Article

Found in translation: Microproteins are a new class of potential host cell impurity in mAb drug products

Marina Castro-Rivadeneyra[1,2†], Ioanna Tzani[1†], Paul Kelly[1], Lisa Strasser[1], Felipe Guapo[1], Ciara Tierney[1], Lin Zhang[3], Martin Clynes[4], Barry L. Karger[5], Niall Barron[1,2], Jonathan Bones[1,2†] and Colin Clarke[1,2†*].

Affiliations

[1]National Institute for Bioprocessing Research and Training, Fosters Avenue, Blackrock, Co. Dublin, Ireland.

[2]School of Chemical and Bioprocess Engineering, University College Dublin, Belfield, Dublin, Ireland.

[3]Bioprocess R&D, Pfizer Inc. Andover, Massachusetts, USA.

[4]National Institute for Cellular Biotechnology, Dublin City University, Dublin 9, Ireland.

[5]Barnett Institute, Northeastern University, 360 Huntington Ave, Boston, Massachusetts 02115, USA.

[†]Equal contribution

*Corresponding author

Correspondence

Email: colin.clarke@nibrt.ie

Phone: +353 1 215 8164

Fax: +353 1 215 8116

Keywords:

Ribosome footprint profiling; Chinese hamster ovary cells; Biopharmaceutical manufacturing; Translational regulation; Ribo-seq; Host cell protein; Short open reading frame; Upstream open reading frame; translation; Microprotein;

Abbreviations: CDS, coding sequence; CHO, Chinese hamster ovary; CHX, cycloheximide; Harr, Harringtonine; HCP, host cell protein; mAb, monoclonal antibody; NGS, Next generation sequencing; NTS, non-temperature shifted; ORF, open reading frame; ouORF, overlapping upstream ORF; PAGE; polyacrylamide gel; Ribosome footprint profiling, Ribo-seq; RPF, Ribosome protected fragment; RPKM, Reads per kilobase mapped; sORF, short open reading frame; TS, temperature shifted; TE, Translational efficiency; uORF, upstream open reading frame; UTR, untranslated region; BPM: Bins per million; AGC, Automatic Gain Control; GO: Gene Ontology; LFQ, Label Free Quantification; DDA, Data Dependent Acquisition; IT, Injection Time;

Highlights

- Analysis of translation initiation and elongation using ribosome footprint profiling provides a refined annotation of the Chinese hamster genome.
- 7,769 novel Chinese proteoforms were identified including those initiating at near cognate start codons.
- 941 N-terminal extensions of annotated genes were identified.
- 5,553 short open reading frames (sORFs) predicted to encode microproteins (i.e., proteins < 100 aa) were also characterised.
- The annotation of non-canonical proteins increases the coverage of MS-based host-cell protein analysis in monoclonal antibody drug products.
- 8 microproteins were found in adalimumab drug product.
- Transcripts annotated as non-coding can contain short open reading frames (sORFs) predicted to encode peptides (or microproteins) which are found to undergo changes in expression and translational regulation at reduced cell culture temperature.
- 95 of the novel proteoforms of which 79 were microproteins were subsequently identified in a second CHO K1 cell line using LC-MS/MS based proteomics. A comparison of protein abundance revealed that 13 microproteins were found to be differentially expressed between the exponential growth and stationary phases of cell culture.

Abstract

Mass spectrometry (MS) has emerged as a powerful approach for the detection of Chinese hamster ovary (CHO) cell protein impurities in antibody drug products. The incomplete annotation of the Chinese hamster genome, however, limits the coverage of MS-based host cell protein (HCP) analysis. In this study, we performed ribosome footprint profiling (Ribo-seq) of translation initiation and elongation to refine the Chinese hamster genome annotation. Analysis of these data resulted in the identification of thousands of previously uncharacterised non-canonical proteoforms in CHO cells, such as N-terminally extended proteins and short open reading frames (sORFs) predicted to encode for microproteins. MS-based HCP analysis of adalimumab with the extended protein sequence database, resulted in the detection of CHO cell microprotein impurities in a mAb drug product for the first time. Further analysis revealed that the CHO cell microprotein population is altered over the course of cell culture and in response to a change in cell culture temperature. The annotation of non-canonical Chinese hamster proteoforms permits a more comprehensive characterisation of HCPs in antibody drug products using MS.

1. Introduction

Chinese hamster ovary (CHO) cells are the predominant mammalian expression host for the production of therapeutic monoclonal antibodies (mAbs), with more than 80% of the new mAbs approved between 2014-2018 manufactured in CHO cell lines (Walsh, 2018). During the cell culture phase of production, CHO cells continually secrete mAb into the supernatant. A series of downstream purification steps are required to recover the product in the harvested cell culture fluid and reduce a range of impurities originating from the host CHO cell line. Host cell proteins (HCPs) present in the final drug product are a particular concern, due to the risk that a HCP could elicit an immune response in the patient or reduce efficacy (Hanania et al., 2015). In addition, the presence of proteolytic HCPs can degrade or affect the stability of the mAb (Li et al., 2021; Luo et al., 2019). Regulatory authorities consider the amount of HCP in the final product to be a critical quality attribute, and require that the total HCP concentration be < 100 ppm (Bracewell et al., 2015). Enzyme-linked immunosorbent assays (ELISA) are typically used for HCP analysis, enabling sensitive quantitation and reasonable throughput. Such assays can be limited in terms of coverage in that those HCPs that are weaky immunogenic or do not elicit an immune response in the species used to generate the antibodies will not be detected (Henry et al., 2017).

Mass spectrometry (MS) has emerged as a complementary HCP detection method (Bracewell et al., 2015) capable of identifying individual HCPs, even those at low concentrations. These data can be used to understand HCP clearance at each stage of downstream purification (Huang et al., 2021), and characterise the populations of HCPs present in different cell culture conditions (Goey et al., 2018). Knowledge of the HCP population can be used to guide process optimisation, or identify targets for cell line engineering to remove unwanted HCPs (Chiu et al., 2017). The publication of the first CHO cell genome (Xu et al., 2011) and availability of CHO cell-specific protein databases (Meleady et al., 2012) have significantly improved the detection of CHO HCP impurities in mAb drug product using MS. The quality of available genomes has steadily improved over time and, with the release of the Chinese hamster PICR-H genome, the field now has a reference assembly comparable to that of model organisms (Hilliard et al., 2020). While annotation of the transcriptome has progressed significantly, characterisation of the proteome is more challenging and remains incomplete, therefore limiting the ability of MS to detect the entire spectrum of HCP impurities.

Until recently, the Chinese hamster proteome was annotated via a combination of *ab initio* computational pipelines, homology, ESTs, and transcriptomics data. The Lewis lab elegantly demonstrated that ribosome footprint profiling (Ribo-seq) can be used to improve annotation of the Chinese hamster genome (Li et al., 2019). Ribo-seq enables transcriptome-wide determination of ribosome occupancy at single nucleotide resolution enabling open reading frame (ORF) annotation, and, when combined with RNA-seq, variations in translational regulation (Ingolia et al., 2009). The technique utilises chemical or physical inhibitors to arrest translation and fixes translating ribosomes in position resulting in the protection ~30 nt of RNA within the ribosome from subsequent enzymatic degradation. The resulting monosomes are purified via sucrose gradient, sucrose cushion or size exclusion chromatography, followed by the isolation of ribosome protected fragments (RPFs) through

size selection, from which a sequencing library is prepared. Sequencing of RPFs and alignment to a reference genome or transcriptome permits the identification and quantitation of regions undergoing active translation.

Over the last decade Ribo-seq has provided compelling evidence that the traditional rules of eukaryotic translation need to be revised. For example, translation initiation at near-cognate codons (CUG, GUG, UUG) is more widespread in mammalian genomes than previously thought (Wright et al., 2021). The analysis of Ribo-seq data has also been essential for the characterisation of a range of non-canonical ORFs, including N-terminal extensions (Ivanov et al., 2011), detecting translation of RNAs previously thought to be non-coding (Ji et al., 2015), and uncovering the regulatory role of ORFs initiating in the 5' leader sequence of mRNAs (i.e., upstream open reading frames) (Zhang et al., 2021). Ribo-seq has also revealed the existence of small open reading frames (sORFs) that produce potentially functional microproteins (classified as proteins < 100 aa) in a diverse range of organisms including *Drosophila* (Aspden et al., 2014), zebrafish (Bazzini et al., 2014), mouse (Ingolia et al., 2011) and human (Chen et al., 2020; Martinez et al., 2020). Studies have so far shown that specific microproteins play a role in a variety of cellular processes such as oxidative phosphorylation (Zhang et al., 2020), mitochondrial translation (Rathore et al., 2018), metabolism (Lee et al., 2015), DNA repair (Slavoff et al., 2014) and can also act as transcription factors (Koh et al., 2021).

In this manuscript, we present a further refinement of the Chinese hamster genome annotation using Ribo-seq to increase the coverage of MS-based HCP identification. The reduction of cell culture temperature ("temperature shift") is a method used extend the viability of some commercial cell culture processes and improve product quality (Masterton and Smales, 2014). Here, we captured Ribo-seq data from a small-scale model of temperature shift to generate a database of new CHO cell proteoforms. A critical advance of this study is the use of multiple translation inhibitors for Ribo-seq to enable not only the capture of information on elongation, but also initiation in CHO cells. These data have enabled us to characterise non-canonical ORFs that begin at AUG and at near cognate start codons (i.e., CUG, GUG and UUG). We have identified a range of novel proteoforms of canonical protein coding genes (e.g., with N-terminal extensions), ORFs in non-coding transcripts, regulatory regions in the 5' leader sequence of mRNAs, and sORFs in CHO cells. The detection of sORF derived microproteins in a mAb drug product confirms that the extended proteome annotation enables more comprehensive MS-based HCP identification. Through the comparative analysis of the transcriptome, translatome and proteome, we further show that microprotein abundance is altered over the course of cell culture and upon alteration of the bioreactor temperature. These results indicate that cell culture optimisation could be used to reduce contamination from unwanted host cell microproteins.

2. Results

2.1 Transcriptome wide analysis of CHO cell translation initiation and elongation using Ribo-seq

Ribo-seq was performed for a monoclonal antibody producing CHO K1 cell line (CHO K1-mAb), previously shown by our laboratory to have decreased growth and altered extracellular lactate and ammonia profiles at sub-physiological temperature (Tzani et al., 2020). We utilised the small-scale cell culture model of temperature shift for this study and the resulting data was used to construct a CHO cell proteoform database, and subsequently for differential translation analysis (Section 2.5). To capture the Ribo-seq data we conducted two identical cell culture experiments for the analysis of translation initiation and elongation. For both experiments, 8 replicate shake flasks were initially grown for 48 hrs at 37°C before the temperature was reduced to 31°C (temperature shifted (TS) group; n=4) while maintaining the remainder at 37°C (non-temperature shift (NTS) group; n=4). Translation was arrested and samples were acquired for further analysis 24 hrs post temperature shift (Figure 1A) at which point there was an average decrease of 30% (initiation experiment) and 24% (elongation experiment) in cell density in the TS sample group (Figure S1; Table S1).

To capture a snapshot of the CHO cell translatome, we performed ribosome footprint profiling experiments using harringtonine (HARR) (n=8), an inhibitor for translation initiation (Ingolia et al., 2011), and cycloheximide (CHX) (n=8) an inhibitor for translation elongation (Figure 1C) (Ingolia et al., 2009). For each harringtonine sample, a matched Ribo-seq sample (n=8) was treated with DMSO and flash frozen to arrest translation (we refer to these data as “No-drug” (ND)). For the CHX samples, matched gene expression profiles were acquired using total RNA-seq (n=8) (Figure 1B) to determine the significant differences in translational efficiency (TE) between the NTS and TS sample groups. Sequencing of the 24 resulting Ribo-seq libraries yielded an average of ~68, ~67 and ~58 million reads across the 8 replicates for the CHX, HARR and ND Ribo-seq, respectively while an average of ~56 million reads per sample were obtained for the 8 RNA-seq libraries. Low quality reads were removed, and adapter sequences trimmed from the raw Ribo-seq and RNA-seq data (Table S2). For Ribo-seq data, an additional filtering stage was carried out to eliminate contamination from non-coding RNA. Reads were mapped to STAR (Dobin et al., 2013) indices constructed from *Cricetulus griseus* rRNA, tRNA and snoRNA sequences obtained from v18 of the RNA Central database (The RNAcentral Consortium, 2019). Reads aligning to any of these indices were discarded from further analysis. This filtering stage removed an average of ~40%, ~46% and ~50% of trimmed reads for the CHX, HARR and ND samples, respectively (Figure S2; Table S2).

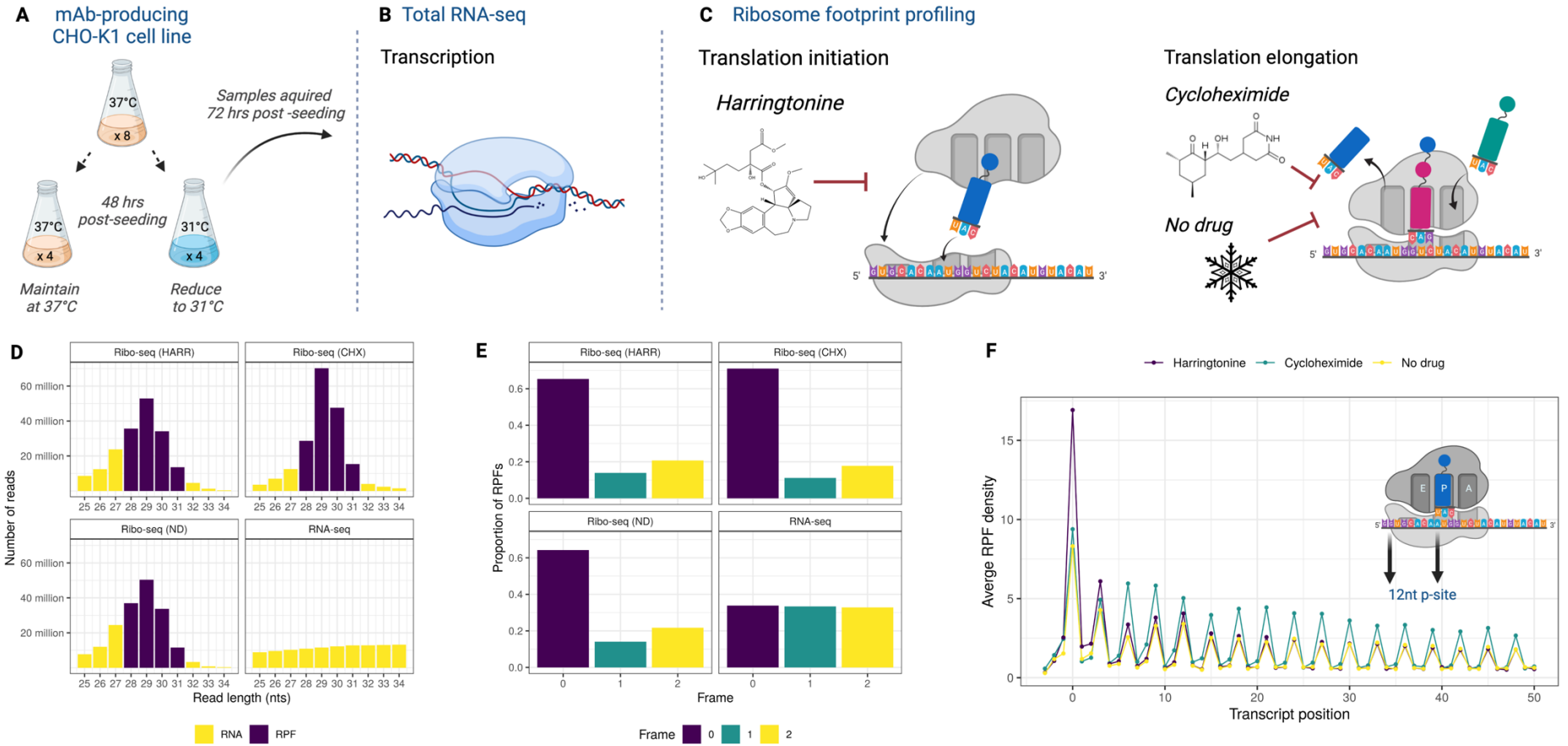


Figure 1: Analysis of sub-physiological temperature induced changes in CHO cell translation using ribosome footprint profiling. (A) 8 replicate shake flasks were seeded with a mAb producing CHO K1 cell line cultured for 48 hrs, at this point the temperature of 4 shake flasks was reduced to 31°C. At 72 hrs, samples were harvested from the non-temperature and temperature shifted cultures. We utilised **(B)** RNA-seq to characterise the transcriptome as well as **(C)** Ribo-seq using different inhibitors to monitor translation initiation (harringtonine) and elongation (cycloheximide and no drug). Following pre-processing of the raw Ribo-seq data, we **(D)** retained reads within the expected size range of RPFs. An optimum P-site offset of 12 nucleotides was selected for all datasets, where **(E)** an average of 60% of RPFs was found to exhibit the expected triplet periodicity. A metagene analysis was conducted for the three Ribo-seq datasets, confirming **(F)** the expected enrichment of RPFs at the TIS of annotated protein coding genes in harringtonine Ribo-seq data when compared to the cycloheximide and no-drug treated samples.

