTTCGCCCTGGCAGTTACAGCA</td>
<td>GCTTACCTGAGTCAAGCGATGGGAACCC</td>
</tr>
<tr>
<td>CRUL3</td>
<td>1</td>
<td>CGCGGGCGACGTGAGGGGGGCG</td>
<td>CGCGGGCGACGTGAGGGGGGCG</td>
</tr>
<tr>
<td>NEMO</td>
<td>1</td>
<td>CCTGGCGCGCCCGCTGCTGACAAATAGTGG</td>
<td>CATTCTTATATGGCAAGCGGGCGGGCG</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>CCGGAATTCGCTAATACGTGAGCGCTGATCTCTTCTAC</td>
<td>GTGAAAGGCTGATCGTTAGTCAGTACACACAGGCG</td>
</tr>
<tr>
<td></td>
<td>1+5</td>
<td>CCTGGCGCGCCCGCTGCTGACAAATAGTGG</td>
<td>CATTCTTATATGGCAAGCGGGCGGGCG</td>
</tr>
<tr>
<td>JUNB</td>
<td>1</td>
<td>GGCGGCCCCCCCTGAGTCTGGAG</td>
<td>CCACGAGTCTAGGGGGGGGCG</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>CGCGGGCAGCTGCGCTGAGCAAGCTGGAGTGG</td>
<td>GTGAAAGGCTGATCGTTAGTCAGTACACACAGGCG</td>
</tr>
<tr>
<td></td>
<td>1+5</td>
<td>CCTGGCGCGCCCGCTGCTGACAAATAGTGG</td>
<td>CATTCTTATATGGCAAGCGGGCGGGCG</td>
</tr>
<tr>
<td>NAP1L1</td>
<td>1</td>
<td>GATATTTGGAGTTCTTACAAGTCCGGAGACATCGGAAAGCAAG</td>
<td>CTTGTTGTGCAATGCTGACAGGAGCCACAAAATAC</td>
</tr>
<tr>
<td></td>
<td>42</td>
<td>TCTGAGAATTTTCTCTCACTGAGGCAAGGTTAAGGTGGCAAGCG</td>
<td>CTGACGCTTATGCAGCTGACAGGGTTAAGGTGGCAAGCG</td>
</tr>
<tr>
<td></td>
<td>43</td>
<td>GTCACTAATGGCAAGCTGCTGCAAGGCGTAACTAGTCTCTTCTTT</td>
<td>AAGAATCTGAGATTTTCTCAGTACATGCGGGCCG</td>
</tr>
<tr>
<td></td>
<td>42+43</td>
<td>CGCGTCTACGCTAATGGCAAGCTGCTGCAAGGCGTAACTAGTCTCTT</td>
<td>AAGAATCTGAGATTTTCTCAGTACATGCGGGCCG</td>
</tr>
<tr>
<td>PNPO</td>
<td>1</td>
<td>GGCGGCCGCCCTGAGTCTGGAG</td>
<td>CCACGAGTCTAGGGGGGGGCG</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>CGCAGCTGCTGGCACGGCCACCCTGGGAC</td>
<td>GTCCACGCTTCAGGGGCGACAGCTGCG</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>CGCCGAGCTAAGGGGAAGGGGAGAGCAGGAAGATCTTGGAG</td>
<td>TCCAGATTATGTGCTACAGGAGGCGACCGCATGCTGACAGGGAGGCGCG</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>CGCGCTCTTGAGCTCAGGCAAGATGGATTAGTGGCTGAGG</td>
<td>GCATCTCACATAGCTGAGTGGGAACATCTGACAGGGGCA</td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>AGCTCTTGGTCAGAAGACGTGCTGCAAGGCTGAGGCGAGC</td>
<td>GCATCTCACAGTGGACGGCTGCAAGGCTGAGGCGAGC</td>
</tr>
<tr>
<td></td>
<td>22+26</td>
<td>CGCGCTCTTGAGCTCAGGCAAGATGGATTAGTGGCTGAGG</td>
<td>GCATCTCACATAGCTGAGTGGGAACATCTGACAGGGGCA</td>
</tr>
<tr>
<td>SF3A</td>
<td>1</td>
<td>GGCCTTCGGAGACAGTGCGCTGAGAATGAGCC</td>
<td>GGCTTTGAGCGTGCTGACAGGAGGCGACCC</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>TGCACTACGCGCGTGGAGAGCAGTACACACCC</td>
<td>GGCTTTGAGCGTGCTGACAGGAGGCGACCC</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>GGAGCTACCAACCCCGAGCTGCTGACAAAAGTGGCTCAGTCTC</td>
<td>AGCAGTTGGTGAGCGTGCTGACAGGAGGCGACCC</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>TCCGCTTCGCTGCGTGGCCAGGCGC</td>
<td>GGCGCGGAGGGAGAGGAGGGGCG</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>CGCGGGAGATCTTGGAGGGCGCG</td>
<td>CGCCGCGTCAGAGTCTCCGGCG</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>GCCCGCGACCCCGCTGCGCTGGGCG</td>
<td>GCCCGCGACCCCGCTGCGCTGGGCG</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>AACCCACAGAGAGGCTGAAATGAGACCC</td>
<td>GGTTCTTGGCTATCCAGCTGCTGACCTGCTGAGG</td>
</tr>
<tr>
<td></td>
<td>27</td>
<td>CCCAGCAGAGAAGAGATGAGACTGAGAGAC</td>
<td>AGTCTTGGACGCCAGCTGCTGACCTGCTGCTGAGG</td>
</tr>
<tr>
<td></td>
<td>25+27</td>
<td>CCCAGCAGAGAAGAGATGAGACTGAGAGAC</td>
<td>AGTCTTGGACGCCAGCTGCTGACCTGCTGCTGAGG</td>
</tr>
</tbody>
</table>

---

**Appendix Table I.2 – Primers used in site-directed mutagenesis experiments.**
Appendix File I.1

The H2G2 Genome Browser can be accessed via the URL [http://h2g2.ugent.be/biobix.html](http://h2g2.ugent.be/biobix.html) using the following login credentials: **Name**: aTIS - **Password**: SITa1052.

In the upper left corner under “Select Project” one can choose “TIS Human” or “TIS Mouse”. The genome browse software lets you zoom from the chromosome to the nucleotide level in a semantic way, as such data is loaded and displayed according to the zooming level and region displayed. The projects hold several visualization tracks: (A) an Ensembl gene track (B) an Ensembl transcript track (visible after mouse-click on the gene track), and tracks holding (C) the annotated TIS within the UniProtKB database split into SwissProt and trEMBL (D) the annotated TIS within the Ensembl database (E) the reported aTIS and dbTIS locations (F) a conservation track based on the PhastCons conservation scores based on alignment of 45 and 29 vertebrate genomes respectively for human and mouse. (G) 3 extra tracks with ribosome profiling information (Ingolia et al., Cell, 2011 data for mESC and Lee et al., PNAS, 2012 for HEK293), one depicting the cycloheximide treated sample, one depicting the harringtonine or lactimidomycin treated sample and one with the translation product prediction based on the 2 aforementioned tracks. Furthermore all genes where an aTIS or dbTIS has been identified by means of the N-terminal COFRADIC experiments are listed in “GeneDigest” reports within the H2G2 genome browser environment. The aTIS “GeneDigest” report lists all genes wherefore an alternative start site is reported, whereas the dbTIS “GeneDigest” report lists all genes wherefore a Swiss-Prot database annotated TIS has been identified. A third “Genedigest” report lists extra translation start sites identified from the N-terminomics experiments searching a protein product database constructed based on ribosome profiling sequence information. This ribosome profiling information is also mapped on the reference genome and presented in the H2G2 genome browser interface as separate visualization tracks.

