University of Nottingham
UK | CHINA | MALAYSIA

Bioinformatic investigation of downstream open reading frames and their role in post-transcriptional regulation of gene expression

Joe Tomlinson

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy

August 2025

Abstract

Downstream open reading frames (dORFs) are short open reading frames in 3' untranslated regions (3’ UTRs) of messenger RNAs (mRNAs) that are proposed to regulate translation. dORF translation has been suggested to increase the translation of a transcript. Ribosomes are present in 3’ UTRs, and this can be influenced by local cellular conditions. The importance of dORFs is highlighted by many diseases, including cancer, which are linked to dysfunction of translational regulation or post-transcriptional regulators.

This project used a bioinformatic approach and publicly available datasets to investigate translational regulation, focussing on dORFs and their role in translational regulation in cancers. Paired RNA sequence (RNAseq) and ribosome profiling (RP) datasets were used to investigate translational regulation of transcripts with and without dORFs and ribosome association with 3’ UTRs and dORFs. Paired healthy and cancer datasets were compared. The abundance of potential dORFs in 3’ UTRs was investigated and comparison of dORF nucleotide, dinucleotide, and trinucleotide composition with 3’ UTRs and coding sequences (CDSs) was determined. The conservation of dORFs was explored, including comparison with flanking sequences and 3’ UTRs more generally. Further RP datasets were analysed to determine whether different cell types, cellular conditions, and disease states influenced ribosomal association with dORFs and dORF-containing 3’ UTRs.

Generally, in healthy tissue, transcripts containing dORFs reported in Wu et al. (2020b) were translationally upregulated compared to transcripts without. The activity of these dORFs reduced in tumour tissue, which is unlikely to be explained by transcript alterations, such as 3’ UTR truncation, in the tumour datasets. dORFs were found to share more similar nucleotide, dinucleotide and trinucleotide composition with the noncoding 3’ UTRs than CDSs. dORFs with ribosome association are widespread, however most dORFs had low ribosomal association and little evidence of translational regulation, in contrast to the dORFs reported in Wu et al. (2020b). dORF ribosomal association was varied and likely influenced by different cell types, disease states and local conditions. The biological importance of dORFs is emphasised by their conservation across species. dORFs are more conserved than flanking sequences and 3’ UTRs generally. Some dORFs may be translational regulators, whereas others may have activity through an encoded peptide. Further investigation is needed to understand the mechanism of action and function of dORFs, which are a novel, interesting, and important feature of mRNAs.

Acknowledgements

I would like to thank Keith Spriggs, my primary supervisor, for his support and advice throughout my PhD, particularly during the challenging points. Thank you for being encouraging and approachable, it helped develop my range of research skills for the future. Additionally, thank you to my secondary supervisors, Adam Blanchard and Anna Piccinini, for their support during the project, especially for their additional insight into bioinformatics and cellular biology.

Thank you to the KAS group members for their useful debate and discussion around my research and the wider field. Thank you to all members of the GRRB lab for their useful questions about my research and for making my time in the office enjoyable. Thank you to the BBSRC DTP team and the School of Pharmacy for the development opportunities and support throughout my project.

I would like to thank my family and friends for their invaluable support and interest in my work, they have helped me enjoy my PhD. Finally, I want to thank my wife, Jessica, for her unwavering belief in me and her constant support. She is an inspiration to me and always supported me in challenging times.