Next, we examined reads within the expected RPF length range (25-34nt), to select the P-site offset (the distance from the 5' end of a read to the first nucleotide of the P-site codon) (Figure 1D). Each Ribo-seq dataset was mapped to the Chinese hamster PICRH-1.0 genome using STAR (Dobin et al., 2013). Then, the Plastid tool (Dunn and Weissman, 2016) was used to assess the P-site offset and subsequently determine the proportion of reads exhibiting triplet periodicity for NCBI-annotated protein coding genes for each offset. Following this analysis, we retained the reads between 28-31 nt for further analysis (Figure 1E). The optimum P-site offset was found to be 12 nt, for which 60% of reads exhibited the expected triplet periodicity for each Ribo-seq dataset. Prior to running ORF-RATER for the *de novo* ORF identification, we confirmed the expected preferential enrichment of ribosomes at the translation initiation sites (TIS) of NCBI annotated protein coding genes for the harringtonine-treated Ribo-seq data in comparison to the cycloheximide and no drug treated Ribo-seq data (Figure 1F).

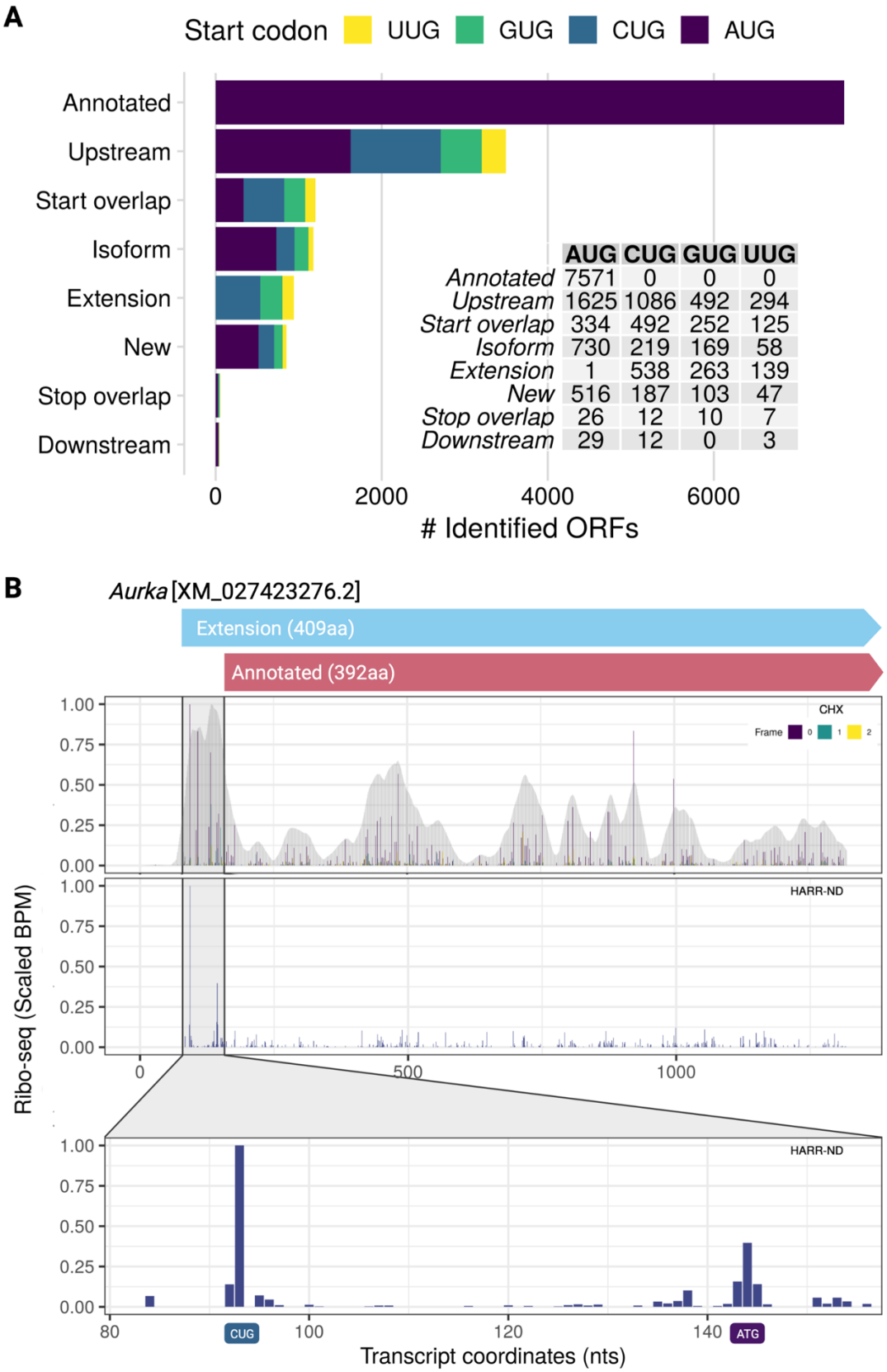
2.2 Ribo-seq enables the characterisation of novel Chinese hamster proteoforms

The Ribo-seq data was used to refine the annotation of translated regions of the Chinese hamster PICRH-1.0 genome by conducting a transcriptome wide analysis using ORF-RATER (Fields et al., 2015). The ORF-RATER algorithm integrates initiation and elongation Ribo-seq data to enable the identification of unannotated ORFs by first finding all potential ORFs beginning at user defined start codons with an in-frame stop codon *in-silico*. The experimental data is then used to confirm occupancy at each TIS and that the ORF is undergoing active translation. To maximise the sensitivity of ORF detection, we merged the RPFs for all replicates in each type of Ribo-seq experiment yielding a total of approximately 136, 161 and 132 million RPFs for the harringtonine, cycloheximide, and no-drug treated Ribo-seq, respectively. Prior to ORF identification, transcripts originating from 4,583 pseudogenes, transcripts with low coverage (n = 19,357) or where RPFs mapped to a small number of positions within the transcript (n = 1,538) were removed from further analysis. For the remaining transcripts, the initial ORF-RATER search was limited to ORFs that began at AUG or near cognate start codons (CUG, GUG and UUG). To determine if a potential TIS was occupied, only the RPF data from the harringtonine-treated Ribo-seq was considered while CHX and ND-treated Ribo-seq data was utilised to determine if putative ORFs were translated by comparing the RPF occupancy of each ORF to the typical pattern of translation elongation observed for annotated mRNAs.

An initial group of 26,606 proteoforms identified by ORF-RATER with an ORF-RATER score of ≥0.5 (Eisenberg et al., 2020; Finkel et al., 2021) and a length ≥ 5 aa was selected for further analysis. The proteoforms identified included those present in the current annotation of the Chinese hamster genome (i.e., Annotated) and N-terminal extensions (i.e., Extension). Two distinct classes of ORFs initiating upstream of the annotated CDS (i.e., the main ORF) were also identified. The first type, called upstream ORFs (i.e., uORFs) initiate upstream and terminate before the start codon of the main ORF. The second upstream ORF type, termed overlapping upstream open reading frames (ouORFs), also initiates in the 5' leader of mRNAs but extends downstream beyond the start codon of the main ORF and is therefore

translated in a different reading frame. We also identified ORFs in transcripts that had both unannotated start and stop codons in the PICRH-1.0 genome ("New" ORFs).

The conditions used to inhibit translation initiation can, in some cases, lead to the identification of false positive internal ORFs due to capture of residual elongating ribosomes (Eisenberg et al., 2020). In our case, we utilised flash freezing in combination with harringtonine, which will also result in the capture of a proportion of RPFs from elongating ribosomes, however, this will almost certainly lead to erroneous identifications. To reduce false positives from internal TIS, we excluded truncated ORF (n=8,856) and internal ORF (n=1,469) classifications from further analysis. In addition, where more than one upstream ORF (uORF), start overlapping uORF (ouORF) or "New" ORF had the same stop codon, we retained only the longest of these ORFs, resulting in the elimination of a further 941 ORFs. Following this stringent filtering process, 15,340 high confidence ORFs were retained (Figure 2A, Table S3), with 49.3% (n=7,769) of the identified ORFs not present in the Chinese hamster PICRH-1.0 annotation. 58% of these new identifications start at near cognate codons (i.e., CUG, GUG or UUG). The ability to identify initiation at non-AUG codons enabled us to identify alternative proteoforms of conventional protein coding genes that would not be possible with previous annotation approaches for the Chinese hamster genome. For instance, 12.1% (n=941) of novel ORFs identified were N-terminal extensions of annotated protein coding transcripts (e.g., Aurora kinase A (Figure 2B)).



	AUG	CUG	GUG	UUG
Annotated	7571	0	0	0
Upstream	1625	1086	492	294
Start overlap	334	492	252	125
Isoform	730	219	169	58
Extension	1	538	263	139
New	516	187	103	47
Stop overlap	26	12	10	7
Downstream	29	12	0	3

Figure 2: Ribo-seq identifies thousands of novel CHO cell proteoforms. In this study, we utilised the ORF-RATER algorithm to identify ORFs initiated at near cognate (i.e., NUG) start codons from the Ribo-seq data. A total of **(A)** 15,340 ORFs were identified including 7,769 that were not previously annotated in the Chinese hamster genome. These new ORFs included N-terminal extensions for protein coding genes. For instance, we identified a CUG initiated extension of **(B)** a transcript of the *Aurka* kinase gene. The CHX coverage of the transcript is shown (full coverage and P-site offset [coloured by reading frame relative to the annotated TIS]) along with the HARR-ND coverage (P-site offset) illustrating the initiation signal at the CUG start codon upstream of the NCBI annotated AUG start codon.

2.3 The Chinese hamster genome harbours thousands of short open reading frames

The ORF-RATER algorithm also identified thousands of previously uncharacterised short open reading frames (sORFs) in the Chinese hamster genome (Figure 3A; Table S3). sORFs are defined as ORFs predicted to produce proteins < 100 aa termed microproteins (Olexiouk et al., 2018). Greater than 90% of the ORFs identified in the 5' region of mRNAs or in transcripts annotated as non-coding were sORFs (Figure 3B). In this study we found 3,497 uORFs (Figure 3D) with an average length of 24 aa (Figure S3A), with AUG (46.4%) the most prevalent start codon, followed by CUG (31%), GUG (14%) and UUG (8.4%). The average length of the ouORFs (Figure 3E) identified (n = 1,203) was 48 aa (Figure S3B), with CUG (40.8%) the most frequent start codon, followed by AUG (27.7%), GUG (20.9%) and UUG (10.3%). The presence of uORFs in 5' leader sequences has been shown to have a repressive effect on the main ORF in multiple species (Chew et al., 2016), and we observed the same tendency in this study (Supplementary Results).

For the "New ORF" class (n=853), the majority of ORFs were found in transcripts annotated as non-coding (Figure 3F). The average length of "New ORFs" was 42 aa (Figure S3C), with AUG (60.4%) the most common start codon, followed by CUG (21.9%), GUG (15.2%) and UUG (5.5%). Upstream ORFs and sORFs in the "New" ORF group, were found to have clear differences in amino acid usage, when compared to annotated proteins with ≥ 100 aa. These results were comparable to a similar analysis conducted for microproteins encoded in the human genome (Martinez et al., 2020). CHO cell sORFs were found to have increased usage of arginine, glycine, and tryptophan as well as a decrease in usage of asparagine, glutamate, lysine, and aspartic acid. Alanine and proline were more prevalent in uORFs in comparison to annotated proteins and sORFs found in ncRNA, while methionine usage was more frequent in the sORFs in ncRNA (Figure 3C; Figure S4).

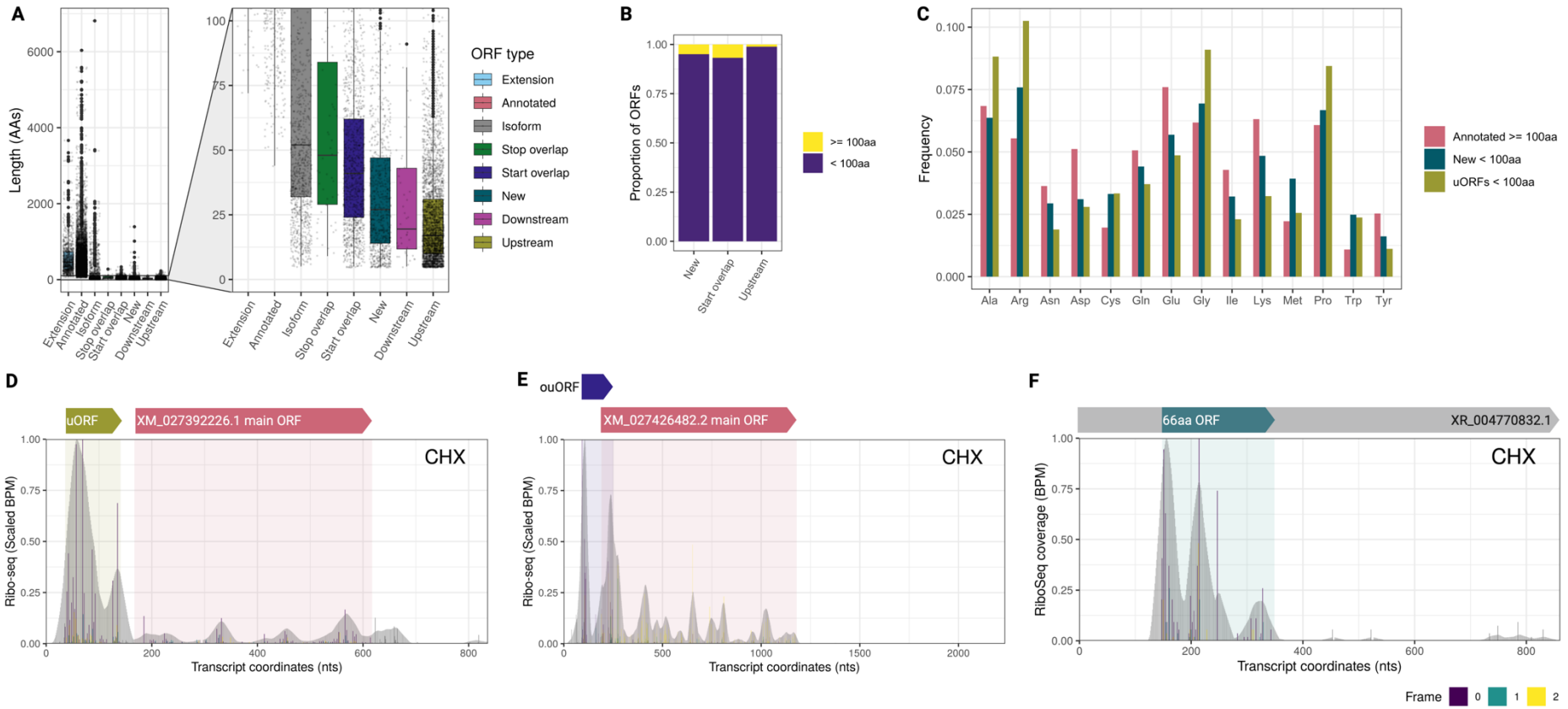


Figure 3: Ribosome footprint profiling uncovers thousands of short open reading frames in the Chinese hamster genome. A considerable number of previously uncharacterised ORFs identified by ORF-RATER were **(A)** predicted to be < 100 aa. In this study, we focused on short open reading frames found in the 5' leader of protein coding transcripts (i.e., upstream ORFs and start overlapping uORFs) as well as ORFs found in non-coding RNAs where **(B)** > 90% of all identified ORFs in each class were > 100 aa. Comparison of **(C)** the amino acid frequencies of uORFs (both uORFs and ouORFs) and ncRNA sORFs to annotated proteins, revealed differences in usage of amino acids including arginine and glycine when compared to conventional protein coding ORFs (≥ 100aa). Examples are shown of **(D)** an uORF found in a Ddit3 transcript, **(E)** an ouORF in Rad51 transcript and **(F)** an sORF found in a long non-coding RNA.