Extra information in the form of static and sample tracks can be dragged into the visualization field from the left. Further information on the usage of the H2G2 genome browser can be found in the manual which can be downloaded here: [http://h2g2.ugent.be/downloads/ManualV1_0.pdf](http://h2g2.ugent.be/downloads/ManualV1_0.pdf).
Appendix Part II

Appendix Figure II.1

A. Representative proteoforms (10) identified in 7 time points displaying a large variety of turnover rates. B. For each peptide, all possible combinations of 3 random points were fitted with an exponential model and used for calculating 50% turnover time. Next, models with R$^2$ coefficient below 0.8 were rejected. The remaining valid models were grouped per peptide and represented as box plots.
Appendix Figure II.2 - For each proteoform quantified in at least 3 time points (1,972), variation of data unexplained by the exponential model was calculated as $(1-R^2)$ and represented in %. The distribution of the unexplained variations is tightly centred around 2% (corresponding to a median $R^2=0.98$) while only a minority of exponential models fitted insufficiently ($R^2<0.8$) and were rejected.

Appendix Figure II.3 - Experimental assignment of Jurkat cell doubling time. Density of cell culture was monitored in triplicate over the course of 48 hours and fitted with an exponential model using the Doubling Time Online Calculator (http://www.doubling-time.com/compute.php). Cell doubling time was calculated from the assigned growth rate (0.0288) as follows: $\ln(2)/0.0288$ and represented in hours.
Appendix Figure II.4

Appendix Figure II.4 - Distribution of turnover values for lysine containing ubiquitinated (N = 348) and non-ubiquitinated dbTIS indicative N-termini (N = 388).

Appendix Figure II.5

Appendix Figure II.5 - SiLAC labelling of Jurkat cells. Jurkat cells cultured in regular RPMI medium were transferred to RPMI SiLAC medium containing Arg⁶ L-arginine and cultured for
7 days. Subsequently, cells were transferred to a RPMI SILAC medium containing Arg\(^0\) L-arginine for next 7 days, to achieve complete unlabelling. The isotope replacement was monitored by quantifying the identified MS-spectra using Mascot Distiller.

Appendix Table II.1

Appendix Table II.2 can be accessed via the URL:

https://goo.gl/hoz76i

Appendix Table II.1 - Stability of N-terminal proteoforms in Jurkat cells. A. Calculations of 50\% turnover times and abundance of 2,578 proteoforms identified across the time points studied. B. A detailed analysis of 1,972 proteoforms identified in at least 3 time points. Information on proteoform-specific measurements of turnover, abundance, disorder, hydropathy and dbTIS association with macromolecular complexes are given. C. Tracking the origin of 2,578 N-terminal proteoforms identified in our study using ribosome profiling and extensive mining of public databases, including Swiss-Prot Isoform, UniProt TrEMBL, Ensembl and TopFIND 3.0. Column headers used in A-C. are additionally explained in legends supplied in the table.

Appendix Table II.2

(see next page)

Appendix Table II.2 - Contribution of ubiquitination to proteoforms stability. Ubiquitination sites previously reported in Jurkat cells [4] were mapped onto dbTIS indicative N-termini identified in this study. As such, we found 11 proteoform pairs of which only the database-annotated variant holds (a) previously identified, and thus possible ubiquitination site(s) with lower stability as compared to its N-terminally truncated counterpart.
<table>
<thead>
<tr>
<th>Accession</th>
<th>Entry name</th>
<th>Start</th>
<th>Stop</th>
<th>% of turnover time in</th>
<th>No. of valid turnover time measurements</th>
<th>Time at stop (dBT15-A15)</th>
<th>Time at start (dBT15-A15)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q43339</td>
<td>TPO4_HUMAN</td>
<td>1</td>
<td>34</td>
<td>Yes</td>
<td>7</td>
<td>28.08</td>
<td>28.08</td>
</tr>
<tr>
<td>P07910</td>
<td>HNRPC_HUMAN</td>
<td>2</td>
<td>14</td>
<td>Yes</td>
<td>7</td>
<td>30.53</td>
<td>30.53</td>
</tr>
<tr>
<td>Q85935</td>
<td>HNRPC_HUMAN</td>
<td>2</td>
<td>92</td>
<td>Yes</td>
<td>7</td>
<td>37.48</td>
<td>37.48</td>
</tr>
<tr>
<td>P36289</td>
<td>PIBPL_HUMAN</td>
<td>1</td>
<td>112</td>
<td>Yes</td>
<td>4</td>
<td>17.67</td>
<td>17.67</td>
</tr>
<tr>
<td>P39019</td>
<td>R511_HUMAN</td>
<td>2</td>
<td>121</td>
<td>Yes</td>
<td>7</td>
<td>36.69</td>
<td>36.69</td>
</tr>
<tr>
<td>P5209</td>
<td>NPL1_HUMAN</td>
<td>42</td>
<td>55</td>
<td>Yes</td>
<td>6</td>
<td>28.43</td>
<td>28.43</td>
</tr>
<tr>
<td>P80778</td>
<td>HNRPC_HUMAN</td>
<td>1</td>
<td>36</td>
<td>Yes</td>
<td>7</td>
<td>21.31</td>
<td>21.31</td>
</tr>
<tr>
<td>Q16833</td>
<td>TLSB_HUMAN</td>
<td>2</td>
<td>36</td>
<td>Yes</td>
<td>7</td>
<td>21.82</td>
<td>21.82</td>
</tr>
<tr>
<td>Q02307</td>
<td>NONO_HUMAN</td>
<td>1</td>
<td>22</td>
<td>Yes</td>
<td>7</td>
<td>21.14</td>
<td>21.14</td>
</tr>
<tr>
<td>Q18519</td>
<td>NONO_HUMAN</td>
<td>1</td>
<td>17</td>
<td>Yes</td>
<td>7</td>
<td>22.14</td>
<td>22.14</td>
</tr>
<tr>
<td>O70450</td>
<td>TNNK1_HUMAN</td>
<td>1</td>
<td>12</td>
<td>Yes</td>
<td>7</td>
<td>10.75</td>
<td>10.75</td>
</tr>
</tbody>
</table>