Table of Contents

List of Figures

List of Tables

Abbreviations

aa	Amino Acid
A	Adenine
A site	Aminoacyl Site
ABC	ATP Binding Cassette
ABCE1	ATP-binding cassette sub-family E member 1
APA	Alternative Polyadenylation
ARE	AU-rich Elements
AUG dORFs	dORFs using an AUG start codon
C	Cytosine
C-Terminus	Carboxyl-terminus
CDS	Coding Sequence
circRNA	Circular RNA
CPE	Cytoplasmic Polyadenylation Elements
CPEB	Cytoplasmic Polyadenylation Element Binding Protein
CpG	Cytosine Preceding Guanine
dORF	Downstream Open Reading Frame
eIF	Eukaryotic Initiation Factor
eRF	Eukaryotic Translation Termination Factor
G	Guanine
G domain	GTP-binding Domain
HC dORFs	Highly Conserved dORFs
IGV	Integrative Genome Viewer
IRE	Iron Response Element
IRES	Internal Ribosome Entry Segment
ISR	Integrated Stress Response
lncRNA	Long Noncoding RNA
m^6A	N6 Methyladenosine
MAPK	Mitogen Activated Protein Kinase
Met-tRNAi	Methionyl-initiator tRNA
MHC	Major Histocompatibility Complex
miRISC	miRNA-induced Silencing Complex
miRNA	Micro-RNA
MRE	miRNA Response Element
mRNA	Messenger RNA
MS	Mass Spectrometry
MSVW dORFs	Mass Spectrometry Validated Wu dORFs
mTOP	Mammalian or Mechanistic Target of Rapamycin
N-Terminus	Amino-Terminus
NBD	Nucleotide-binding Domain
NMD	Nonsense-mediated Decay
OCC	Oral cavity carcinoma
ORF	Open Reading Frame
P site	Peptidyl Site
PI3K	Phosphatidylinositol 3-kinase
PIC	Preinitiation Complex
PABP	Poly(A) Binding Protein
Poly(A)	Polyadenosine

RNA	Ribonucleic Acid
RNAseq	RNA Sequence
RP	Ribosome Profiling
RPF	Ribosome Profiling Footprint
RPKM	Reads Per Kilobase of transcript, per Million mapped reads
SD	Standard Deviation
SEP	Short Encoded Peptide
sORF	Short Open Reading Frame
SRA	Sequence Read Archive
TC	Ternary Complex
tRNA	Transfer RNA
uORF	Upstream Open Reading Frame
U	Uracil
UTR	Untranslated Region
Wu dORFs	dORFs published by Wu et al. (2020b)

Chapter 1: Introduction

Downstream open reading frames (dORFs) are novel gene regulators. Regulation of gene expression is essential in healthy biological function and organism survival. Intricate machinery, including ribosomes, RNAs, proteins and amino acids, is required to accurately produce protein (Hershey, Sonenberg and Mathews, 2019). Regulation can occur at different stages of gene expression. For example, transcription, which is the production of messenger RNA (mRNA) using DNA as a template, or translation, which is the production of a protein using mRNA as a template. Regulation of translation plays an important role in diseases and response to cell stress.

This research focuses on translational regulation. There are many translational, or post-transcriptional, regulators, including upstream open reading frames (uORFs), internal ribosome entry segments (IRESs) and micro-RNAs (miRNAs). These regulators are demonstrably important in regulating gene expression and highlight the significance of translational regulation. This research investigated ribosome occupancy and possible translation within 3’ untranslated regions (UTRs), a noncoding region of mRNAs, focussing on a recently reported translational regulator, dORFs, characterised by the presence of open reading frames (ORFs) within 3’ UTRs (Wu *et al.*, 2020b). This project used bioinformatic tools and publicly available datasets, evidencing the strength of bioinformatic analysis and the importance of data sharing in health research.

1.1 mRNA biogenesis

Translation can only occur once mRNA has been transcribed from DNA. Following transcription (itself tightly regulated), mRNAs undergo several processing steps. mRNAs are capped at the 5’ end and polyadenylated at the 3’ end, alongside internal modifications such as splicing and methylation. mRNA is a single-stranded sequence of nucleotides that contains both coding, and noncoding, regions. Mature mRNA consists of a 5’ cap preceding the 5’ UTR, followed by the coding sequence (CDS), before the 3’ UTR and polyadenylated tail (poly(A) tail). The 5’ cap is important in the initiation of translation through recognition by ribosomal machinery, and also to prevent degradation of the mRNA. The 5’ UTR is a noncoding region before the CDS that interacts with ribosomal machinery and contains regulatory elements that can influence mRNA stability and translation. The CDS is the coding region of mRNA that is made up of codons and is translated by ribosomes into the protein. The 3’ UTR also contains regulatory elements. The 3’ UTR is a noncoding region that follows the CDS and can influence mRNA stability and translation. The poly(A) tail is important to prevent mRNA degradation and also to support nuclear export and translation initiation. Importantly, in addition to transcription and mRNA abundance, post-transcriptional mechanisms involving interactions of mRNAs with proteins and other RNAs also affect the quantity and activity of proteins in cells (Mitchell and Parker, 2014; Ivanov, Kedersha and Anderson, 2019; Peer *et al.*, 2019). mRNA processing steps and modifications can influence translation and different mRNAs can have varying translation rates (Biswas *et al.*, 2019; Ingolia, Hussmann and Weissman, 2019). Alternative promoters and splicing can produce different transcripts with varied translational efficiencies, highlighting the interplay between mRNA processing and translational regulation (Tazi, Bakkour and Stamm, 2009). These processing steps can affect 3’ UTRs and any regulators present.