2.4 Detection of host cell microprotein contamination in adalimumab and trastuzamab drug products

Next, we aimed to determine if sORF derived microproteins are present in mAb drug product and are therefore a potential source of host cell protein (HCP) impurities. For this purpose, we utilised liquid chromatography-tandem mass spectrometry (LC-MS/MS) to analyse the HCP content of adalimumab. We searched the data against a protein sequence database comprised of previously annotated Chinese hamster proteins from UniProt (n=56,478) and the sORFs identified from the Ribo-seq data in this study (n=5, 645) (Figure 4B). For proteins ≥ 100aa, we retained only identifications comprised of ≥ 2 peptides, while for proteins with < 100 aa we retained identifications comprised of 1 unique peptide. The analysis of adalimumab resulted in the identification of 32 HCPs (Figure 4C; Table S4). 24 of the identified proteins were ≥ 100 aa, 23 of which were annotated in UniProt. A novel 1392 aa ORF in a transcript (accession: XM_027419483.2) of the *Notch3* gene with both a previously unannotated start and stop codons was also identified. We detected 8 microproteins in adalimumab drug product (Figure 4D) originating from sORFs found in the 5' leader sequence of protein coding transcripts (i.e., uORFs and ouORFs) and a non-coding RNA. The microproteins identified range from 26-89 aa in length with 6 of 8 microproteins found to initiate at near cognate start codons (i.e., CTG and TTG). Two or more peptides were detected for 3 microproteins, with a single peptide identified for the remaining microproteins. For the predicted 59 aa sORF found on XR_003481490.2 lncRNA transcript (Figure 4E), we were able to identify 4 peptides representing 85% coverage of the microprotein (Figure 4F, Figures S5-S8).

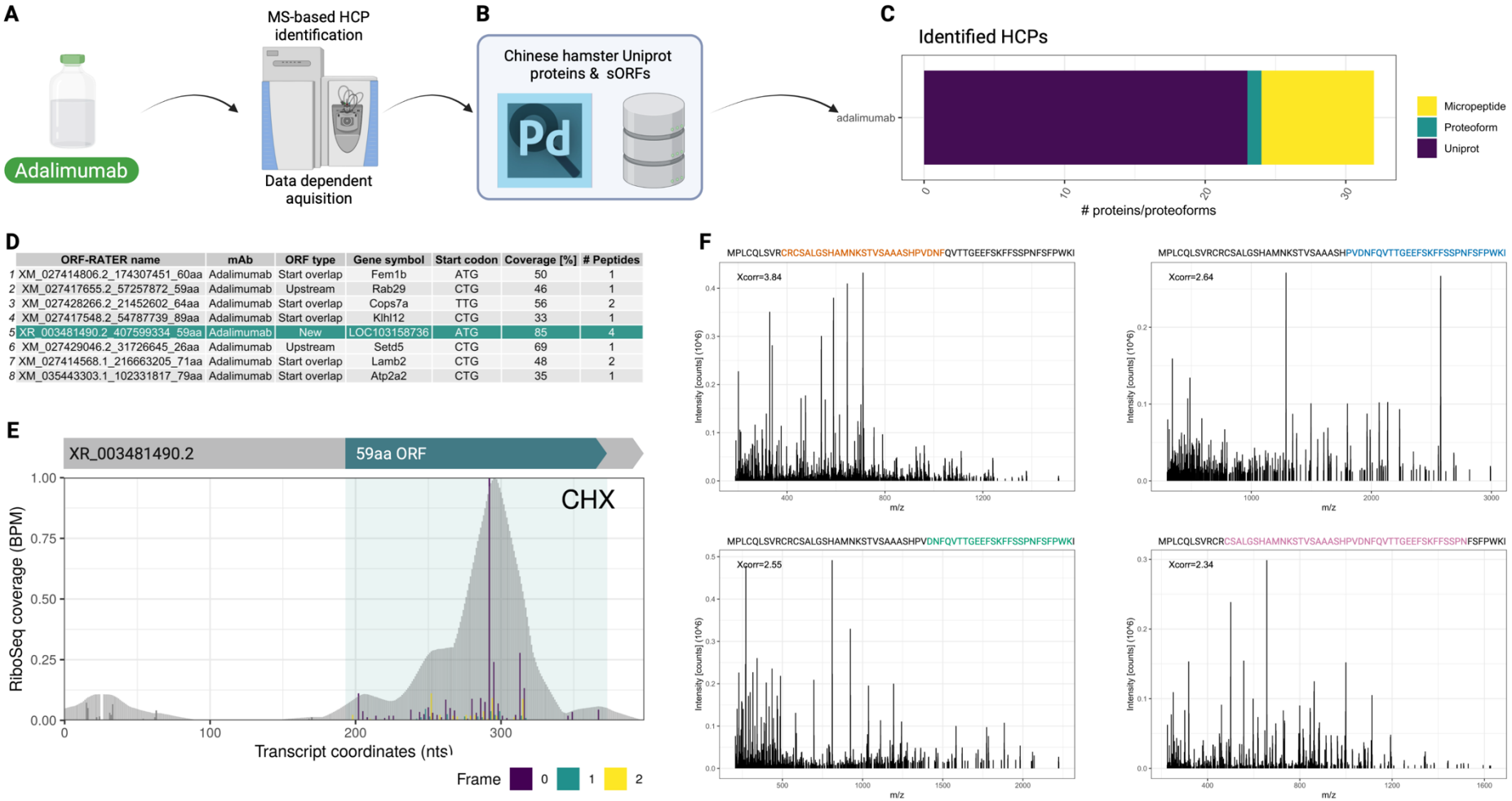


Figure 4: Microproteins are a new class of host cell impurity in mAb drug products. We **(A)** conducted LC-MS/MS based HCP analysis of adalimumab drug product. The resulting data were searched against a protein database comprised of **(B)** Chinese hamster proteins in UniProt and proteoforms from uORFs, ouORFs and non-coding RNA encoded sORFs identified using Ribo-seq. In total, we **(C)** identified 32 HCPs - 23 canonical proteins annotated in UniProt, 1 novel ORF > 100 aa in a transcript of the *Notch3* gene and **(D)** 8 microproteins. For **(E)** a lncRNA encoded sORF, we identified **(F)** 4 peptides, representing 85% coverage of the predicted microprotein sequence. The microprotein sequence is shown, and the peptide sequence corresponding to each spectrum shown.

2.5 The translation efficiency of sORFs found in non-coding RNA genes is altered in response to mild hypothermia in CHO cells.

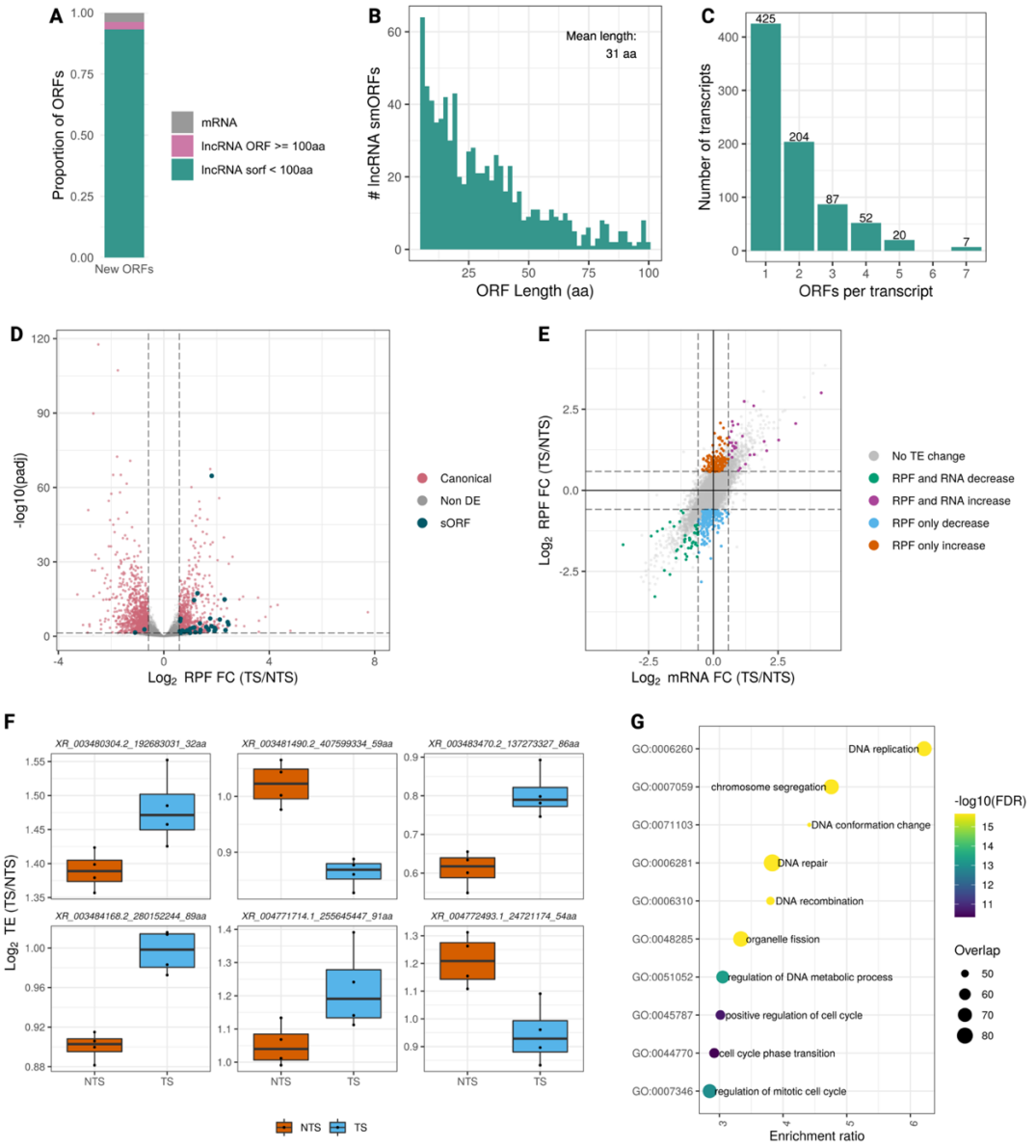
The reduction of cell culture temperature is a method used to extend the viability of some commercial cell culture processes and improve product quality (Masterton and Smales, 2014). Several studies have reported that mild hypothermia can change abundance of CHO cell HCPs (Goey et al., 2018, 2017; Jin et al., 2010; Tait et al., 2013). Ribo-seq enables the protein synthesis rate to be inferred by calculating the translation efficiency of each ORF. Translation efficiency is calculated by normalising the RPF occupancy by RNA abundance (Ingolia et al., 2009). Significant differences in translational regulation can then be determined for each ORF following the comparison of translation efficiency between conditions (Ingolia et al., 2009). Here, we wished to determine if sub-physiological temperature altered the translation efficiency of a selected cohort of sORFs and assess if translatome analysis can provide additional valuable information to complement transcriptome and proteome analysis.

For the differential gene expression and differential translation analysis, we utilised the CHX-treated Ribo-seq data for the TS and NTS sample groups, along with the matched RNA-seq data (Figure S9). We focused only on the “New” ORF sORFs identified in non-coding RNA transcripts. 821 ORFs identified by ORF-RATER were found on transcripts annotated as non-coding, and 795 of these ORFs were predicted to produce a protein < 100 aa (Figure 5A). The average length of these putative microproteins found in non-coding RNA transcripts was 31 aa (Figure 5B). Most transcripts encoded 1 or 2 sORFs, although there were instances of up to 7 ORFs present in a single non-coding RNA transcript identified (Figure 5C). To ensure compatibility with the Plastid cs algorithm (Dunn and Weissman, 2016), we retained only the longest sORF per non-coding transcript (collapsed to 462 genes for counting) prior to merging with the annotated canonical ORFs for differential expression and differential translation analysis. During the read counting process, we excluded the first 5 and last 5 codons for ORFs ≥ 100aa and the first and last codons for ORFs < 100 aa to reduce potential bias from the accumulation of ribosomes at the beginning and end of the CDS. Prior to differential expression, genes with < 20 counts on average across the 8 samples were eliminated.

We initially conducted separate analyses of the RNA-seq and Ribo-seq data using DESeq2, to identify differences in RNA abundance (Figure S10A) and RPF occupancy (Figure S10B), as well as to determine the extent to which significant changes in abundance between both data types were correlated (Table S5). Following comparison of the TS and NTS samples, 1,880 ORFs were found to have significantly different RPF counts (1,846 canonical & 34 sORFs). 53.8% of the ORFs with significantly altered RPF density, also had significant change in RNA abundance in the same direction (Figure 5D & Figure S10C). The expression and RPF occupancy of 18 sORFs were found to be correlated (Figure S11). To determine if there were alterations in translational efficiency (ΔTE), we retained only those genes which had an average read count of 20 in the RNA-seq and Ribo-seq datasets. DESeq2 was again utilised to assess the differences in RPF counts; however for this analysis the RNA expression was included as an interaction term in the model. This approach allowed us to identify changes in ribosome occupancy that were altered independently of transcription. 374 ORFs (368 canonical & 6 sORFs) were found to be differentially translated (≥ 1.5-fold increase or decrease in

331 ΔTE; adjusted p-value < 0.05) upon the reduction of cell culture temperature (Figure 5E & Figure 5F;
332 Table S5)).

333 To assess if there was an overrepresentation of biological processes for canonical ORFs, we conducted
334 a GO enrichment analysis. For this analysis, we first combined the ORFs that were differentially
335 expressed in the same direction from the Ribo-seq and RNA-seq data with translationally regulated
336 ORFs. Forty-nine GO terms including processes related to DNA repair, cell cycle and apoptosis were
337 found to be significantly enriched (Figure 5G, Table S6A). A separate enrichment analysis was also
338 carried out for translationally regulated canonical ORFs. In this case, a single biological process, DNA
339 repair, was found to be enriched (FDR = 6.07×10^{-5}) (Table S6B). In total, 9 of the 26 genes overlapping
340 with the DNA repair GO process (Figure S12A), including *Brca1* ($\log_2$ ΔTE = -1.25 [padj = 6.3×10^{-14}])
341 (Figure S12B), were found to undergo a significant reduction in translation efficiency. These results
342 therefore demonstrate, that Ribo-seq can further enhance our understanding of CHO biology through
343 the identification of changes in translation regulation.



344
345 **Figure 5 (caption next page)**

Figure 5: Temperature shift induces alterations in translation regulation of canonical ORFs and sORFs in CHO cells. To characterise the impact of reducing cell culture temperature, we carried out differential expression and translation analysis of canonical ORFs and non-coding RNA sORFs. Examination of the 821 ORFs classified as "New" by ORF-RATER, revealed that **(A)** the majority of ORFs found in non-coding RNAs are sORFs. The average length of these sORFs is **(B)** 31aa with **(C)** as many as 7 sORFs encoded on a transcript. We identified **(D)** 1,011 ORFs (993 canonical & 18 sORFs) that had correlated differential expression in both RNA-seq and Ribo-seq data. We also identified **(E)** 374 (368 canonical & 6 sORFs) genes where the translational efficiency was altered. **(F)** The translational efficiency of 4 sORFs was found to increase, while 2 sORFs decreased. To understand the biological processes affected by temperature shift, we conducted **(G)** an enrichment analysis for GO biological processes against the canonical ORFs that were significantly altered (both transcriptionally and translationally).

2.6 Microproteins are differentially expressed between the exponential and stationary phases of CHO cell culture

Next, to determine if microproteins were altered over the course of cell culture we performed LC-MS/MS-based proteomics employing label-free quantification (LFQ). For this analysis, a non-mAb producing CHO-K1GS cell line was cultured for 7 days, with samples acquired at two timepoints: (1) when the cells were undergoing exponential growth (Day 4) and (2) when the cells had reached stationary phase (Day 7) (Figure 6A). The cellular lysate from 4 biological replicates of each timepoint was subjected to a SP3 protein clean-up procedure and tryptic digestion, before LC-MS/MS analysis was carried out (Figure 6B). The resulting MS data from the 8 samples were searched against the protein sequence database comprised of annotated Chinese hamster proteins from UniProt and sORFs identified in this study (Figure 6B). Proteins ≥ 100 aa with 2 unique peptides, and proteins < 100 aa with 1 unique peptide identified, were confidently detected, and retained for further analysis.

In total, we identified 5,167 proteins that fulfilled these criteria across the Day 4 and Day 7 samples (Table S7A). We were able to identify 95 of the novel proteoforms identified from the Ribo-seq analysis (Figure 6C), found to initiate at AUG (n = 30), CUG (n = 31), GUG (n = 26) and UUG (n = 8). Of these novel proteoforms, 79 were microproteins derived from uORFs and ouORFs as well as "New" microproteins from an annotated protein coding gene (e.g., *Wnk4*) and transcripts annotated as non-coding (n = 11) (Figure 6D; Table S7B). To determine if microproteins were differentially expressed between the Day 4 and Day 7 samples we first $\log_2$ transformed and median normalised the abundance of each protein. The proDA (Ahlmann-Eltze and Anders, 2020) algorithm was utilised to fit a probabilistic drop out model to these data. Following the observation of global protein differences between the two conditions (Figure 6E), we then performed differential expression analysis. We identified 1,824 differentially expressed proteins (≥ 1.5 absolute fold change, adjusted p-value < 0.05, protein detected in at least 4 samples). 21 novel proteoforms, 13 of which were microproteins, were found to significantly change in abundance (6 upregulated and 7 downregulated) between exponential growth and stationary phase of culture (Table S7C).