Sequence with underline: NGN2 - desdication site at Up (N)
Appendix Part III

Appendix Figure III.1 - Examples of improved identifications in the shotgun proteomics experiment. The addition of ribo-seq data to the proteomics experiment improved the identification and score significance for 69 proteins and three representative examples are depicted here. The left column shows the Clustal Omega alignment of the ribo-seq-derived amino acid sequences to the Swiss-Prot sequences with the relevant peptide identifications highlighted in cyan. The column on the right shows the corresponding fragmentation spectra and peptide sequence fragmentations.
Appendix Figure III.2 - Correlation plots of protein abundance based on NSAF values and RPF counts for the proteins uniquely identified in Swiss-Prot. Some transcripts were not contained in our custom database because the LTM treatment and/or TIS calling failed to identify these TISs. Correlations could still be calculated as the CHX treatment did result in detectable coverage for these transcripts. The number of data points used in every plot was lower than the total number of unique Swiss-Prot identifications (312), because whenever a Swiss-Prot protein corresponded to multiple transcripts only the transcript with the highest normalized RPF value was used. Top left: all transcripts; top right: transcripts with ribo-seq coverage in all exons; bottom left: all transcripts with an RPF count ≥ 200; bottom right: transcripts with both coverage in all exons and an RPF count ≥ 200. The regression line is shown in green. For each plot, the number of data points used (i.e. the number of dbTIS transcripts) as well as the corresponding Pearson correlation coefficient ($r^2$) is shown.
Appendix Table III.1

Appendix Table III.1 can be accessed via the URL:

https://goo.gl/hoz76i

Appendix Table III.1 - General overview of peptide and protein identifications. A. List of all 2,816 human protein products identified in HCT116 cell lysates. B. List of all 1,289 human protein N-terminal peptides identified in HCT116 cell lysates.

Appendix Table III.2

Appendix Table III.2 can be accessed via the URL:

https://goo.gl/hoz76i

Appendix Table III.2 – A gene ontology enrichment analysis of 312 proteins missed by the ribo-seq experiment, but identified thanks to the inclusion of Swiss-Prot in the search space.
Appendix Figure IV.1 - Ribosome density profiles of transcripts with actively translated u(-o)ORFs. Full length ribosome CHX reads of si-Ctrl and si-ElF1 samples were visualized across mRNA sequences as positive (blue) and negative (red) values, respectively. Only the 5'
proximal part of the 3’UTR sequence was included. Both uniquely and non-uniquely mapped reads were included. CDS and u(-o)ORFs were marked using grey or green/red/blue/yellow bars, respectively. Direction of expression changes and start codon of the most 5’ proximal u(-o)ORF were indicated (see figure). Potential regulatory u(-o)ORFs were visualized for *UBE2S, RPL12, MIEF1, PCNXL4, GAPDH* and *TOMM7*. These u(-o)ORFs contributed to significant Z-score TE uORF/ TE CDS changes, except in the case of *UBE2S*, which fell below our significance threshold, but displayed significant translational regulation of CDS expression. Genes presented in the two upper rows were reported to undergo uORF-directed translational regulation upon arsenite treatment or glucose and oxygen deprivation conditions in previous ribo-seq studies (*UBE2S*[6], *RPL12, MIEF1* and *PCNXL4*[7]). Additionally, we presented the ribosome profile of *POLR2M*. Although *POLR2M* was unaffected by eIF1-mediated translational regulation, we confirmed the expression of its u-oORFs with two independent peptide identifications (3 in-frame u-oORFs were detected by ribo-seq holding the identified peptides, see Supplementary Table 2).

Appendix Figure IV.2

**Appendix Figure IV.2 – Analysis of the length of overlap between u-oORF and CDS.** A. Translation efficiency of significantly regulated CDSs was generally not affected by the overlap with of u-oORFs. However, the 68 cases of differentially regulated u-oORF / CDS pairs were characterised by shorter than average overlap (B).
Appendix Figure IV.3 – WebLogo analysis [8] representing frequencies of nucleotides surrounding start sites. uTIS context sequences were represented for regulatory u(-o)ORFs with AUG (A.) and non-AUG (B.) uTIS (Z-score TE uORF/TE CDS ≥ 2.58 as upregulated and Z-score TE uORF/TE CDS ≤ -2.58 as downregulated). C. Context analysis of all aTIS identified in our study. WebLogo analysis revealed very weak nucleotide consensus sequence surrounding upregulated AUG uTIS ((C/T)TGATGT, lacking the -3 purine and +4 guanine) in comparison to optimal Kozak context in the aTIS group ((A/G)CCATGG).
Appendix Table IV.1

Appendix Table IV.1 can be accessed via the URL:

https://goo.gl/hoz76i

**Appendix Table IV.1 – Ribosome profiling and proteomics readout of translation upon eIF1 knockdown in HCT116 cells.** Quantification of gene expression changes at several OMICS levels: mRNA, translation, translation efficiency and protein was supplemented with a detailed analysis of upstream (-overlapping) ORFs.

Appendix Table IV.2

Appendix Table IV.2 can be accessed via the URL:

https://goo.gl/hoz76i

**Appendix Table IV.2 – Customized protein library is a source of novel peptide identifications.** Examples of novel proteoforms and non-synonymous variants identified in our study using ribosome profiling (for delineation of in vivo translated ORFs and generation of a custom protein sequence database) in combination with label-free shotgun proteomics.

Appendix Table IV.3

Appendix Table IV.3 can be accessed via the URL:

https://goo.gl/hoz76i

**Appendix Table IV.3 – List of primers and oligonucleotides.**
Appendix File IV.1

Appendix File IV.1 can be accessed via the URL:

https://goo.gl/hoz76i

Appendix File IV.1 - BED file of peptides identified in the “custom database” search of the HCT116 label-free shotgun proteomics data mapped onto the human GRCh38 reference genome (Ensembl annotation bundle 82).

References:

ADDITIONAL PUBLICATIONS
PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration

Jeroen Crappé, Elvis Ndah, Alexander Koch, Sandra Steyaert, Daria Gawron, Sarah De Keulenaer, Ellen De Meester, Tim De Meyer, Wim Van Criekinge, Petra Van Damme and Gerben Menschaert

1Lab of Bioinformatics and Computational Genomics, Department of Mathematical Modeling, Statistics and Bioinformatics, Faculty of Bioscience Engineering, Ghent University, Ghent, Belgium. 2Department of Medical Protein Research, Flemish Institute of Biotechnology, Ghent, Belgium. 3Department of Biochemistry, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium

ABSTRACT
An increasing amount of studies integrate mRNA sequencing data into MS-based proteomics to complement the translation product search space. However, several factors, including extensive regulation of mRNA translation and the need for three- or six-frame-translation, impede the use of mRNA-seq data for the construction of a protein sequence search database. With that in mind, we developed the PROTEOFORMER tool that automatically processes data of the recently developed ribosome profiling method (sequencing of ribosome-protected mRNA fragments), resulting in genome-wide visualization of ribosome occupancy. Our tool also includes a translation initiation site calling algorithm allowing the delineation of the open reading frames (ORFs) of all translation products. A complete protein synthesis-based sequence database can thus be compiled for mass spectrometry-based identification. This approach increases the overall protein identification rates with 3% and 11% (improved and new identifications) for human and mouse, respectively, and enables proteome-wide detection of 5'-extended proteoforms, upstream ORF translation and near-cognate translation start sites. The PROTEOFORMER tool is available as a stand-alone pipeline and has been implemented in the galaxy framework for ease of use.