1.2 Composition of 3’ UTRs

The composition and length of 5’ UTRs and 3’ UTRs varies from dozens to several thousand nucleotides (Pesole et al., 2001; Mignone et al., 2002). In humans, 3’ UTRs (often over 1000 nucleotides) are usually longer than 5’ UTRs (around 100-200 nucleotides) (Pesole et al., 2001). When comparing guanine (G) and cytosine (C) composition, 3’ UTRs have reduced GC composition compared to 5’ UTRs (Pesole et al., 2001; Larizza et al., 2002; Mignone et al., 2002). Increased GC composition is often associated with greater secondary structure, due to G-C, compared to adenine (A)–Uracil (U), base pairing having greater stability with increased hydrogen bonding (Shabalina, Ogurtsov and Spiridonov, 2006), suggesting 3’ UTRs may have reduced secondary structure compared to 5’ UTRs. Longer 3’ and 5’ UTRs are often associated with reduced GC nucleotide composition (Pesole, Bernardi and Saccone, 1999; Pesole et al., 2001; Mignone et al., 2002).

Transcripts with premature stop codons can be targeted for degradation by nonsense-mediated decay (NMD), preventing translation of truncated proteins (Chang, Imam and Wilkinson, 2007; Silva and Romão, 2009; Kervestin and Jacobson, 2012; Schweingruber *et al.*, 2013; Wei-Lin Popp and Maquat, 2013; Zahdeh and Carmel, 2016). NMD also regulates gene expression through interaction with alternative splicing (Lewis, Green and Brenner, 2003; Wollerton *et al.*, 2004; McGlincy and Smith, 2008; Ge and Porse, 2014; Zahdeh and Carmel, 2016). NMD has importance in healthy cellular function and dysfunctional NMD is associated with many genetic disorders and cancers (Frischmeyer and Dietz, 1999; Holbrook *et al.*, 2004; Khajavi, Inoue and Lupski, 2006; Zahdeh and Carmel, 2016). Nucleotide composition is suggested to trigger NMD through increased mRNA structure in the 3' UTR or around the stop codon, such as increased G density upstream of the stop codon (Zahdeh and Carmel, 2016).

Cytosines preceding guanine residues (CpG) are the most common location for DNA methylation in mammalian cells (Portela and Esteller, 2010; Jang *et al.*, 2017). CpG dinucleotides are more common in 5' UTRs than 3' UTRs (Pesole *et al.*, 1997, 2001). DNA methyltransferases methylate CpG sites in DNA at the fifth carbon of the pyrimidine ring of cytosine residues (Jaenisch and Bird, 2003; Law and Jacobsen, 2010; Smith and Meissner, 2013; Jang *et al.*, 2017). Methylation of CpG dinucleotides is important as methyl-C undergoes deamination, causing C-to-T mutations which over time reduces the incidence of CpG dinucleotides in genomic regions that are not subject to strong evolutionary selection (Ehrlich and Wang, 1981; Karlin and Mrázek, 1997; Takata *et al.*, 2017). Methylation of CpG sites can both promote and repress gene expression, influenced by the site location (Portela and Esteller, 2010; Jang *et al.*, 2017). Transcription can be reduced when promoters or enhancers contain methylated CpG sites, preventing binding of transcription activators (Esteller, 2008; Kuroda, 2009; Lister *et al.*, 2009; Chodavarapu *et al.*, 2010; Portela and Esteller, 2010; Shukla *et al.*, 2011; Jang *et al.*, 2017). Increased GC content in 5' UTRs, compared to 3' UTRs, may relate to CpG islands; CpG islands are defined as where CpG dinucleotides are grouped (Esteller, 2008), and are genomic regions of with: 50% or more GC nucleotides, 200 nucleotides or more, occurring above the expected CpG frequency (Portela and Esteller, 2010). CpG islands are regions of high CpG content that are retained due to evolutionary pressure (Antequera and Bird, 1999; Antequera, 2003; Illingworth and Bird, 2009).