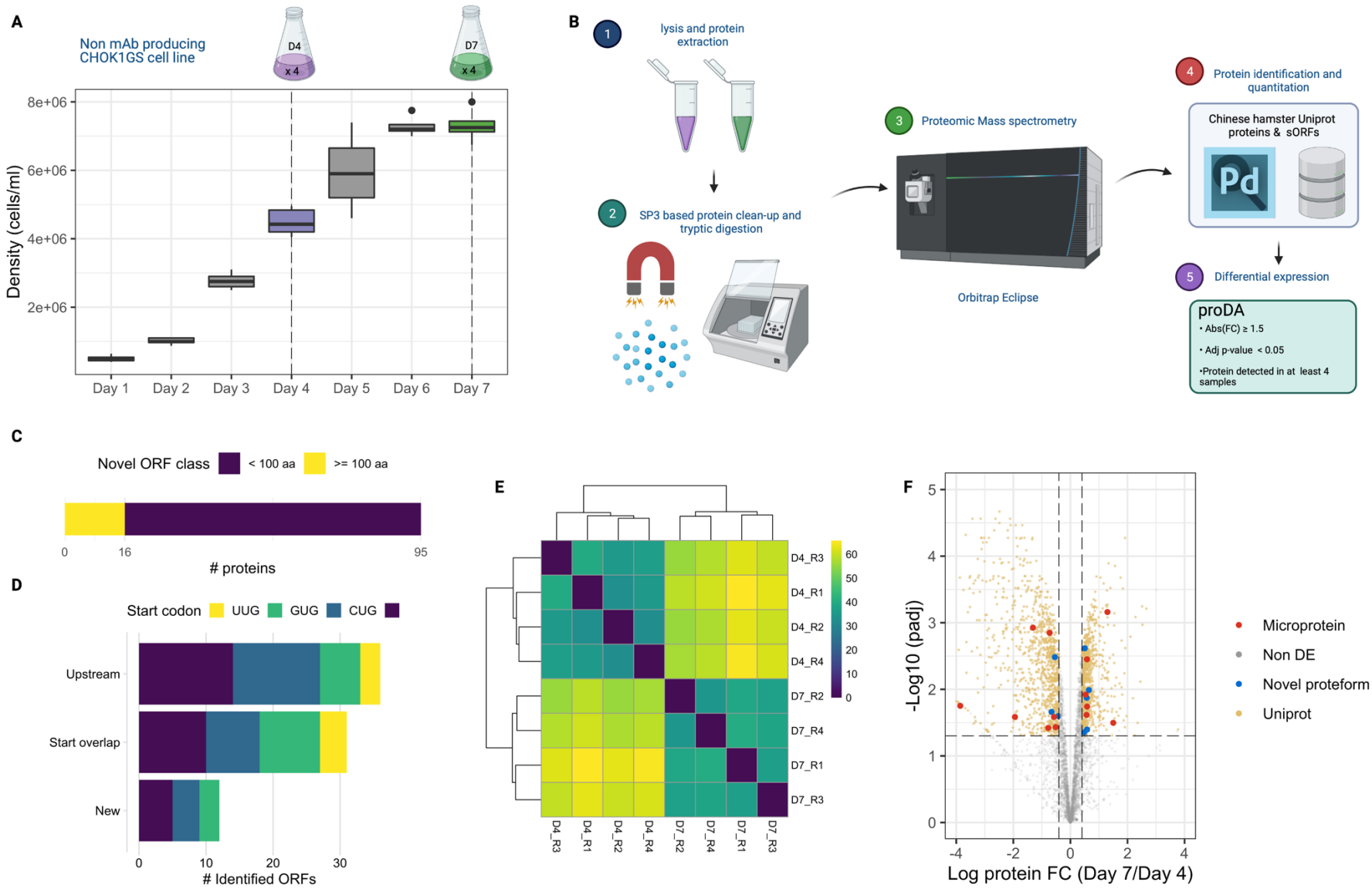


Figure 6 (caption next page)

Figure 6: Microproteins are differentially expressed between the exponential growth and stationary phases of CHO cell culture. To determine if proteoforms predicted from the Ribo-seq could be identified at the protein level, we conducted LC-MS based proteome analysis. For this experiment, we utilised a non mAb-producing CHOK1GS cell line, and **(A)** captured samples at the exponential growth (Day 4) and stationary phases (Day 7) for proteomics. Proteins were extracted from cell lysates (4 biological replicates for each condition) and **(B)** a SP3-based protein cleanup method followed by tryptic digestion was used to prepare samples for MS analysis. The resulting data was searched against a combined database of Chinese hamster proteins from Uniprot and ORF-RATER identifications using Proteome Discover 2.5, and label-free quantification performed. Only those proteins ≥ 100 aa with 2 unique peptide hits and those proteins < 100 aa with 1 unique peptide, were retained for further analysis. This analysis resulted in the identification of 5,167 protein groups. For the novel proteoforms identified in this study, **(C)** 95 (16 ≥ 100 aa & 79 < 100 aa) were identified by mass spectrometry. We found sORF derived proteins that initiated at **(D)** the four start codons considered in study. uORF derived microproteins were the most prevalent (n=36), followed closely by ouORF microproteins (n=31). We also detected 11 microproteins from transcripts annotated as non-coding RNA and one "New" microprotein derived from a transcript of the *Wnk4* gene. Next, we median normalised the $\log_2$ abundance for each protein, and assessed **(E)** the global difference in the proteome between Day 4 and Day 7 of cell culture. The probablistic dropout model of proDA was employed to assess the differential expression of identified proteins, with those proteins with an absolute fold change ≥ 1.5, an adjusted p-value < 0.05 and found in at least 4 samples considered to be differentially expressed. A total of **(F)** 1,824 proteins were found to be altered. Of the novel ORFs identified in this study, we found 13 microproteins that were differentially expressed between the exponential and stationary phases of CHO cell culture.

3. Discussion

Here, we present the findings of a ribosome footprint profiling experiment where both translation initiation and elongation were captured at single nucleotide resolution in CHO cells for the first time. The utilisation of harringtonine to arrest translation resulted in an enrichment of RPFs at the TIS and enabled transcriptome wide identification of ORFs including those that started at near cognate codons. We found that the use of alternative initiation sites is widespread across the CHO cell transcriptome with ~29% of all new ORFs identified beginning at non-AUG start codons (in agreement with TISs for human present in the TISdb (Lee et al., 2012)). For previously annotated protein coding transcripts, we were able to identify 685 extended proteoforms that begin at near cognate start codons. While AUG initiated translation is thought to result in the highest rate of protein synthesis (Kearse and Wilusz, 2017), it is possible, as with other species (Liang et al., 2014), that non-AUG initiated N-terminal extensions play a role in CHO cell biology. We also found thousands of novel sORFs predicted to encode microproteins in the 5' leader sequence of Chinese hamster mRNAs and ncRNA transcripts.

While the work conducted in this study has allowed us to significantly expand the annotation, it is likely that there remain further undiscovered ORFs in the Chinese hamster genome. In addition, our work is potentially limited by the combined use of harringtonine and flash freezing, which likely lead to residual elongating ribosomes and subsequent identification of potential false positive translation initiation sites. We eliminated those classes of ORFs that are liable to be affected (i.e., truncations) entirely from further analysis and conservatively assessed the remaining classes to limit false positive identifications (at the expense of potentially increasing the false negative rate). Future studies utilising Chinese hamster tissues as well as different CHO cell lines grown under a variety of conditions producing a range of mAb and other protein formats will enable the identification of additional proteoforms. In addition, performing Ribo-seq experiments with different inhibitors such as lactimidomycin and puromycin in the future could not only enable new ORFs to be identified but also allow quantitative comparison of CHO cell translation initiation in different conditions (Lee et al., 2012; Zhang et al., 2017).

The identification of CHO cell microproteins in this study permitted the use of a more comprehensive proteomic database for mass spectrometry, resulting in an enhanced assessment of HCP impurities in

adalimumab mAb drug product. We identified 1 novel protein > 100 aa and 8 novel microproteins from the LC-MS/MS data. Previous reports have shown that the population of HCPs is affected by the cell culture process. Our findings also show that microprotein abundance is altered over the course of cell culture and by a change in the cell culture environment. Process optimisation could, therefore, be utilised in the future to control microprotein impurities in the final product.

It should be noted that we make no claims regarding any risk to the patient or impact on efficacy of the mAbs arising from host cell microprotein impurities. In fact, the safety and effectiveness of 100 mAbs approved to date (Mullard, 2021), the majority of which are manufactured in CHO cells, provides convincing evidence that microproteins do not cause widespread issues, if present, in approved drug products. Nevertheless, CHO cell microproteins are a new class of host cell impurity and future studies to evaluate if, in certain circumstances, these HCPs could elicit an immune response, affect mAb stability or how they escape the purification process would be valuable for the industry. To facilitate these efforts, we have made the protein sequence database used for MS analysis freely available (download).

There has been considerable interest in gaining a deeper understanding of the CHO cell biological system and utilising the knowledge acquired to guide process development and cell line engineering strategies to improve manufacturing performance (Kuo et al., 2018). The results of this study could also prove useful for future studies in this area. For example, the most prevalent ORF type identified in this study was found in regulatory regions found in the 5' leader of mRNAs. We have shown that these upstream ORFs tend to have a repressive effect on translation of the main ORF and in some cases can also affect the abundance of the transcript. Manipulating uORFs or ouORFs has the potential to provide new routes for host cell line engineering to control the synthesis of CHO cell proteins. The knowledge gained from the study of endogenous CHO cell uORFs could also be used, similar to previous studies with synthetic uORFs (Ferreira et al., 2013), to precisely control the production of a recombinant protein.

The identification of differentially translated genes, upon a reduction of cell culture temperature, demonstrates that a component of the divergence observed in RNA abundance and RPF density is explained, in part, by differences in translation regulation between the two conditions. For example, we showed the reduced translation efficiency of proteins involved in DNA repair, a number of these genes, including *Brca1* were altered solely at the level of translation. Analysing translation regulation using Ribo-seq therefore provides a more complete understanding of the biological system than is possible with transcriptional profiling alone. In addition, differences in expression and translation of a number of sORFs encoded in non-coding transcripts (a phenomenon also observed in other species (Ji et al., 2015)) were responsive to sub-physiological cell culture temperature. Mass spectrometric analysis of a different CHO cell line than that used for the Ribo-seq confirmed the existence of a total of 95 novel proteoforms of which 79 were microproteins. The expanded protein database resulting from our work enhances the application of proteomics for CHO cell biology studies.

4. Conclusion

We have refined the annotation of the Chinese hamster genome by identifying proteoforms of annotated proteins initiating at non-AUG start codons, as well as characterising upstream ORFs and sORFs in RNAs previously annotated as non-coding predicted to encode for microproteins. The resulting protein sequence database enhances MS-based HCP analysis, and we have shown for the first time that microproteins can be found in mAb drug product and therefore represent a new class of host cell impurity. We also show that Ribo-seq is also a powerful approach for monitoring CHO cell translational regulation, providing an additional layer of biological understanding that is not possible with transcriptomics alone. The identification of new proteoforms and extending the annotation also enhances the utility of mass spectrometry-based proteomics to study CHO cell biology.

5. Materials and Methods

5.1 Cell culture

5.1.1 CHO K1 mAb

To generate the samples for Ribo-seq and RNA-seq, a mAb producing CHOK1 cell line (CHO K1 mAb) was seeded at a density of 2×10^5 cells/ml in 50ml SFM-II media (Gibco, 12052098) in 8 replicate shake flasks in a Kuhner orbital shaker at 170rpm at 5% CO_2. The cultures were grown at 37°C for 48hr post-seeding, at which point the temperature of 4 of the shake flasks was reduced to 31°C, while the remaining 4 shake flasks per experimental condition were maintained at 37°C (Figure 1A). Samples for library preparation were acquired 72hrs post-seeding. The procedure was repeated in two separate experiments, the first was used to generate Ribo-seq and matched total RNA-seq libraries from cycloheximide-treated cells (8 samples) and the second to generate Ribo-seq libraries from harringtonine-treated (8 samples) and matched no drug-treated cells.

5.1.2 CHOK1GS

For proteomics analysis (Section 5.4) of stationary and exponential phases of cell culture the CHOK1-GS cell line was seeded at a density of 2×10^5 cells/ml in 30ml CD FortiCHO™ medium (Gibco, cat.no. A1148301) supplemented with 4mM L-glutamine (L-Glutamine, cat.no. 25030024) in 250ml Erlenmeyer shake flasks in 8 replicates. The cultures were maintained at 37°C, 170 rpm, 5% CO_2 and 80% humidity in a shaking incubator (Kuhner) for 4 or 7 days. On day 4 and 7 cells were counted, pelleted, and resuspended in fresh media supplemented with cycloheximide to a final concentration of 100µg/ml (Sigma, cat.no. C4859-1ml). Following a 5-minute incubation at 37°C, cells were centrifuged at 300g for 5 minutes at room temperature and the media was removed. The cell pellets were washed with ice cold PBS with cycloheximide (100µg/ml) and stored at -80°C until analysis.

5.2 Ribosome footprint profiling

5.2.1 Translation Initiation sample preparation

72 hours post seeding 2 × 10^5 cells/ml (per replicate) were treated with harringtonine (2 µg/ml) (or DMSO) for 2 minutes at 31°C or 37°C. The cultures were transferred to 50 ml tubes and following centrifugation at 1,000 rcf for 5 minutes at room temperature the media was removed, and the cells were resuspended in ice cold PBS supplemented with harringtonine or DMSO respectively). Following a 5-minute centrifugation at 1,000 rcf at 4°C, the PBS was removed, and the pellet was flash frozen in liquid nitrogen. Frozen pellets were resuspended in 400µl 1X Mammalian Polysome buffer (Illumina TruSeq Ribo Profile (mammalian) kit) prepared according to manufacturer's guidelines. Cell lysates were incubated on ice for 10 minutes, centrifuged at 18,000 rcf for 10 minutes at 4°C to pellet cell debris and the supernatant was used for ribosome-protected fragment (RPF) isolation and library preparation.

5.2.2 Translation elongation sample preparation

72 hours post seeding a total of 25 × 10^6 cells (per replicate) were pelleted and resuspended in 20ml of fresh CHO-S-SFMII media supplemented with cycloheximide at a final concentration of 0.1 mg/mL and incubated at 37°C or 31°C for 10 min. Cells were subsequently pelleted, washed in 1 mL of ice-cold PBS containing 0.1 mg/mL of CHX, clarified and lysed. Prior to the generation of ribosomal footprints, part of the lysate was used for total RNA extraction and RNA-seq library preparation with the TruSeq Ribo Profile (mammalian) kit. PAGE Purified RPFs were used for ribosome profiling library preparation with the Illumina TruSeq Ribo Profile (mammalian) kit.

5.2.3 Library preparation

To prepare RNA-seq and Ribo-seq libraries for sequencing, the TruSeq Ribo Profile (Mammalian) Kit (Illumina) was used in accordance with the manufacturer's specifications. For Ribo-seq samples RNase treatment was performed with 10µl of TruSeq Ribo Profile Nuclease per 200µl lysate at room temperature for 45 minutes with gentle shaking. Digestion was stopped with 15µl SUPERaseIn (20U/µl) (Ambion, cat. No. AM2696). Monosomes were isolated with size exclusion chromatography using the Illustra MicroSpin S-400 HR Columns (GE life sciences, cat. no. 27514001) according to manufacturer's instructions. Ribosomal RNA was removed with the RiboZero-Gold rRNA removal Kit (Illumina, cat. No MRZG12324). Ribosome protected fragments were size selected from a 15% denaturing urea polyacrylamide gel (PAGE) following electrophoresis (7M urea, acrylamide (19): bis-acrylamide (1)). A gel extraction step (from 15% denaturing PAGE gels) for the isolation of linker ligated ribosome protected fragments, was added to the protocol after the linker ligation reaction as in Ingolia's protocol (Ingolia et al., 2012) for the Harringtonine and No-drug treated samples, to avoid high concentration of linker dimers contaminating the final library. Following reverse transcription, cDNA was extracted from 7.5% denaturing urea PAGE gels. PCR amplified libraries were purified from 8% PAGE gels and subsequently analysed with the Agilent High Sensitivity DNA assay (Agilent, Bioanalyzer).

5.2.4 Sequencing

The libraries for translation initiation and elongation analysis were sequenced on an Illumina NextSeq configured to yield 75bp and 50bp single end reads respectively.

5.3 Ribo-seq and RNA-seq data analysis

5.3.1 Pre-processing

Adapter sequences were trimmed from the Ribo-seq and RNA-seq datasets using Cutadapt v1.18 (Martin, 2011), and Trimmomatic v0.36 (Bolger et al., 2014) was used to remove low quality bases. To remove contaminants from the Ribo-seq data Chinese hamster rRNA, tRNA and snoRNA sequences were downloaded from v18 of the RNAcentral v18 database (The RNAcentral Consortium, 2019) and an individual STAR v2.7.8a (Dobin et al., 2013) index was built for each type of RNA. The Ribo-seq reads were aligned against each index using the following parameters: --seedSearchStartLmaxOverLread .5 --outFilterMultimapNmax 1000 --outFilterMismatchNmax 2. Reads that mapped to rRNA, tRNA or snoRNA were discarded.