INTRODUCTION
The integration of next-generation transcriptome sequencing and highly sensitive mass spectrometry (MS) has emerged as a powerful strategy for the fast and comprehensive profiling of mammalian proteomes (1). Protein sequence database search tools (2) typically use publicly available protein databases, such as Swiss-Prot and Ensembl, to match MS spectra to peptides. Because these reference databases only contain experimentally verified and/or predicted protein sequences, it is very unlikely that they give a comprehensive assessment of the expressed protein pool of a given sample. Translation product prediction based on messenger RNA sequencing (mRNA-seq) data gives a more representative state of the protein repertoire expressed and aids the protein identification process by eliminating unexpressed gene products from the search space (3). On top of that, transcript data additionally provides sequence variation information, such as single nucleotide polymorphisms (SNPs) and RNA-splice and editing variants (4), which improve the chances of identifying novel protein forms (5,6).

Despite the benefits of adding mRNA-seq information to proteomics experiments, this approach has some shortcomings. First, mRNA levels are not a perfect proxy for protein expression levels since the translation of mRNA is subject to extensive regulation (7). Furthermore, there are several factors, such as internal ribosome entry sites, non-AUG start codons and non-sense read-through (8), that hinder the prediction of the exact protein product(s) translated from the transcript sequence. Also, inclusion of mRNA-seq information requires three- or six-frame-translation of the derived sequences, dramatically expanding the protein search space and hence decreasing the search sensitivity and specificity (9).

Recently a new strategy, termed ribosome profiling (RIBO-seq), was introduced that overcomes these shortcomings (8). By using the property of translating ribosomes to protect mRNA fragments from nuclease digestion it is possible to directly monitor the in vitro synthesis of mRNA-encoded translation products measured at the genome-wide level (10). In contrast to polysome profiling, often used for

© The Authors 2014. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
analyzing gene expression, RIBO-seq enables delineation of the genomic positions of translating ribosomes with subcodon to single-nucleotide precision (11). Furthermore, (alternative) translation initiation sites (TIS) can be accurately predicted by exploiting the abilities of antibiotics, such as harringtonine (HAR) or lactimidomycin (LTM), that halt ribosomes at sites of translation. However, as some non-coding transcripts show association with ribosomes (12), MS-assisted validation is in many cases still indispensable (13).

The presented PROTEOFORMER tool processes RIBO-seq data allowing genome-wide visualization of protein synthesis, and moreover enables the delineation of in vivo proteoforms (14) building an optimal protein sequence search database for peptide to MS/MS matching (15–18) (Figure 1). PROTEOFORMER starts with the mapping of ribosome-protected fragments (RPFs) and quality control of subsequent alignments. It further includes modules for identification of transcripts undergoing protein synthesis, positions of translation initiation with subcodon specificity and SNPs. We used PROTEOFORMER to create protein sequence search databases from publicly available mouse (8) and in-house performed human RIBO-seq experiments and evaluated these with matching proteomics data. We demonstrate that this approach results in an increase of the number of protein/peptide identifications, leads to the identification of novel protein forms and aids in the re-annotation of the genome.

MATERIALS AND METHODS

The PROTEOFORMER pipeline (Figure 1) is made up of six major steps: (i) the alignment of the RPF reads to a reference genome, (ii) a quality control of the alignments, (iii) assignment of transcripts with evidence of translation, (iv) identification of TIS, (v) inclusion of SNP information and (vi) finally generation of a RIBO-seq derived translation product database that can be used as a search space for MS-based proteomics studies, either independently or combined with a canonical protein database. All input parameters for the different steps of the PROTEOFORMER pipeline are user-definable in order to allow research-specific optimization. A more detailed description of the parameter settings is available via the readme file and website (Supplemental File S2 and http://www.biobix.de/proteoformer).

Sequence processing and alignment

For the mouse and human sequences we use respectively the Ensembl (19) release 72 and 70 genome annotation (assembly GRCh38 and GRCh37) from the iGenome repository (http://support.illumina.com/sequencing/sequencing-software/igenome.html).

RIBO-seq-derived reads can be aligned using both a STAR (20) (2.3.0c_r291) or TopHat (21) (v2.0.9) based pipeline. The STAR-based workflow sequentially aligns the reads to STAR indices composed of the following sequences: (i) the PhiX bacteriophage genome, (ii) Mus musculus or Homo sapiens rRNA (obtained using BioMart, filtered on Mt_rRNA and rRNA gene types) and (iii) Mus musculus or Homo sapiens complete genome (obtained from the corresponding iGenome repository). The STAR internal clipping function is used to clip the 3' adaptor, up to two mismatches are allowed for the alignment, the option seedSearchStartLmaxOverRead is set to 0.5 and no introns are allowed for the alignment against the PhiX genome. The TopHat-based workflow uses Bowtie (v2.1.0) to sequentially align sequencing reads to Bowtie indices composed of the PhiX bacteriophage and the rRNA sequences (see above) using the 'sensitive-local' option, whereas TopHat itself is used for the complete genome alignment using default settings except for 'segment-length' that is set to 15. Since TopHat does not have an internal clipping functionality, the clipper from the FASTX Toolkit (0.0.13) is used to clip the 3' adaptor sequence prior to mapping. For the RPF distribution plots and quality controls, only uniquely mapped reads are accounted for whereas for the custom DB creation multi-mapping reads (up to 15 locations) are additionally considered. Only reads with a length between 26 and 34 bases (i.e. relevant RPFs) are retained for further genome-coordinate mapping. RPF alignments are assigned to the current ribosomal P-site, based on the length of the fragment. The offset from the 5'-end of the alignment is +12, +13 and +14, respectively, for alignments ≤30 bases long, 31–33 bases long or ≥34 bases long (8). The alignment and RPF density information are returned as output by PROTEOFORMER (BedGraph format) making it easy to upload and visually evaluate the data in a genome browser environment of choice (22).

Quality control: metagenic functional classification

As a first quality assessment, the obtained ribosomal footprints are classified using a species-specific Ensembl annotation bundle (converted to SQLite format). First, a metagenic functional annotation of the uniquely mapped footprints is determined using the Ensembl annotation of all transcripts. Here, translation associated annotation (i.e. 5' untranslated region (UTR), exon, intron or 3' UTR) is only defined for transcripts with a 'protein-coding' biotype. The RPFs not assigned to protein-coding transcripts are assigned to non-protein-coding transcripts (i.e. 'other biotypes'). The remaining unassigned footprints are classified as 'intergenic'. The resulting classification counts are available in a tab-separated table and summarized as a pie chart (Supplementary Figure S1a). For the ribosome footprints classified as 'other biotypes', a second table and accompanying pie chart is created, depicting the biotype distribution of these footprints (Supplementary Figure S1b).

Gene distribution

The quality is also assessed by determining the uniquely mapped ribosomal footprint counts per gene (using available Ensembl annotation). In total, three summarizing plots are available: (i) a gene abundance plot ranging from the highest to the lowest covered genes, (ii) a cumulative gene distribution plot ranging form the highest to the lowest covered genes and (iii) a gene density plot (for more details, see Supplementary Figure S2). These results are also stored as tabular files.

226
Transcript calling based on elongating ribosome coverage

Profiles of ribosomal footprints along a transcript are obtained by summing the number of footprints assigned to each genomic position of the coding sequence (CDS). The CDS of each known transcript is assembled using a species-specific Ensembl annotation bundle (converted to SQLite format). For non-protein-coding transcripts, the CDS is defined as the full exonic region of that specific transcript. For protein-coding transcripts, UTR-information is available, allowing us to determine the start and stop codons and to define the CDS as the exonic region between these two codons.

To remove variability in ribosomal footprint density due to RPF accumulation at start and stop codons (8,23), we additionally restrict the region where RPFs are counted by excluding the 15 nucleotides following each start codon and 15 nucleotides preceding each stop codon. For each transcript, the ribosomal footprint count is normalized based on the CDS length for which RPFs are taken into account (total CDS length – 30 bps). In order to identify the actual translated transcript isoforms, we examine the normalized footprint coverage of each of their exons. A transcript is denoted as truly translated if at least 85% of its exons have a coverage higher than or equal to a predetermined threshold. This transcript-specific threshold was set at an intuitive and robust value, namely, its mean exonic footprint coverage divided by 5. This excludes non-translated transcripts isoforms as well as allows (to some extent) possible variability in the ribosomal footprint density of real translated transcripts. Only transcripts that hold a fairly uniform footprint density throughout their CDS are subsequently classified as truly translated.

TIS calling

The mapped profiles from the initiating ribosomes, obtained after harringtonine (HARR) or lactimidomycin (LTM) treatment, are accumulated at AUG or near-cognate start codons using a ±1 nt window, hence tackling the subcodon resolution issue (8,10). Profiles that do not map within this window relative to the first position of a start codon are disregarded during TIS calling. These accumulated peak positions have to comply with a number of criteria in order to be withheld as a true translation start site (10): (i) the identified TIS should have the maximal number of reads (HARR and/or LTM) within a window of 7 nucleotides (i.e. one codon up- and downstream), (ii) the combined number of ribosome profiles for the TIS should exceed a minimal profile count threshold and (iii) the TIS should have a $R_{LTM/HARR} - R_{CHX}$ value equal or higher than a certain threshold, where

$$R_k = (X_k/N_k) \times 10 \quad (k = LT\text{ Mor HARR, CHX})$$

$$X_k = \text{number of reads on position} \ X \ \text{for data} \ k$$

$$N_k = \text{total number of reads on transcript for data} \ k$$

We opted for a categorized approach based on TIS localization: 5' UTR, aTIS, CDS, 3' UTR and no translation (TIS within non-protein-coding transcripts), aTIS that do not comply with the aforementioned criteria are also
taken into account if the Ensembl transcript shows elongating ribosome occupancy. Hence, aTIS identifications are further divided in three subcategories; (i) those demonstrating accumulated TIS LTM/HARR coverage and compliant with all rules (TRUE), (ii) those having accumulated TIS LTM/HARR coverage, but not compliant with all rules (FALSE) and (iii) those without accumulated TIS LTM/HARR coverage (NO DATA). TIS in the other four categories that do not comply with these rules are discarded.

**SNP calling**

Variants are extracted from the mapped RIBO-seq reads using SAMtools (24) (v.0.1.19) and by comparing the read mismatches to the NCBI dbSNP (25) data (build 137). The Picard toolkit (v.1.102: http://picard.sourceforge.net) is used to remove duplicates. Next, the variants are extracted using SAMtools mpileup coupled to BCFtools and the vcftools.pl tool (both part of the SAMtools toolkit). To reduce the chances of missing variants with SAMtools, we also compare every mismatch in the mapped reads to the variants in dbSNP and any mismatch found in dbSNP is retained in the final set of variants. To keep the size of the search database manageable, the number of dbSNP-matched mismatches is calculated per transcript (based on Ensembl annotation release 72 and 70 for mouse and human, respectively) and whenever this number is higher than five, the mismatches in this transcript are removed from the final variant list.

**Translation assembly; PROTEOFORMER-DB construction and integration with a canonical protein database (e.g. Swiss-Prot)**

Fast assembly of the translated sequences is made possible by a binary reading technique; fetching the CDS exon sequences from the corresponding chromosome sequence files (available from the iGenome repository). The aforementioned proteoform information (transcript isoform, TIS, SNP) is translated into the resulting amino acid sequence. Noteworthy is that only information on non-synonymous variations is presented in the translation product description. A custom, non-redundant translation product database for MS/MS-based protein/peptide identification is generated in FASTA format. The transcripts can be mapped to a known canonical protein database (e.g. Swiss-Prot) either by using the Biomart framework (26) (ID-based mapping) or by Basic Local Alignment Search Tool (BLAST) searching (sequence-based mapping).

Redundant sequences are eliminated based on the ranking of the annotations (in decreasing order of likeliness aTIS, 5' UTR, CDS, 3' UTR, no translation). If two or more transcript IDs have the same sequence, the transcript ID with the most plausible annotation is retained. If SNP information is included and two transcripts have the same annotation type and sequence, then the transcript with SNP information is retained. If two or more sequences satisfy all the constraints then one is chosen randomly. All sub-sequences (i.e. sequences completely contained in another sequence) are also eliminated from the database.

The ID-based mapping only considers those transcripts with annotation types aTIS or 5' UTR transcripts, the other annotation types (CDS, 3' UTR, no translation) are mapped by BLAST search. The ID-based mapping option simply maps a given Ensembl transcript ID to a corresponding canonical ID using the Biomart framework. If two or more transcripts have the same sequence (then the transcript with an existing canonical ID (Biomart-mapped) is retained. If two or more transcripts have a Biomart mapping then one with a higher annotation ranking is retained. The transcripts without any pre-mapped ID could then be mapped by the sequence-based methods. In the sequence-based mapping, redundant transcripts are removed based on their annotation ranking and length. If two transcripts have the same sequence then the one with the most highly ranked annotation is retained and subsequently all sub-sequences are removed. The non-redundant sequences can then be mapped to known canonical proteins by performing a BLAST search against the canonical protein database (e.g. Swiss-Prot).

**PROTEOFORMER implementation**

All information on the different implementations of the PROTEOFORMER method is available via http://www.biobix.be/proteoformer. A script-based (Perl 5) version and a Galaxy instance implementation are made available for download. These can respectively be deployed on a Unix system and implemented on a Galaxy instance (27). Furthermore, a customized virtual machine (Ubuntu 12.04 LTS) with all script dependencies and a Galaxy server already installed can be downloaded. A manual describing the aforementioned implementations (including prerequisites and dependencies) is made available on the website and as Supplementary Files S2 and S3.

**Supplementary methods**

Additional information on the experimental procedures, MS data analysis and correlation analysis can be found in Supplementary Methods S1.

**RESULTS**

In order to test the performance of the PROTEOFORMER method, we optimized different modules (mapping, TIS calling and SNP analysis) specifically toward the creation of a protein-synthesis based sequence database, using available mouse embryonic stem cell (mESC) RIBO-seq data (8). Matching shotgun and N-terminal COFRADIC (28) proteomics data served to evaluate this setup. While the former proteomics strategy gives a global assessment of the expressed proteome, the latter technique enables the isolation of N-terminal peptides, making it very appropriate for the validation of the by RIBO-seq observed (alternative) TIS.