5.3.2 Read alignment

The pre-processed Ribo-seq and RNA-seq data aligned to the NCBI CriGri-PICRH 1.0 genome and transcriptome (GCA_003668045.2) (Hilliard et al., 2020) with STAR v2.7.8a using the following parameters: --outFilterMismatchNmax 2 --outFilterMultimapNmax 1 --outFilterMatchNmin 16 --aligEndsType EndToEnd.

5.3.3 Ribo-seq P-site offset identification and selection of RPFs

The P-site offset (the number of nucleotides between the 5' end of a Ribo-seq read and the P-site of the ribosome footprint that was captured) was determined using Plastid v0.4.8 (Dunn and Weissman, 2016) by first defining the genomic region around annotated Chinese hamster CDS using the metagene generate programme with default settings. The P-site tool was then used to assess the P-site for different read lengths around the expected mammalian RPF size (27-32nt) for those CDS that had at least 10 mapped reads to the start region. Using Plastid the estimated P-site offsets for each read length was determined for the CHX, HARR and No-drug Ribo-seq data. Only those read lengths where ≥ 60% of the reads were found to have the expected triplet periodicity with a P-site offset of 12 were retained for further analysis.

5.3.4 ORF identification

The 8 replicates from each Ribo-seq type were merged to increase sensitivity before the ORF-RATER pipeline (Fields et al., 2015) was used to identify ORFs in the Chinese hamster genome. Annotated pseudogenes were removed from the reference with only those transcripts with a minimum of 64 mapped RPFs from the CHX and ND Ribo-seq data were considered for ORF identification. The ORF search was limited to NUG codons with only the Harr Ribo-seq data used to identify the translation initiation sites, while the CHX and ND RPFs were used to assess translation at putative ORFs. Identified ORFs with an ORF-RATER score ≥ 0.5 (Eisenberg et al., 2020; Finkel et al., 2021) and with a length ≥ 5aa were retained. Visualisation of transcripts with novel Chinese hamster ORFs was accomplished using deeptools bamCoverage (Ramírez et al., 2016). Where one or more Ribo-seq type was displayed on the same figure, the bins per million (BPM) value was scaled between 0 and 1.

5.3.5 Transcript-level quantitation

The RNA abundance and RPF density in reads per kilobase mapped (RPKM) of annotated and novel ORFs was determined for each CHX-treated Ribo-seq replicate from the NTS and TS samples using the Plastid cs programme (Dunn and Weissman, 2016). Reads and RPFs aligning to the first 5 or last 15 codons of each CDS were eliminated for ORFs ≥ 100aa while for those ORFs < 100aa the first and last codon counts were excluded (Martinez et al., 2020). The translation efficiency for CDS regions was calculated by dividing the RPF RPKM value by that of the RPKM of the matched RNA-seq sample.

5.3.6 Gene-level differential expression and differential translation analysis

To conduct gene-level differential translation analysis the reference protein coding annotation was merged with selected ORFs found to be encoded by non-coding RNAs. Prior to counting Plastid cs generate was used to collapse transcripts that shared exons, remove regions comprised of more than 1 same-strand gene and create position groups corresponding to exons, CDS, 5' leader and 3'UTR. An identical codon masking procedure to transcripts was also utilised for gene level analyses. The counts corresponding to CDS regions were analysed by DESeq2 (Love et al., 2014) to identify differences between the RNA-seq and RPF counts for the TS and NTS groups. For differential translation analysis, the RNA-seq data was used as an interaction term within the DESeq2 model to enable the identification of changes in RPF density independently of RNA abundance. For all analyses, an absolute fold change ≥ 1.5 and Benjamini Hochberg adjusted p-value < 0.05 were considered significant. Gene level coverage was determined using deepTools bamCoverage and the wiggleplotr R-package v1.16.0 (Alasoo, 2021) was used to display the track and corresponding gene model.

5.3.7 Enrichment Analysis

The overrepresentation of gene ontology (GO) biological processes in differentially expressed and/or differentially translated genes were identified with the R WebGestaltR package (Wang et al., 2017). Where no gene symbol was available the Chinese hamster gene name was mapped to the NCBI Mus musculus GRCm39 annotation, and the corresponding mouse gene symbol was used. GO biological processes with a Benjamini-Hochberg adjusted p-value of < 0.05 were considered significant.

5.4 Proteomic analysis of the CHOK1GS cell lysate

5.4.1 Sample preparation for reversed phase liquid chromatography-tandem mass spectrometry (RPLC-MS/MS)

Eight samples comprised of 4 biological replicates of CHOK1GS cells at day 4 and day 7 of culture were prepared for proteomics using a semi-automated version of the SP3 protocol (Hughes et al., 2019). Briefly CHO-GS cells were pelleted via centrifugation at 300 × g for 5 mins. Following a wash step with 1 × PBS, cells were lysed using 1 × RIPA buffer (Cell Signalling Technology, Dublin, Ireland) containing 1 × protease inhibitor (cOmplete™, Mini, EDTA-free Protease Inhibitor Cocktail, Sigma, Wicklow, Ireland) followed by sonication. After removing cell debris, protein concentration was determined and an aliquot of the sample containing 50 µg of protein was used for tryptic digestion over night as described before (Strasser et al., 2021). Following digestion, magnetic beads were removed, and samples were acidified by adding 0.1% (v/v) formic acid before subsequent analysis using LC-MS

as described below. Note: An identical sample preparation procedure was carried out for HCP analysis (Section 5.5.2) of adalimumab (provided by St. Vincent's University Hospital in Dublin, Ireland).

5.4.2 RPLC-MS/MS analysis of CHOK1GS cell lysates

Mass spectrometric analysis was performed using an Orbitrap Eclipse™ Tribid™ mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) coupled to an UltiMate™ 3000 RSLCnano system by means of an EASY-Spray™ source (Thermo Fisher Scientific, Germering, Germany). 2 µg per sample were loaded onto a C18 Nano-Trap Column followed by separation using an EASY-Spray Acclaim PepMap 100, 75 µm × 50 cm column (Thermo Fisher Scientific, Sunnyvale, CA, United States) maintained at 45.0°C at a flow rate of 250.0 nL/min. Separation was achieved using a gradient of (A) 0.10% (v/v) formic acid in water and (B) 0.10% (v/v) formic acid in acetonitrile (LC-MS optima, Fisher Scientific). Gradient conditions were as follows: 5% B for 5 min, followed by a linear gradient of 5-25% in 95 min, followed by another increase to 35% B in 20 min. The separation was followed by 2 wash steps at 90% B for 5 min and the column was re-equilibrated at 5% B for 15 min.

MS analysis was performed in positive ion mode. Full scans were acquired in the Orbitrap at a resolution setting of 120,000 (at m/z 200) with a scan range of m/z 200-2,000 using a normalized automatic gain control (AGC) target of 100% and an automatic maximum injection time in centroid mode. Using a 3 s cycle time, ions were selected for HCD fragmentation using a collision energy setting of 28%. Fragment scans were acquired in the Orbitrap using a resolution setting of 30,000 (at m/z 200). The AGC target was set to 200%. For isolation of precursor ions, an isolation window of 1.2 m/z was used. An intensity threshold of 5.0e4 was applied while unassigned charge states as well as charges >6 were excluded. The dynamic exclusion time was set to 60 s with ± 5 ppm tolerance.

5.4.3 Proteome discoverer data analysis

Analysis of acquired raw data was performed in Proteome Discoverer version 2.5 (Thermo Fischer Scientific). Each dataset was searched against a protein sequence database comprised of 56,478 Cricetulus griseus (taxon identifier 10029) proteins downloaded from UniProt in November 2021 (UniProt Consortium, 2021) and 5,645 novel proteoforms (derived from uORFs, ouORFs and ORFs encoded on NCBI annotated non-coding RNAs).

5.4.4 Protein detection and quantitation in cell lysate samples

For protein identification as well as label-free quantitation (LFQ) in cell lysate samples, two Sequest HT searches were performed using fixed value PSM validator (#1 and #2) and an extra Sequest HT (#3) search was conducted using Percolator. Detailed settings of the database search can be found in Table S8A.

5.4.5 Identification of differentially expressed proteins

LFQ data was log2 transformed, median normalised and the proDA algorithm (Ahlmann-Eltze and Anders, 2020) was then utilised to fit a probabilistic drop out model to these data prior to differential expression analysis. Proteins with a ≥ 1.5 absolute fold change, adjusted p-value < 0.05 and detected in at least 4 samples were considered differentially expressed.

5.5 Host cell protein analysis

5.5.1 RPLC-MS/MS analysis of adalimumab drug product

Following tryptic digestion as described above (Section 5.4.1), HCP analysis of adalimumab samples was performed using a Q Exactive™ Plus Hybrid quadrupole-Orbitrap™ mass spectrometer on-line hyphenated to an UltiMate™ 3000 RSLCnano system by means of an EASY-Spray™ source. Therefore, 5 µg of tryptic peptides were separated using the same column and solvents mentioned above. Gradient conditions were as follows: 5% to 25% B in 140 min, increased to 35% B in 30 min followed by two wash steps at 90% B. The used flowrate was 250 nL/min, and the column temperature was maintained at 40ºC.

Data-dependent acquisition (DDA)-MS2 analysis was performed in positive ion mode. Full scans were acquired at a resolution of 35,000 (at m/z 200) for a mass range of m/z 300-2,000 using an AGC target of 1.0×10^6 with a maximum injection time (IT) of 120 ms. Fragment scans were acquired using a resolution setting of 17,500 with an AGC target of 1.0×10^5 and a maximum IT of 200 ms, an isolation window of m/z 1.2 and a signal intensity threshold of 4.2×10^4. Fragmentation of top 15 most abundant precursor ions was done using a normalised collision energy set to 29 using a dynamic exclusion for 60 s and charge exclusion set to unassigned and > 8. Additionally, adalimumab derived peptides were set on an exclusion list allowing for a 10 ppm mass tolerance to aid low abundant HCP detection.

5.5.2 Detection and quantitation of HCPs in drug product samples

For protein detection and LFQ in drug product samples, two Sequest HT searches were performed using Fixed value PSM validator. Detailed settings of the database search can be found in Table S8B.

5.6 Data availability

The Ribo-seq and RNA-seq data from the harringtonine, cycloheximide and no-drug treated cells have been in the Sequence Read Archive (SRA) with accession code PRJNA778050. The code required to reproduce the results presented in this manuscript is available at https://github.com/clarke-lab/CHO_cell_microprotein_analysis. The proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD030186. The protein sequence database used for HCP and proteomics analysis is available at https://doi.org/10.5281/zenodo.5801357.

Conflict of interest

MCR, IT, PK, CT, FG, LS, BLK, MC, NB, JB and CC declare no competing interests. LZ is an employee of Pfizer Inc.

Author Contributions

IT and CC conceived the study and designed experiments; Cell culture and Ribo-seq were carried by IT and PK. Ribo-seq data analysis was performed by MCR and CC. CT, FG, LS and JB performed the

proteomics analysis. MCR, IT, LZ, MC, BLK, NB, JB and CC wrote the manuscript. All authors reviewed the paper.

Acknowledgements

The authors gratefully acknowledge funding from Science Foundation Ireland (grant references: 15/CDA/3259 & 13/SIRG/2084). Figures were created with BioRender.com.

References

Ahlmann-Eltze, C., Anders, S., 2020. proDA: Probabilistic dropout analysis for identifying differentially abundant proteins in label-free mass spectrometry. Biorxiv 661496.

Alasoo, K., 2021. wiggleplotr: Make read coverage plots from BigWig files.

Aspden, J.L., Eyre-Walker, Y.C., Phillips, R.J., Amin, U., Mumtaz, M.A.S., Brocard, M., Couso, J.-P., 2014. Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. eLife 3, e03528. https://doi.org/10.7554/eLife.03528

Bazzini, A.A., Johnstone, T.G., Christiano, R., Mackowiak, S.D., Obermayer, B., Fleming, E.S., Vejnar, C.E., Lee, M.T., Rajewsky, N., Walther, T.C., Giraldez, A.J., 2014. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. EMBO J. 33, 981–993. https://doi.org/10.1002/embj.201488411

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinforma. Oxf. Engl. 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bracewell, D.G., Francis, R., Smales, C.M., 2015. The future of host cell protein (HCP) identification during process development and manufacturing linked to a risk-based management for their control. Biotechnol. Bioeng. 112, 1727–1737. https://doi.org/10.1002/bit.25628

Chen, J., Brunner, A.-D., Cogan, J.Z., Nuñez, J.K., Fields, A.P., Adamson, B., Itzhak, D.N., Li, J.Y., Mann, M., Leonetti, M.D., Weissman, J.S., 2020. Pervasive functional translation of noncanonical human open reading frames. Science 367, 1140–1146. https://doi.org/10.1126/science.aay0262

Chew, G.-L., Pauli, A., Schier, A.F., 2016. Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish. Nat. Commun. 7, 11663. https://doi.org/10.1038/ncomms11663

Chiu, J., Valente, K.N., Levy, N.E., Min, L., Lenhoff, A.M., Lee, K.H., 2017. Knockout of a difficult-to-remove CHO host cell protein, lipoprotein lipase, for improved polysorbate stability in monoclonal antibody formulations. Biotechnol. Bioeng. 114, 1006–1015. https://doi.org/10.1002/bit.26237

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dunn, J.G., Weissman, J.S., 2016. Plastid: nucleotide-resolution analysis of next-generation sequencing and genomics data. BMC Genomics 17, 958. https://doi.org/10.1186/s12864-016-3278-x

Eisenberg, A.R., Higdon, A.L., Hollerer, I., Fields, A.P., Jungreis, I., Diamond, P.D., Kellis, M., Jovanovic, M., Brar, G.A., 2020. Translation Initiation Site Profiling Reveals Widespread Synthesis of Non-AUG-Initiated Protein Isoforms in Yeast. Cell Syst. 11, 145-160.e5. https://doi.org/10.1016/j.cels.2020.06.011

Ferreira, J.P., Overton, K.W., Wang, C.L., 2013. Tuning gene expression with synthetic upstream open reading frames. Proc. Natl. Acad. Sci. U. S. A. 110, 11284–11289. https://doi.org/10.1073/pnas.1305590110

Fields, A.P., Rodriguez, E.H., Jovanovic, M., Stern-Ginossar, N., Haas, B.J., Mertins, P., Raychowdhury, R., Hacohen, N., Carr, S.A., Ingolia, N.T., Regev, A., Weissman, J.S., 2015.

A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. Mol. Cell 60, 816–827. https://doi.org/10.1016/j.molcel.2015.11.013

Finkel, Y., Mizrahi, O., Nachshon, A., Weingarten-Gabbay, S., Morgenstern, D., Yahalom-Ronen, Y., Tamir, H., Achdout, H., Stein, D., Israeli, O., Beth-Din, A., Melamed, S., Weiss, S., Israely, T., Paran, N., Schwartz, M., Stern-Ginossar, N., 2021. The coding capacity of SARS-CoV-2. Nature 589, 125–130. https://doi.org/10.1038/s41586-020-2739-1

Goey, C.H., Bell, D., Kontoravdi, C., 2018. Mild hypothermic culture conditions affect residual host cell protein composition post-Protein A chromatography. mAbs 10, 476–487. https://doi.org/10.1080/19420862.2018.1433977

Goey, C.H., Tsang, J.M.H., Bell, D., Kontoravdi, C., 2017. Cascading effect in bioprocessing-The impact of mild hypothermia on CHO cell behavior and host cell protein composition. Biotechnol. Bioeng. 114, 2771–2781. https://doi.org/10.1002/bit.26437

Hanania, N.A., Noonan, M., Corren, J., Korenblat, P., Zheng, Y., Fischer, S.K., Cheu, M., Putnam, W.S., Murray, E., Scheerens, H., Holweg, C.T.J., Maciuca, R., Gray, S., Doyle, R., McClintock, D., Olsson, J., Matthews, J.G., Yen, K., 2015. Lebrikizumab in moderate-to-severe asthma: pooled data from two randomised placebo-controlled studies. Thorax 70, 748–756. https://doi.org/10.1136/thoraxjnl-2014-206719

Henry, S.M., Sutlief, E., Salas-Solano, O., Valliere-Douglass, J., 2017. ELISA reagent coverage evaluation by affinity purification tandem mass spectrometry. mAbs 9, 1065–1075. https://doi.org/10.1080/19420862.2017.1349586

Hilliard, W., MacDonald, M.L., Lee, K.H., 2020. Chromosome-scale scaffolds for the Chinese hamster reference genome assembly to facilitate the study of the CHO epigenome. Biotechnol. Bioeng. 117, 2331–2339. https://doi.org/10.1002/bit.27432

Huang, Y., Molden, R., Hu, M., Qiu, H., Li, N., 2021. Toward unbiased identification and comparative quantification of host cell protein impurities by automated iterative LC–MS/MS (HCP-AIMS) for therapeutic protein development. J. Pharm. Biomed. Anal. 200, 114069. https://doi.org/10.1016/j.jpba.2021.114069

Hughes, C.S., Moggridge, S., Müller, T., Sorensen, P.H., Morin, G.B., Krijgsveld, J., 2019. Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. Nat. Protoc. 14, 68–85. https://doi.org/10.1038/s41596-018-0082-x

Ingolia, N.T., Brar, G.A., Rouskin, S., McGeachy, A.M., Weissman, J.S., 2012. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat. Protoc. 7, 1534–1550. https://doi.org/10.1038/nprot.2012.086

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., Weissman, J.S., 2009. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. Science 324, 218–223. https://doi.org/10.1126/science.1168978

Ingolia, N.T., Lareau, L.F., Weissman, J.S., 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell 147, 789–802. https://doi.org/10.1016/j.cell.2011.10.002

Ivanov, I.P., Firth, A.E., Michel, A.M., Atkins, J.F., Baranov, P.V., 2011. Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences. Nucleic Acids Res. 39, 4220–4234. https://doi.org/10.1093/nar/gkr007

Ji, Z., Song, R., Regev, A., Struhl, K., 2015. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. eLife 4, e08890. https://doi.org/10.7554/eLife.08890

Jin, M., Szapiel, N., Zhang, J., Hickey, J., Ghose, S., 2010. Profiling of host cell proteins by two-dimensional difference gel electrophoresis (2D-DIGE): Implications for downstream process development. Biotechnol. Bioeng. 105, 306–316. https://doi.org/10.1002/bit.22532

Kearse, M.G., Wilusz, J.E., 2017. Non-AUG translation: a new start for protein synthesis in eukaryotes. Genes Dev. 31, 1717–1731. https://doi.org/10.1101/gad.305250.117

Koh, M., Ahmad, I., Ko, Y., Zhang, Y., Martinez, T.F., Diedrich, J.K., Chu, Q., Moresco, J.J., Erb, M.A., Saghatelian, A., Schultz, P.G., Bollong, M.J., 2021. A short ORF-encoded transcriptional regulator. Proc. Natl. Acad. Sci. U. S. A. 118, e2021943118. https://doi.org/10.1073/pnas.2021943118

Kuo, C.-C., Chiang, A.W., Shamie, I., Samoudi, M., Gutierrez, J.M., Lewis, N.E., 2018. The emerging role of systems biology for engineering protein production in CHO cells. Curr. Opin. Biotechnol. 51, 64–69. https://doi.org/10.1016/j.copbio.2017.11.015

Lee, C., Zeng, J., Drew, B.G., Sallam, T., Martin-Montalvo, A., Wan, J., Kim, S.-J., Mehta, H., Hevener, A.L., de Cabo, R., Cohen, P., 2015. The mitochondrial-derived peptide MOTS-c promotes metabolic homeostasis and reduces obesity and insulin resistance. Cell Metab. 21, 443–454. https://doi.org/10.1016/j.cmet.2015.02.009

Lee, Sooncheol, Liu, B., Lee, Soohyun, Huang, S.-X., Shen, B., Qian, S.-B., 2012. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proc. Natl. Acad. Sci. U. S. A. 109, E2424-2432. https://doi.org/10.1073/pnas.1207846109

Li, S., Cha, S.W., Heffner, K., Hizal, D.B., Bowen, M.A., Chaerkady, R., Cole, R.N., Tejwani, V., Kaushik, P., Henry, M., Meleady, P., Sharfstein, S.T., Betenbaugh, M.J., Bafna, V., Lewis, N.E., 2019. Proteogenomic annotation of the Chinese hamster reveals extensive novel translation events and endogenous retroviral elements. J. Proteome Res. 18, 2433–2445. https://doi.org/10.1021/acs.jproteome.8b00935

Li, X., An, Y., Liao, J., Xiao, L., Swanson, M., Martinez-Fonts, K., Pavon, J.A., Sherer, E.C., Jawa, V., Wang, F., Gao, X., Letarte, S., Richardson, D.D., 2021. Identification and characterization of a residual host cell protein hexosaminidase B associated with N-glycan degradation during the stability study of a therapeutic recombinant monoclonal antibody product. Biotechnol. Prog. 37, e3128. https://doi.org/10.1002/btpr.3128

Liang, H., He, S., Yang, J., Jia, X., Wang, P., Chen, X., Zhang, Z., Zou, X., McNutt, M.A., Shen, W.H., Yin, Y., 2014. PTENα, a PTEN isoform translated through alternative initiation, regulates mitochondrial function and energy metabolism. Cell Metab. 19, 836–848. https://doi.org/10.1016/j.cmet.2014.03.023

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550. https://doi.org/10.1186/s13059-014-0550-8

Luo, H., Tie, L., Cao, M., Hunter, A.K., Pabst, T.M., Du, J., Field, R., Li, Y., Wang, W.K., 2019. Cathepsin L Causes Proteolytic Cleavage of Chinese-Hamster-Ovary Cell Expressed Proteins During Processing and Storage: Identification, Characterization, and Mitigation. Biotechnol. Prog. 35, e2732. https://doi.org/10.1002/btpr.2732

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10–12. https://doi.org/10.14806/ej.17.1.200

Martinez, T.F., Chu, Q., Donaldson, C., Tan, D., Shokhirev, M.N., Saghatelian, A., 2020. Accurate annotation of human protein-coding small open reading frames. Nat. Chem. Biol. 16, 458–468. https://doi.org/10.1038/s41589-019-0425-0

Masterton, R.J., Smales, C.M., 2014. The impact of process temperature on mammalian cell lines and the implications for the production of recombinant proteins in CHO cells. Pharm. Bioprocess. 2, 49–61. https://doi.org/10.4155/pbp.14.3

Meleady, P., Hoffrogge, R., Henry, M., Rupp, O., Bort, J.H., Clarke, C., Brinkrolf, K., Kelly, S., Müller, B., Doolan, P., Hackl, M., Beckmann, T.F., Noll, T., Grillari, J., Barron, N., Pühler, A., Clynes, M., Borth, N., 2012. Utilization and evaluation of CHO-specific sequence databases for mass spectrometry based proteomics. Biotechnol. Bioeng. 109, 1386–1394. https://doi.org/10.1002/bit.24476

Mullard, A., 2021. FDA approves 100th monoclonal antibody product. Nat. Rev. Drug Discov. 20, 491–495. https://doi.org/10.1038/d41573-021-00079-7

Olexiouk, V., Van Criekinge, W., Menschaert, G., 2018. An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. Nucleic Acids Res. 46, D497–D502. https://doi.org/10.1093/nar/gkx1130

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., Manke, T., 2016. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 44, W160-165. https://doi.org/10.1093/nar/gkw257

Rathore, A., Chu, Q., Tan, D., Martinez, T.F., Donaldson, C.J., Diedrich, J.K., Yates, J.R., Saghatelian, A., 2018. MIEF1 Microprotein Regulates Mitochondrial Translation. Biochemistry 57, 5564–5575. https://doi.org/10.1021/acs.biochem.8b00726

The RNAcentral Consortium, 2019. RNAcentral: a hub of information for non-coding RNA sequences. Nucleic Acids Res. 47, D221–D229. https://doi.org/10.1093/nar/gky1034

Slavoff, S.A., Heo, J., Budnik, B.A., Hanakahi, L.A., Saghatelian, A., 2014. A human short open reading frame (sORF)-encoded polypeptide that stimulates DNA end joining. J. Biol. Chem. 289, 10950–10957. https://doi.org/10.1074/jbc.C113.533968

Strasser, L., Boi, S., Guapo, F., Donohue, N., Barron, N., Rainbow-Fletcher, A., Bones, J., 2021. Proteomic Landscape of Adeno-Associated Virus (AAV)-Producing HEK293 Cells. Int. J. Mol. Sci. 22, 11499. https://doi.org/10.3390/ijms222111499

Tait, A.S., Tarrant, R.D.R., Velez-Suberbie, M.L., Spencer, D.I.R., Bracewell, D.G., 2013. Differential Response in Downstream Processing of CHO Cells Grown Under Mild Hypothermic Conditions. Biotechnol. Prog. 29, 688–696. https://doi.org/10.1002/btpr.1726

Tzani, I., Monger, C., Motheramgari, K., Gallagher, C., Hagan, R., Kelly, P., Costello, A., Meiller, J., Floris, P., Zhang, L., Clynes, M., Bones, J., Barron, N., Clarke, C., 2020. Subphysiological temperature induces pervasive alternative splicing in Chinese hamster ovary cells. Biotechnol. Bioeng. 117, 2489–2503. https://doi.org/10.1002/bit.27365

UniProt Consortium, 2021. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 49, D480–D489. https://doi.org/10.1093/nar/gkaa1100

Walsh, G., 2018. Biopharmaceutical benchmarks 2018. Nat. Biotechnol. 36, 1136–1145. https://doi.org/10.1038/nbt.4305

Wang, J., Vasaikar, S., Shi, Z., Greer, M., Zhang, B., 2017. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. Nucleic Acids Res. 45, W130–W137. https://doi.org/10.1093/nar/gkx356

Wright, B.W., Yi, Z., Weissman, J.S., Chen, J., 2021. The dark proteome: translation from noncanonical open reading frames. Trends Cell Biol. S0962-8924(21)00226–9. https://doi.org/10.1016/j.tcb.2021.10.010

Xu, X., Nagarajan, H., Lewis, N.E., Pan, S., Cai, Z., Liu, X., Chen, W., Xie, M., Wang, W., Hammond, S., Andersen, M.R., Neff, N., Passarelli, B., Koh, W., Fan, H.C., Wang, Jianbin, Gui, Y., Lee, K.H., Betenbaugh, M.J., Quake, S.R., Famili, I., Palsson, B.O., Wang, Jun, 2011. The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. Nat. Biotechnol. 29, 735–741. https://doi.org/10.1038/nbt.1932

Zhang, H., Wang, Y., Wu, X., Tang, X., Wu, C., Lu, J., 2021. Determinants of genome-wide distribution and evolution of uORFs in eukaryotes. Nat. Commun. 12, 1076. https://doi.org/10.1038/s41467-021-21394-y

Zhang, P., He, D., Xu, Y., Hou, J., Pan, B.-F., Wang, Y., Liu, T., Davis, C.M., Ehli, E.A., Tan, L., Zhou, F., Hu, J., Yu, Y., Chen, X., Nguyen, T.M., Rosen, J.M., Hawke, D.H., Ji, Z., Chen, Y., 2017. Genome-wide identification and differential analysis of translational initiation. Nat. Commun. 8, 1749. https://doi.org/10.1038/s41467-017-01981-8

Zhang, S., Reljić, B., Liang, C., Kerouanton, B., Francisco, J.C., Peh, J.H., Mary, C., Jagannathan, N.S., Olexiouk, V., Tang, C., Fidelito, G., Nama, S., Cheng, R.-K., Wee, C.L., Wang, L.C., Duek Roggli, P., Sampath, P., Lane, L., Petretto, E., Sobota, R.M., Jesuthasan, S., Tucker-Kellogg, L., Reversade, B., Menschaert, G., Sun, L., Stroud, D.A., Ho, L., 2020. Mitochondrial peptide BRAWNIN is essential for vertebrate respiratory complex III assembly. Nat. Commun. 11, 1312. https://doi.org/10.1038/s41467-020-14999-2

Supplementary data

Results

Upstream open reading frames repress the translation efficiency of CHO cell mRNAs.
Download: https://app.box.com/s/j9afj0io2cp1w68ndrubwqo1kim5pvbk

Tables

Table S1: Cell densities at 72 hrs post-seeding for the NTS and TS sample groups for both the initiation and elongation experiments. A significant reduction in cell density was observed in the temperature shifted groups for both experiments.
Download: https://app.box.com/s/n3c8nwx25fmu7lku17bk18bo9fdh59bn

Table S2: Read pre-processing metrics for Ribo-seq and RNA-seq data. The number of reads removed following adapter trimming and quality assessment are included for each dataset. For the Ribo-seq data the number of reads eliminated following alignment to contaminating RNA species (rRNA, tRNA and snoRNA) as well as phasing analysis are shown.
Download: https://app.box.com/s/un8az9pkxgchoaz9amzusvl3rxp8f89y

Table S3: ORFs identified in this study. ORF-RATER was used to annotate ORFs using the 3 types of Ribo-seq data. This table contains both annotated and novel ORFs identified. For each ORF, the ORF-RATER ID, ORF-type, ORF-RATER score, gene symbol, gene name, transcript family, transcript ID, start codon, amino acids, whether the start or stop codons were previously annotated, transcript and genome coordinates and strand are included.
Download: https://app.box.com/s/o115ekhaht8fzhr9udr8kbja7phpho7y

Table S4: Adalimumab Host cell protein identifications. LC-MS/MS analysis identified 32 HCPs from adalimumab drug product, including 8 microproteins. Shown are the UniProt and ORF-RATER accessions, sequence coverage, number of peptides, length, Sequest score, and LFQs for 3 technical replicates.
Download: https://app.box.com/s/pxbgacaz17n735ouq2dox8ialz1r184e

Table S5: Significant alterations in RNA abundance and translational regulation following temperature shift. DESeq2 was used to identify changes in RNA abundance, RPF occupancy and translational efficiency for **(A-C)** canonical ORFs and **(D-F)** sORFs found in ncRNA. The ID assigned by plastid, the NCBI Gene ID, gene symbol, Gene name, baseMean of DESeq2 normalised counts, $\log_2$ p-value and BH adjusted p-value are shown for each gene.
Download: https://app.box.com/s/e8wfp451glmg6r1wwmg5lag54yr6fbod

Table S6: GO biological process enrichment analysis. (A) differentially expressed and **(B)** differentially translated genes**.** The GO ID, description, number of genes in category, overlap, enrichment ratio, p-value, FDR, and genes overlapping are shown for each overrepresented biological process.
Download: https://app.box.com/s/j4i07c7h0fg1iax44nvh1uzkyrnw1rx2

Table S7: LC-MS/MS analysis of the CHOK1GS proteome. (A) Detected proteins across the Day 4 and Day 7 samples, **(B)** novel ORF-RATER proteoforms identified and **(C)** differentially expressed proteins identified using proDA.
Download: https://app.box.com/s/k68rlke9tnsstg0335hifehnv715kqnv

Table S8: Database search settings for (A) CHOK1GS cell lysate and **(B)** adalimumab HCP analysis**.**
Download: https://app.box.com/s/992v70eyt2uajpynrfwlfbxq5hq954fl

Figures

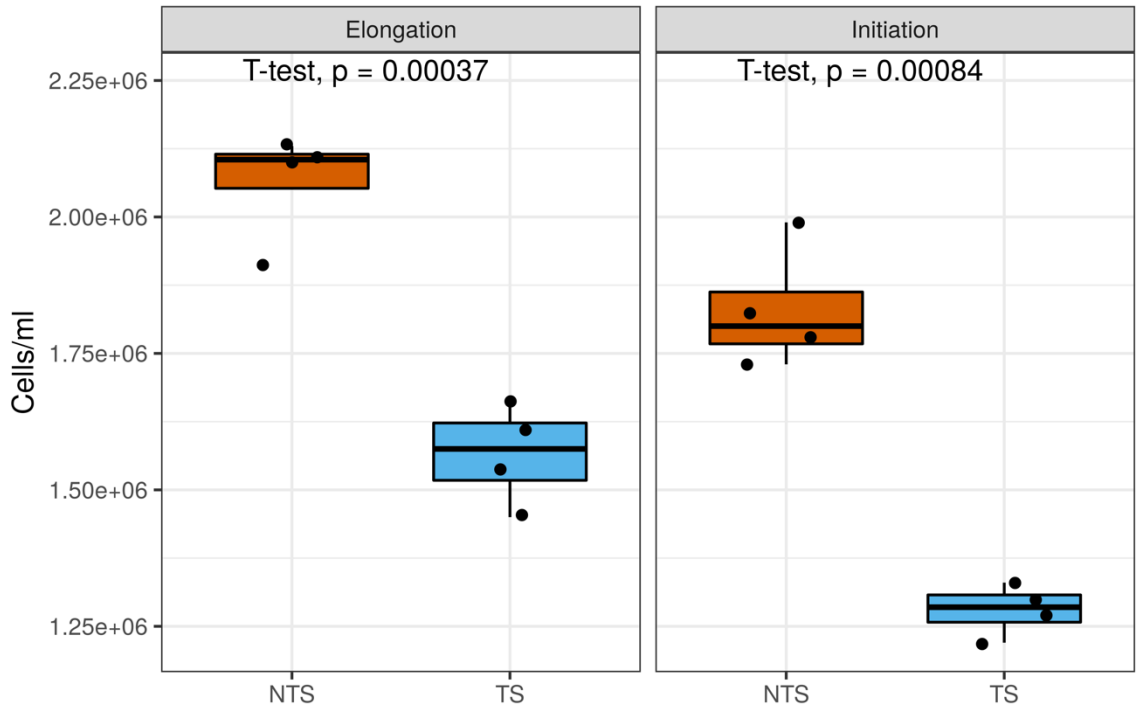


Figure S1: Reduction of cell culture temperature to 31°C decreases CHO cell growth rate. Separate cell culture experiments of the temperature shift model were carried out to generate samples for both elongation Ribo-seq (CHX and RNA-seq) and initiation Ribo-seq (Harr and ND). In both experiments, a significant decrease in cell density of ~25% (elongation) and 31% (initiation) for TS samples was observed 24hrs post-temperature shift (72hrs post-seeding).