**Optimization**

Two different mapping tools (STAR (20) and TopHat2 (21)) were evaluated and both performed similarly in terms of the percentage of reads mapped onto the reference genome (Supplementary Table S2). However, STAR was selected for
the rest of the analysis because it aligned slightly more relevant RPFs (i.e. with length between 26 and 34 bases), providing an increase of 2.8% and 4.6% for RPF of elongating and initiating ribosomes, respectively (Supplementary Figure S3). It also outperformed TopHat2 in terms of speed.

To optimize the PROTEOFORMER TIS calling algorithm for aTIS transcripts, we varied the two main TIS calling parameters: i.e. the minimum profile count (min count) and the difference in the normalized reads between the treated and untreated samples ($R_{UTM/HRR}$ - $R_{CHX}$). By varying the min count and $R_{UTM/HRR}$ - $R_{CHX}$ values we evaluated their impact on the downstream peptide identification rates on the mESC data. To do this we compiled non-redundant tryptic peptide search spaces for a range of different $R_{UTM/HRR}$ - $R_{CHX}$ (0.01–0.15) and min count (1–20) values and used these for spectral matching and database searching. The best TIS calling parameters were selected based on the number of confident tryptic peptides identified at a PEP (Posterior Error Probability) cutoff of 0.2 as this corresponds to an False Discovery Rate (FDR) of 1% (Figure 2a and b). With the min count set to 5 and the $R_{UTM/HRR}$ - $R_{CHX}$ values varying from 0.01 to 0.15, we observed that as the $R_{UTM/HRR}$ - $R_{CHX}$ value decreased the number of identified peptides increased and converged to a maximum. Below an $R_{UTM/HRR}$ - $R_{CHX}$ value of 0.01 the number of identifications started decreasing indicating that more noise was allowed into the data and that it became difficult for the peptide identification algorithm to clearly distinguish the good hits from the bad ones. This was also observed below a value of 5 when the min count varied from 1 to 20 while setting the $R_{UTM/HRR}$ - $R_{CHX}$ value fixed at a constant value of 0.01. For these reasons, a combination of $R_{UTM/HRR}$ - $R_{CHX}$ = 0.01 and min count = 5 was used for further analysis of aTIS transcripts. Furthermore, the rule-based TIS calling clearly outperformed a Support Vector Machine (SVM) algorithm (8) in compiling a comprehensive list of TISs in our setup (Figure 2c).

For other TIS categories, more stringent threshold settings were used in order to limit the amount of false-positive RIBO-seq-derived transcripts. This is especially important for downstream CDS TISs (using a rule-based TIS calling approach), as this region is very prone to false positives because of high ribosomal occupancy levels. However, excluding less stringent CDS TISs does not have a great impact on the final protein sequence database. During translation assembly, and in order to eliminate redundancy, the majority of CDS TISs-based transcripts are removed anyway (see Materials and Methods). Moreover, a more stringent approach for non-annotated TISs also ensures that TISs, and subsequent transcripts, that still pass parameter settings have a much greater chance to be true positives and are definitely worth further investigation. Thresholds for a TIS located in the 5' UTR were set to 10 (min count) and 0.05 ($R_{UTM/HRR}$ - $R_{CHX}$); for a TIS located in the downstream CDS 15 and 0.15 were used; for a TIS within the 3' UTR or a TIS within a non-protein-coding transcript these thresholds were set to 10 and 0.05.

With these optimal parameters identified we then generated varying non-redundant tryptic peptide databases based on inclusion of non-synonymous mutation information obtained from the RIBO-seq data using different strategies. These databases were compared alongside a tryptic peptide database generated from the mouse Swiss-Prot protein sequences. A search space built from the combination of RIBO-seq-derived sequences with mutation information derived from SAMtools and Swiss-Prot performed better than one derived from RIBO-seq, Swiss-Prot and mutation information from dbSNP (25) in terms of the number of tryptic peptide identifications (Figure 2e). This indicated that SAMtools is able to capture mutation information brought about by RIBO-seq, which is lacking in dbSNP. These settings also proved optimal in analyzing the human colorectal cancer cell line (HCT116) RIBO-seq data (Figure 2d).

Evaluation on mESC and HCT116 cell line material

To evaluate the deep proteome coverage of the PROTEOFORMER pipeline, it was applied to the mESC and HCT116 RIBO-seq data sets. Combining the RIBO-seq-derived protein sequences with Swiss-Prot (mouse and human individually), 3771 mouse and 2853 human protein identifications were obtained from the shotgun experiments at a 1% FDR threshold (Figure 3a and b and Supplementary Table S1a and b). The supplemental (RIBO-seq-derived) sequences in the search space contributed to respectively 323 and 20 (8.6% and 0.7%) new and 124 and 65 (3.3% and 2.3%) improved protein identifications for the mouse and human data sets. These so-called new identifications were not contained in Swiss-Prot and originated from peptide identifications that (partly) overlapped an N-terminal extension, an exon region of an alternative spliced isoform, a mutation site or alternatively, an upstream open reading frame (uORF) (Figure 3a and b and Supplementary Figure S4a and b). Due to the increased protein coverage, these phenomena also accounted for a substantial increase of identifications with an improved protein score.

Correlation of the translational outcome based on ribosome profiling (RPF count) with the label-free protein abundance measures of the shotgun experiments (emPAI and NSAIF) demonstrated that these technologies are highly complementary. Positive Pearson's correlation coefficients reaching up to 0.714 and 0.643 were obtained for mouse and human (18), respectively (Supplementary Figures S5 and S6), exceeding the correlation of the same MS spectral count-based measures with mRNA FPKM counts (1,6,29,30).