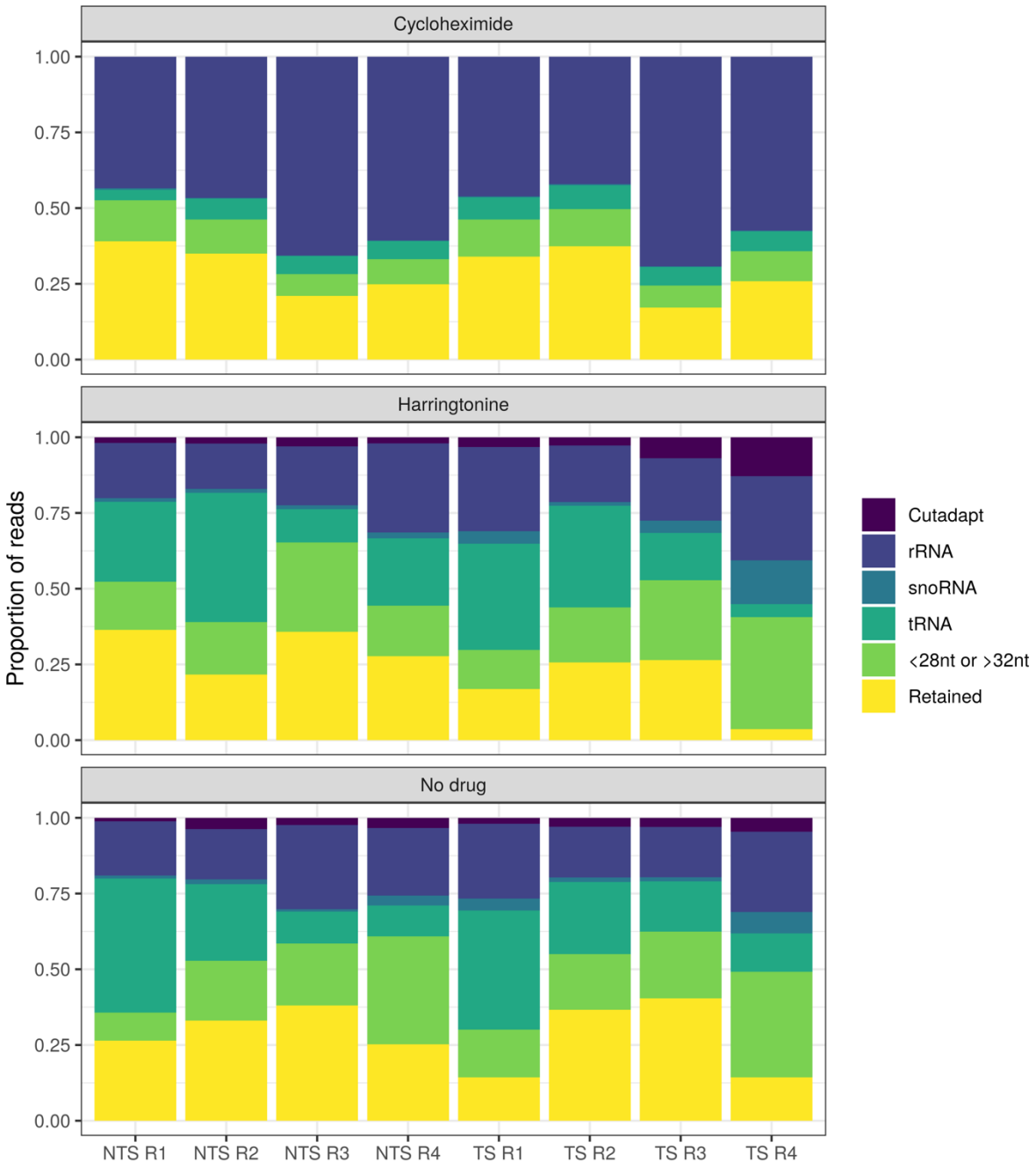
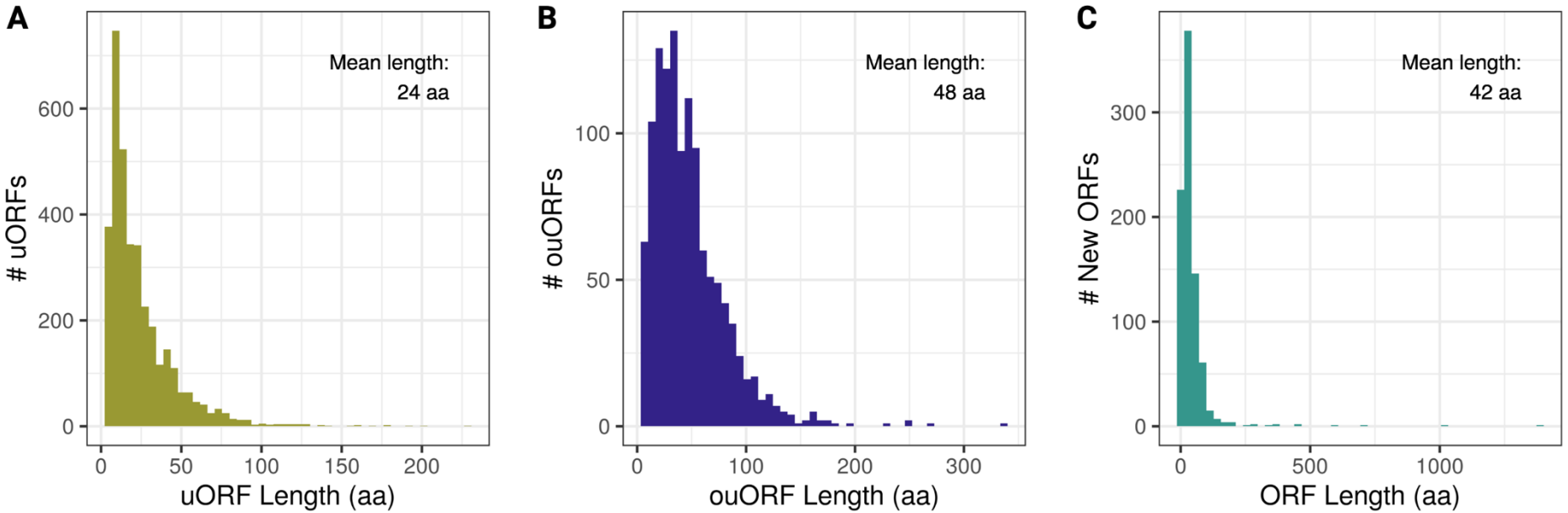
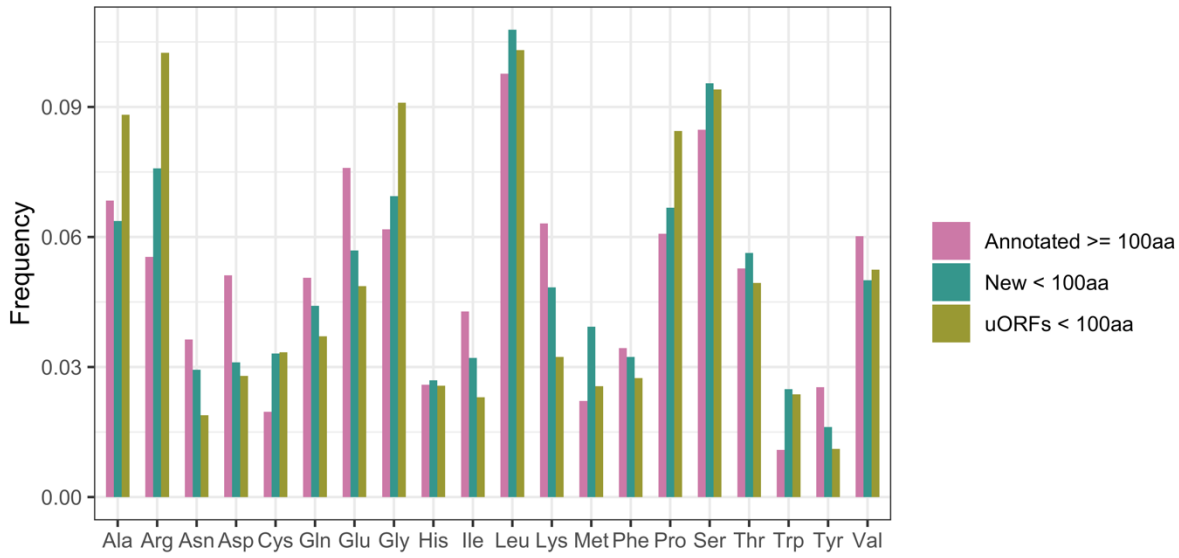


Figure S2: Pre-processing of Ribo-seq data. Prior to analysis adapters were removed from the raw sequencing reads using Cutadapt. *Note*: The Ribo-seq CHX data was obtained from the sequencing provider with adapter sequences removed. Reads mapping to contaminating RNA species (i.e., rRNA, snoRNA or tRNA) were filtered. Finally, only the reads lengths 28-31nt where 60% of reads were in frame with P-site offset =12 were retained for further analysis.



933

934 **Figure S3: Length distribution of novel Chinese hamster proteoforms classes comprised primarily of short open reading frame proteins.** >90% of **(A)** uORFs, **(B)** ouORFs and **(C)** New
935 ORFs identified by ORF-RATER were classified as short ORFs. The average length of each type is shown. Note New ORFs here are found in both protein coding and non-coding transcripts.



936

937 **Figure S4**: **Amino acid frequency of annotated and short ORFs.** The amino acid frequencies of the 20aas for uORFs (both
938 uORFs and ouORFs) and ncRNA sORFs along with annotated protein coding genes were determined, revealing differences
939 between each of the groups.

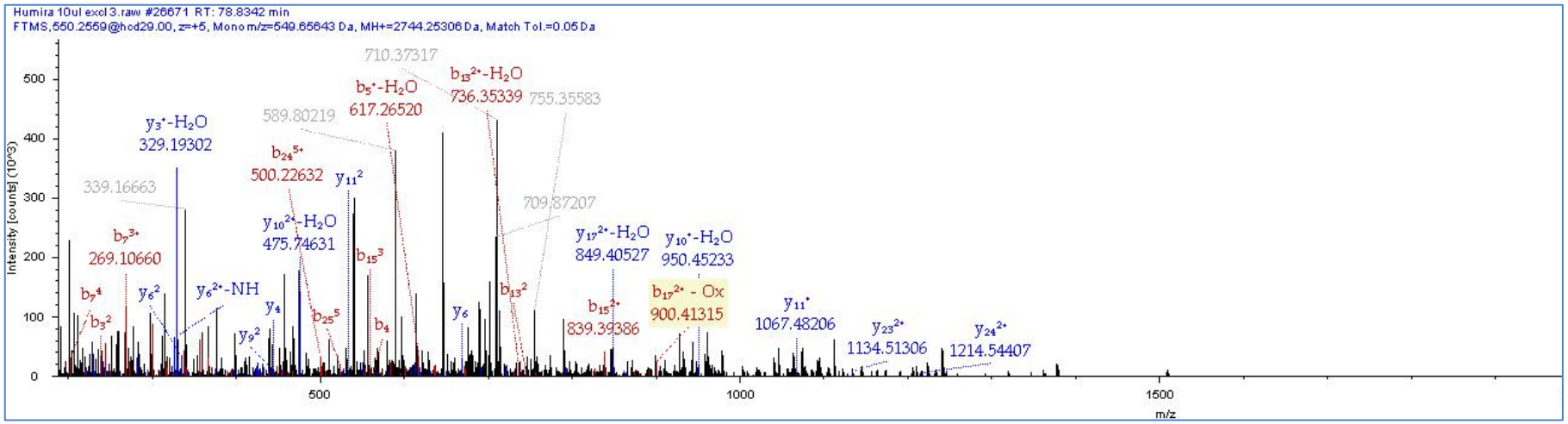


Figure S5 Annotated mass spectrum of peptide #1 of the XR_003481490.2_407599334_59aa microprotein found in the adalimumab drug product sample analysed in our laboratory.

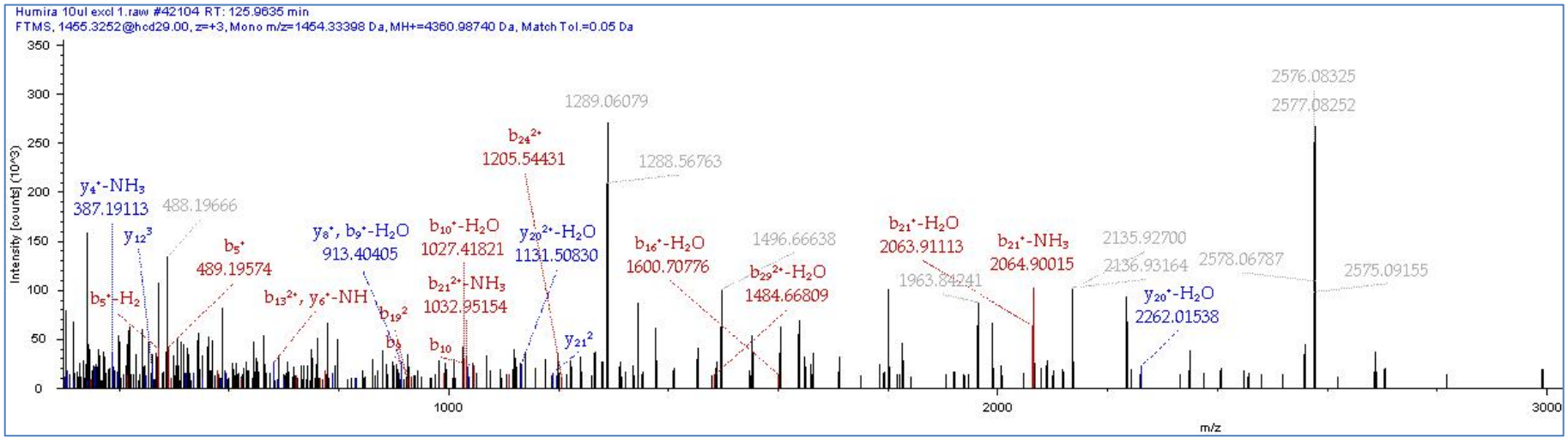


Figure S6 Annotated mass spectrum of peptide #2 of the XR_003481490.2_407599334_59aa microprotein found in the adalimumab drug product sample analysed in our laboratory.

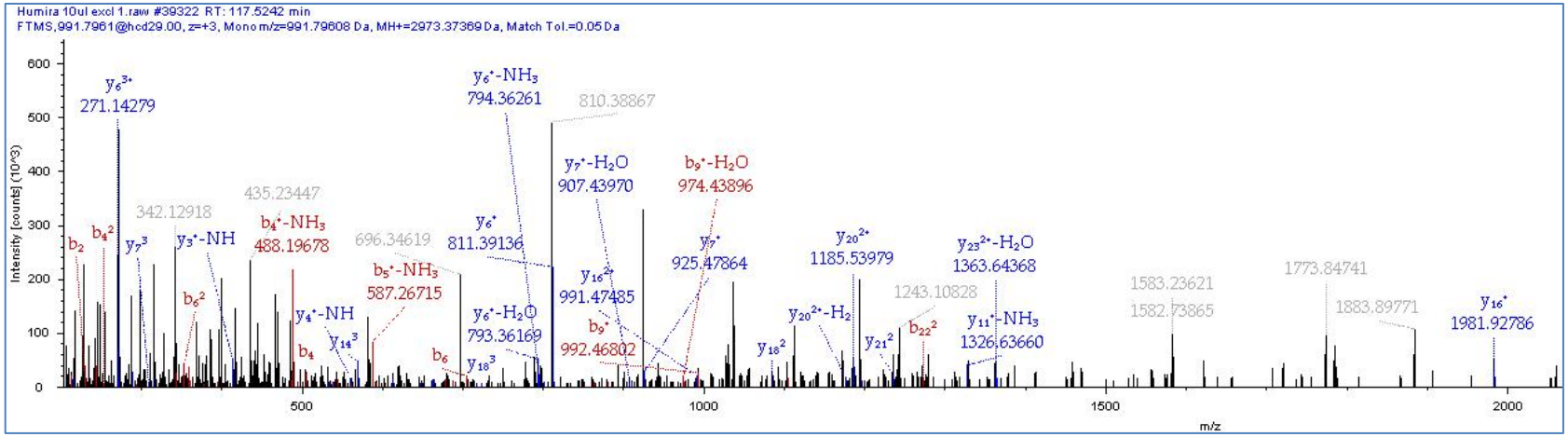


Figure S7 Annotated mass spectrum of peptide #3 of the XR_003481490.2_407599334_59aa found in the adalimumab drug product sample analysed in our laboratory.

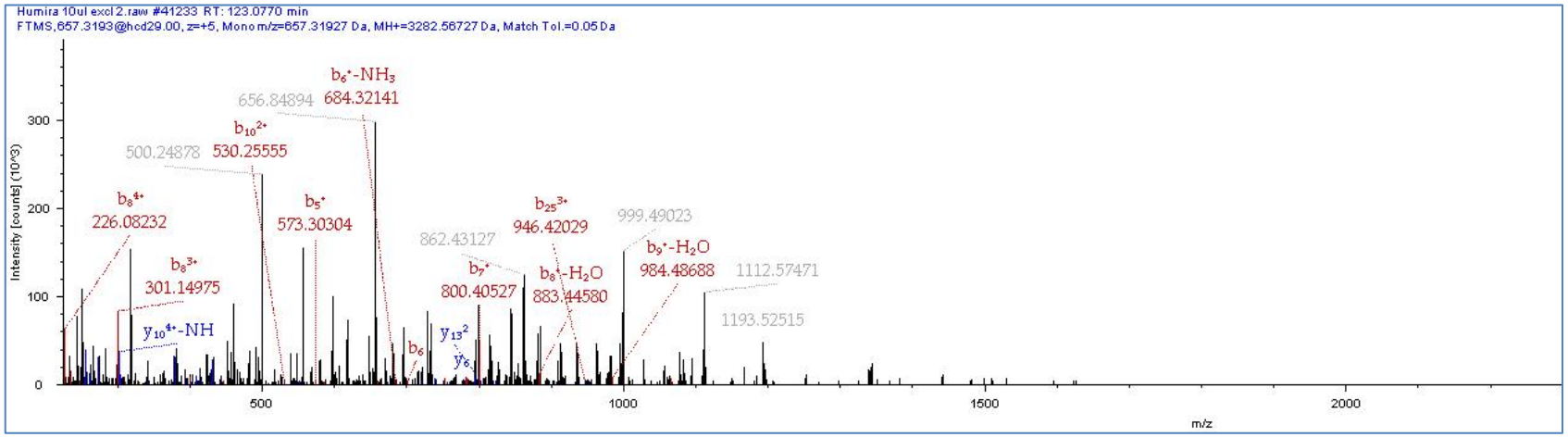


Figure S8 Annotated mass spectrum of peptide #4 of the XR_003481490.2_407599334_59aa found in the adalimumab drug product sample analysed in our laboratory.

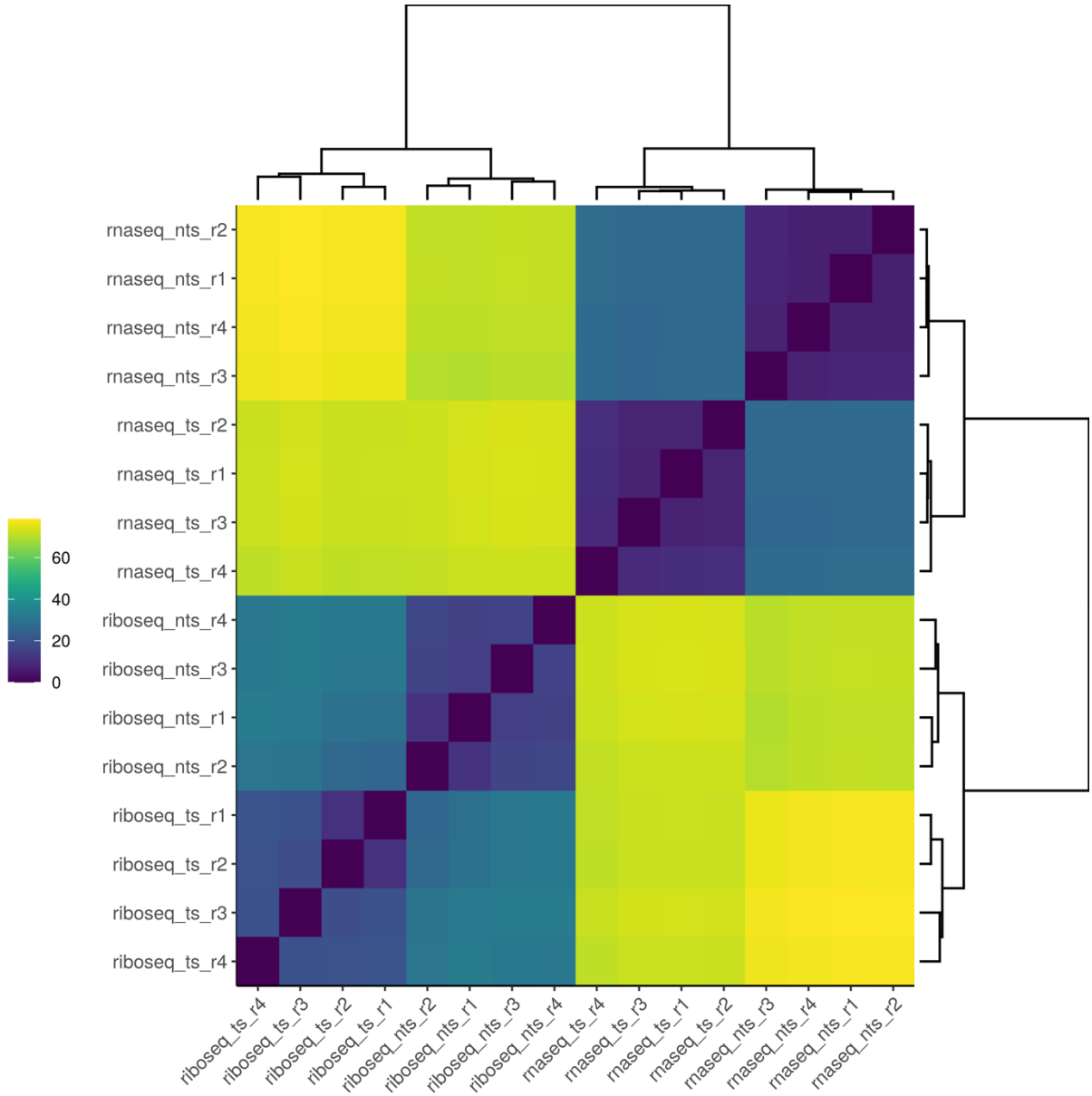
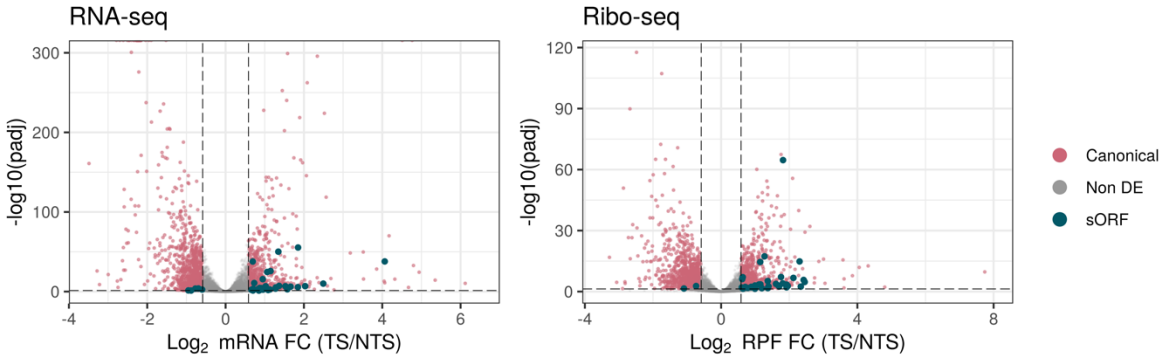


Figure 9: Hierarchical cluster analysis of RNA-seq and CHX Ribo-seq gene-level counts. While the most significant difference was between the Ribo-seq and RNA-seq data we also observed that there was a clear difference between the NTS and TS sample groups.



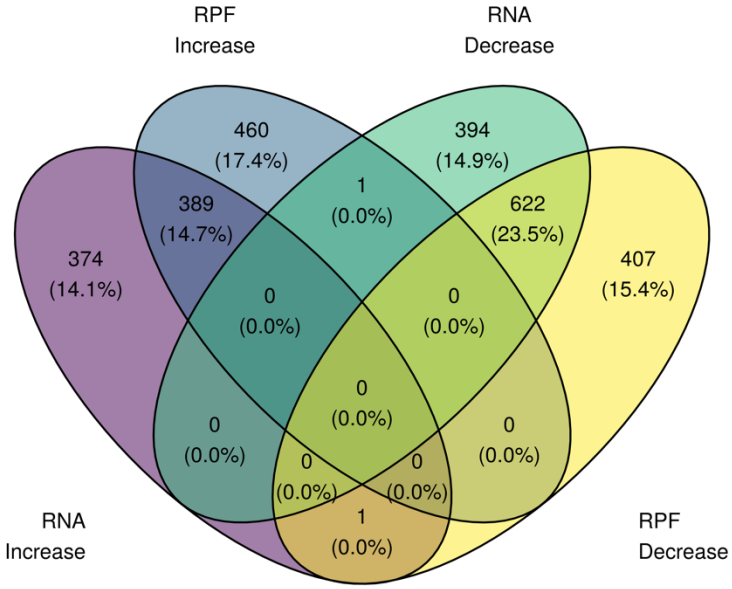


Figure S10: Differential expression and RPF occupancy for canonical and sORFs found in non-coding RNA. DESeq2 was utilised to identify differences in canonical and sORFs that occurred upon a reduction of cell culture temperature from separate analysis of RNA-seq and Ribo-seq data. A total of **(A)** 1,781 ORFs were found to be differentially expressed from the RNA-seq data and **(B)** 1,880 from the Ribo-seq data. **(C)** 1011 ORFs were found to change in the same direction in both datasets.

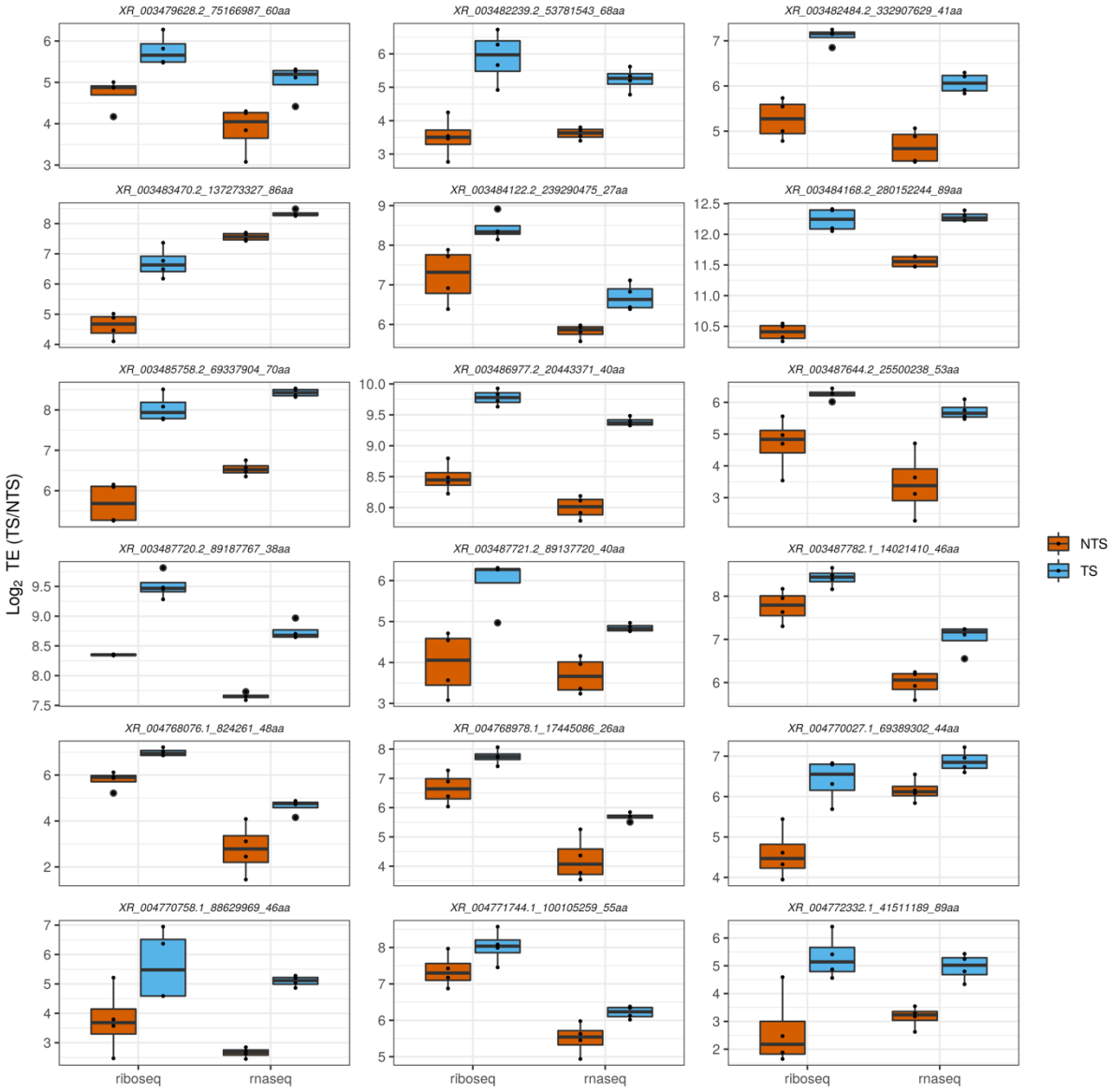


Figure S11: 18 sORFs were found to be upregulated at sub-physiological temperature in both the RNA-seq and Ribo-seq data.

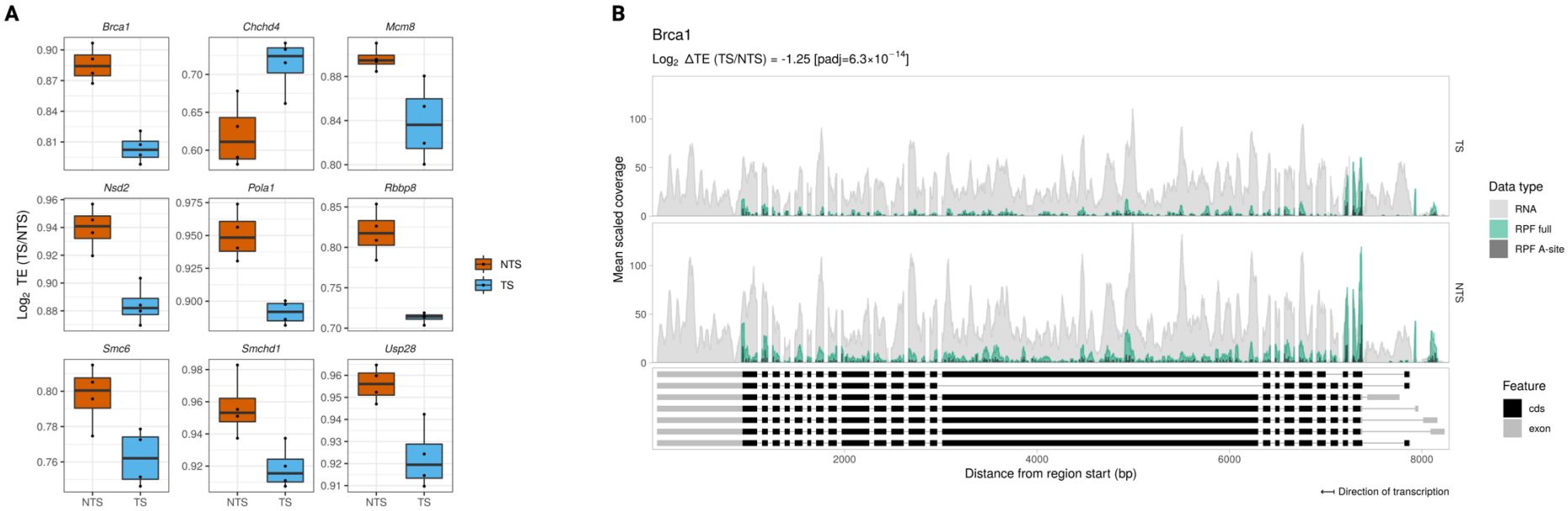


Figure S12: Differential translation efficiency of canonical ORFs involved in DNA repair. GO enrichment analysis revealed the significant overrepresentation of genes involved in the DNA repair. **(A)** Translation efficiency 9 of the 26 genes related in the DNA repair biological process including **(B)** *Brca1* were found to be altered by a reduction in cell culture temperature.