The N-terminal COFRADIC experiments resulted in the identification of different classes of N-termini (Figure 3c and d and Supplementary Table S1c and d). The majority of peptides mapped canonical start sites or Swiss-Prot database annotated TIS (dtTIS): 1346 mouse and 1089 human N-termini (i.e. 84.7% and 83.0% of all identified N-termini), 223 and 213 (14.0% and 16.2%) started downstream of the annotated TIS (dtTIS: past protein position 2 in reference to Swiss-Prot). Interestingly, 18 and 11 peptides pointed to N-terminally extended proteoforms in mouse and human. Another two N-terminal peptides pointed to the translation of uORF completely within the 5' UTR or out-of-frame and overlapping with canonical CDS for mouse. Moreover, analysis of N-terminal COFRADIC data using the PROTEOFORMER pipeline
Figure 2. PEP distributions of the number of identified tryptic peptides from shotgun proteome analyses. The searches were performed on a database holding a non-redundant set of tryptic peptides based on the RIBO-seq-derived sequences having annotated TIS (eTIS). These plots demonstrate the impact of the database creation parameters of PROTEOFORMER on downstream MS/MS identification. The cumulative number of peptides identified is plotted on the y-axis and the corresponding PEP (i.e. the probability that a peptide-to-spectrum match is a chance event) is plotted on the x-axis. (a) mESC shotgun data: The effect of changing the $R_{\text{UM} \text{t} \text{M} \text{HARR}} - R_{\text{UMT} \text{HARR}}$ in the TIS calling procedure on the number of tryptic peptides identified with 'minimum profile count' (TIS calling) set to 10. The number of identified tryptic peptides decreases with increasing $R_{\text{UM} \text{t} \text{M} \text{HARR}} - R_{\text{UMT} \text{HARR}}$ value. There is a marked increase in the number of highly confident matches (for PEP < 0.2) at lower values of $R_{\text{UM} \text{t} \text{M} \text{HARR}} - R_{\text{UMT} \text{HARR}}$. (b) mESC shotgun data: The effect of 'minimum profile count' on the number of identified tryptic peptides at constant $R_{\text{UM} \text{t} \text{M} \text{HARR}} - R_{\text{UMT} \text{HARR}}$ of 0.01. The number of highly confident identifications decreases with increasing number of 'minimum profile count'. At a confidence of about 80% (PEP < 0.2) the number of identified peptides is about the same for 'minimum profile count' 1 and 5. (c) mESC shotgun data: Comparison of the peptide identification numbers using different database versions. From the PEP distributions it is clear that searches using the RIBO-seq-derived database outperformed those using solely Swiss-Prot. With SNP information (RIBO-seq (SAMtools)) included, the number of identification increases even more, with the best result obtained using a search space combining RIBO-seq-derived sequences (SNP information inclusive) and Swiss-Prot entries at an 80% confidence validation threshold. It is also clear that the rule-based algorithm outperformed the SVM algorithm applied in Ingolia et al. (8). (d) HCT116 shotgun data: The number of peptide identifications using only RIBO-seq-derived sequences as a search space is lower than searching Swiss-Prot. Yet a significant increase is notable when searching against a combined database (RIBO-seq derived + Swiss-Prot).
Figure 3. PROTEOFORMER enables deep proteome coverage. Pie charts representing the number of protein and peptide identifications obtained from the shotgun proteomics and N-terminal COFRADIC experiments based on searching the PROTEOFORMER + Swiss-Prot database for both mouse ESC cells and human HCT116 cells using a 1% FDR threshold. (a) Shotgun proteomics results (mouse). A total of 377 proteins were identified. (b) Shotgun proteomics results (human), identifying a total of 2853 proteins. For the shotgun experiments, a categorization was made based on the fact that the protein can be picked up using the PROTEOFORMER and/or Swiss-Prot sequence database. Also, the improved and new protein identifications were further classified into the following categories: new, isoform/homolog, SNP variant, 5' extension and uORF. (c) N-terminal COFRADIC (mouse) experiment resulting in 1589 N-terminal peptide identifications. (d) N-terminal COFRADIC results (human). Here, 1312 N-termini were identified. The N-termini were categorized as either dTIS (database annotated TIS), dTIS (downstream TIS), 5' extension or uORF.

provided us with evidence of translation initiation at near-cognate start sites (non-AUG codons recoded to initiator methionines). Peptide-to-spectrum matches (PSMs) corresponding to peptides located in a uORF region were manually validated (Supplementary File S1) and possibly hint at true translation of these uORFs, although it cannot be ruled out that an unpredicted extended proteome exists comprising this translated uORF sequence.

Interestingly, refined gene models can be built based on novel peptide identifications resulting from our PROTEOFORMER approach. These can be categorized into new exons (pointing to new isoforms, see Supplementary Table S1a and b), N-terminal extensions (see Supplementary Table S1 and examples of the human dcaf12 and the orthologous hdaa1 gene illustrated in respectively Figure 4a and Supplementary Figure S7) and translation of uORFs (see the example of an uORF contained in the S3a:35a gene shown in Figure 4b). This uORF could also be categorized as a new gene product (resulting in a translation product of 103 AA, see Supplementary File S1). These findings suggest that the PROTEOFORMER approach can help to refine the annotation of the genome.

DISCUSSION

PROTEOFORMER is the first publicly available analysis pipeline that provides a complete bioinformatics workflow for the analysis of RIBO-seq NGS data and that enables the construction of a customized protein sequence search space to allow integration with MS facilitating the capture of the proteome complexity. By combining the information from elongating and initiating ribosomes, it is able to create an optimal search space for matching MS experiments. The integration of PROTEOFORMER within the Galaxy framework provides a user-friendly interface for analysis of RIBO-seq data (in combination with proteomics data), resulting in new and improved identifications.

Noteworthy are the overall lower identification rates for the human sample. This can be attributed to (i) the fact that only heavy labeled peptides were considered in the human MS setup (Supplementary Methods S1), (ii) the overall better annotation of the human proteome (represented by the lower number of lower non-Swiss-Prot identifications) and (iii) the higher number of identifications not present in our RIBO-seq-derived sequence pool (i.e., identifications matching Swiss-Prot entries only) for the human sample. Whereas only 148 (3.9%) identifications are not captured based the RIBO-seq strategy for the mouse data, this number increases to 253 (8.9%) for the human data. Inspection of the metagenic RPF abundance plots (Supplementary Figure S2) shows an expected dynamic range of expression. The quantitative correlation between RPF abundance and spectral count-based measures for the non-custom Swiss-Prot proteins (Supplementary Figure S8) demonstrates that this lower performance is not attributable to the CHX-treated HCT116 sample sequencing coverage. Finally, the distribution of the RmRhoR - RmRho values (used in the TIS calling procedure, see Materials and Methods and Supplementary Figure S9) pointed to an overall lower genome-wide coverage of initiating ribosomes, attributable to either biases introduced in the library preparation of the LTM-treated HCT116 sample or suboptimal conditions of the LTM treatment. Consequently, proteomics enables a quality assessment of RIBO-seq, which is typically lacking.

RIBO-seq-based studies also showed ribosome occupancy of long ncRNAs (IncRNAs) (8), possibly hinting towards their protein coding potential. However, most IncRNAs do not function through encoded proteins (31) demonstrating that RIBO-seq on its own is not a perfect proxy for protein synthesis and that MS validation is often indispensable (13). New RIBO-seq approaches as the Fragment Length Organization Similarity Score (32) and Ribosome Release Score (33) in combination with MS validation using, for example, PROTEOFORMER, will prove very useful in RIBO-seq-based protein identification (18).

By increasing the size of the sequence search space (e.g., a database derived from a six-frame translation of nucleotide sequences (based on mRNA-seq)), MS database search engines will underestimate the confidence assigned to the PSMs leading to fewer identifications at the estimated FDR and PEP thresholds (34) using a typical target-decoy approach. PROTEOFORMER only requires one-reading-frame translation in contrast to methods based on regular mRNA sequencing, thus limiting the search space explo-
Figure 4. Examples of new proteoforms. (a) N-terminal extension of DCA13_HUMAN and (b) translated uORF of S35A4_MOUSE that were picked up by the proteogenomics analysis and validated by N-terminal COFRADIC. The UCSC genome browser was used to create a view of the RIBO-seq and COFRADIC data. The different tracks are from top to bottom: CDX-treated RIBO-seq data, LTM/HARR-treated RIBO-seq data, N-terminal COFRADIC data, UCSC genes, RefSeq genes and mRNA. The zoomed-in images show the alternative start site (i, alternative start site; ii, canonical start site), while the MS/MS spectra and sequence fragmentation plots display the confidence and quality of the N-terminal peptide identifications.
sion and keeping the confidence distribution of the search against the PROTEOFORMER database similar to standard Swiss-Prot searches (Figure 2c and d). We also envision that more efficient MS scoring algorithms (9) will be set in place to even better cope with the increasing search space sizes inherent to next-generation sequencing-based methods.

Through user-definable parameter settings, PROTEOFORMER provides the flexibility to tailor the creation of a translatome-based sequence database to the research question at hand. Downstream TIS identification or unbiased TIS calling are, for example, possible, but would need appropriate optimization for the different TIS categories. PROTEOFORMER makes use of iGenomes reference sequences and annotation from Ensembl for mapping, and custom Ensembl SQLite annotation databases (available on the PROTEOFORMER web page). It can already handle Ribo-seq-derived sequencing data of *M. musculus*, *Homo sapiens*, *Drosophila melanogaster* and *Arabidopsis thaliana*, and we are currently working on incorporating other species. This is done on a case-by-case basis as species-specific adaptations, for example, to RPF parsing (23), are often desired. Furthermore, we are also continuously improving our pipeline including other TIS calling algorithms (8), SNP calling tools (35) and Ribo-seq specific measures (32,36).

In conclusion, we developed a new analysis pipeline, termed PROTEOFORMER. It enables the processing of Ribo-seq data and can be optimized based on user-definable parameter settings in order to be useful in answering a plethora of different research questions. The tool includes a mapping module enabling genome-wide visualization of ribosome occupancy on a genome browser of choice. It also includes a TIS calling algorithm that allows for the delineation of the ORFs of all translation products, based on initiating ribosome footprint accumulation obtained upon LFM/HARR treatment. A complete translatome-based sequence database, also including SNP information, can thus be compiled, for spectral database matching. We further showed that optimization toward the use of PROTEOFORMER in a proteogenomic approach, enables deep proteome coverage (including S′ extended proteoforms, alternative spliced isoforms and uORFs) resulting in an increase in overall protein identification rate when searching matching MS data sets.

A stand-alone version (Supplementary File S2) and a galaxy implementation (Supplementary File S3 and Supplementary Figure S10) of our approach are available at http://www.biobix.be/proteofomer next to all relevant information on the installation and underlying requirements.

**SUPPLEMENTARY DATA**

Supplementary Data are available at NAR Online.

**ACKNOWLEDGEMENTS**

**Authors Contributions:** G.M. and P.V.D. designed research; G.M., J.C., E.N., A.K. and S.S. developed the PROTEOFORMER script-based version and Galaxy implementation; G.M. and E.N. analyzed proteomics data; G.M. and J.C. analyzed ribosome profiling data; D.G., P.V.D., S.D.K. and E.D.M. performed ribosome profiling; P.V.D. performed proteome analyses; J.C., E.N. and G.M. wrote the paper; E.N. created the website and virtual machine; P.V.D. and G.M. supervised the research. W.V.C. and T.D.M. advised on research. All authors read and approved the final manuscript.

**FUNDING**

Postdoctoral Fellows of the Research Foundation - Flanders (FWO-Vlaanderen) [to P.V.D.]. The Research Foundation - Flanders (FWO-Vlaanderen) [G.0269.13N to P.V.D.]. PhD grant from the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen) [to J.C. and D.G.]. ‘Vlaamse Liga tegen Kanker’ [to A.K.]. IUAP [to S.S.]. Funding for open access charge: Research Foundation - Flanders (FWO-Vlaanderen), Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen).

Conflict of interest statement. None declared.

**REFERENCES**


ACKNOWLEDGEMENTS
ACKNOWLEDGEMENTS

Eventually there is a moment to look back and acknowledge everyone who helped me.

First and foremost, enormous gratitude is due to Prof. Petra Van Damme – my promoter and supervisor for the past 5 years who has been outstanding in her support and guidance. This work wouldn’t be possible without you. You are brilliant and forever my inspiration.

I pass my warmest thanks to Prof. Kris Gevaert who entrusted me with this PhD project and provided countless opportunities for me to grow as a scientist. It was a privilege to work with you. Thank you for giving me the chance to come to Belgium, it was an amazing experience.

I would like to thank the members of my PhD jury - Prof. dr. Elfride De Baere, Dr. Maarten Dhaenens, Prof. dr. Kim De Keersmaecker, Dr. Audrey Michel, Prof. dr. ir. Katleen De Preter, Prof. dr. ir. Jo Vandesompele - for devoting their time and effort to critically evaluate and improve my thesis.

I owe a great thank you to the members of our group. Veronique, it was a pleasure to work with you at the bench. Thank you for your invaluable help, advice, humour and friendship. Elvis, I am really grateful for your support and stimulating discussions we shared. I wish you all the success with your PhD thesis! Thank you Sebastian for being rational, thoughtful, kind and always helpful. Ursula, thank you for being a great friend, I wish you and Ingacio all the best, you are very talented and surely you have great scientific careers ahead.

I am greatly indebted to our wonderful collaborators at BioBix. Thank you Gerben, Steven, Alexander, Jeroen, Sandra and Volodimir for the enormous help throughout the years. I think we made together a great interdisciplinary team. I learned so much from you, thanks!

Words of gratitude to all my colleagues for creating an amazing atmosphere at work. Thank you Masa and Ludger for the adventures and conversations we shared, thank you for words of encouragement, good luck with you PhDs! Giel, Kevin, Sven, Francis, Eva, Emmy, Sriram, Luis, Fabien - I keep my fingers crossed for your future scientific endeavours – I am convinced you will do great. An, Evy, Jarne, Delphi, Jonathan, Pieter-Jan, I really appreciate the time and expertise you devoted to helping me with proteomics experiments.

Warm thanks to all the (ex-) colleagues I got to share my VIB experience with. Silke, Elisabeth, Kim, Delphine, Noortje, Michela, Alan, Hans, Bart R., Mathias, Arun. You make this place special, I will miss you guys.

Belgium holds a special place in my heart for letting me meet some wonderful people. Thank you Ania T. for your unconditional support and friendship, what would I do without you? Thank you to Ania Ch., Ania Ś., Asia, Kasia Ch., Kamila, Ilse, Elke, Bram, Simon, Tom, Mumin and Radek, I feel lucky to call you my friends.
Thank you Kasia W. - some friendships last forever. I wish you and James all the best.

Najważniejsze podziękowania zostawiam na koniec. Dziękuję Mamo i Tato za nieprzerwane wsparcie, waszą miłość i mądrość. Obdarzyliście mnie zaufaniem i wolnością by rozwijać swoje pasje. Tyle wam zawdzięczam.

Dziękuję Babci Uli i Lidii za wsparcie i kochające serca. Dziękuję, że mogę na was liczyć!

Dziękuję Marzenie, Robertowi, Zofii i Dominikowi Fijałkowskim za ciepło i nieustające słowa otuchy, jesteście nieocenieni.

Dziękuję Igorowi - mojemu kochanemu mężowi a zarazem najlepszemu przyjacielowi i największemu sprzymierńcowi na świecie. Za wszystko.

Thank you to all the readers of this thesis. It was my pleasure to present this work to you.