1.3 Conservation of 3' UTRs

This research will compare 3' UTRs between species. 3' UTR conservation has been reported since the early 1980s (Fournier *et al.*, 1994; Knee, Pitcher and Murphy, 1994; Silverman, 1994; Lipman, 1997). Highly conserved 3' UTR regions could be involved in regulatory processes, as RNA or protein binding sites (Bashirullah, Cooperstock and Lipshitz, 1998; Conne, Stutz and Vassalli, 2000; Grzybowska, Wilczynska and Siedlecki, 2001; Pesole *et al.*, 2001; Mignone *et al.*, 2002; Shabalina *et al.*, 2004). Small AU-rich elements (AREs) are associated with mRNA degradation and are often seen in conserved 3' UTR regions (Ho *et al.*, 1995; Lipman, 1997). Although less conserved than CDSs, in vertebrates, 3' UTRs are more conserved than 5' UTRs (Siepel *et al.*, 2005; Litterman *et al.*, 2019). In vertebrates, 3' UTRs are conserved best in development-associated genes and those with lower GC composition, also seen in genes required for cell survival (Duret, Dorkeld and Gautier, 1993; Shabalina *et al.*, 2003; Litterman *et al.*, 2019). In zebrafish embryos, mRNA destabilization is associated with GC-rich 3' UTR elements and 3' UTR sequences with greater GC composition (Rabani *et al.*, 2017; Litterman *et al.*, 2019). The link between GC composition and mRNA structure suggests that reduced 3' UTR structure is associated with effective protein production (Litterman *et al.*, 2019).

The nucleotides immediately before and after the stop codon have greater conservation (Shabalina *et al.*, 2004). Whereas, the 30 or so nucleotides following the stop codon, in particular UGA and UAG, have greater GC content and less conservation than the remaining 3' UTR (Shabalina *et al.*, 2004). UGA is the most common stop codon in mammals (Jacobs *et al.*, 2002; Shabalina *et al.*, 2004), and this increased flanking GC content may help release ribosomes and prevent further scanning (Shabalina *et al.*, 2004). There is also greater conservation before the poly(A) tail, in the 10-20 nucleotides following the AAUAAA motif, which is found in the polyadenylation signal to support 3' UTR cleavage and mRNA polyadenylation (Proudfoot and Brownlee, 1976; Tian and Graber, 2012; Shi and

Manley, 2015), and this region often has greater GU content (Shabalina *et al.*, 2004). Humans have the longest 3' UTRs and as species become more distantly related the length of 3' UTRs reduces to an average of 400 nucleotides in invertebrates or 200 nucleotides in plants (Pesole et al., 2002; Sood et al., 2006; Mayr, 2016; Wang et al., 2019; Hong and Jeong, 2023)

1.4 Current Understanding of Translation

Translation can be described in four stages: initiation, elongation, termination, and ribosome recycling. The major eukaryotic translation components are 40S and 60S ribosomal subunits, which combine to perform translation (Hershey, Sonenberg and Mathews, 2019). To begin translation, the ternary complex (TC) assembles from methionyl-initiator tRNA (Met-tRNA$_i$), GTP, and eukaryotic initiation factor 2 (eIF2) (Figure 1.1) (Renz, Valdivia Francia and Sendoel, 2020). Following TC formation, the 43*S* preinitiation complex (PIC) forms from the 40S ribosomal subunit, eIFs 1, 1A, 3, and 5, and the TC (Figure 1.1) (Renz, Valdivia Francia and Sendoel, 2020). The PIC is recruited to the 5' UTR cap by the eIF4F complex, containing eIFs 4E, 4G, and 4A (Sonenberg *et al.*, 1978; Joshi, Yan and Rhoads, 1994; Haghighat *et al.*, 1995; Mader *et al.*, 1995; Jackson, Hellen and Pestova, 2010; Hershey, Sonenberg and Mathews, 2019; Bartish *et al.*, 2023). The PIC associates with the 5' cap, before scanning the 5′ UTR in the 5′ to 3′ direction to find the start codon (Merrick and Pavitt, 2018; Renz, Valdivia Francia and Sendoel, 2020). The start codon base pairs with the Met-tRNA$_i$ anticodon, preventing further PIC scanning and triggering release of many initiation factors, allowing the 60S ribosomal subunit to join, forming the 80S initiation complex ready for the elongation stage (Figure 1.1) (Merrick and Pavitt, 2018). The flanking Kozak consensus sequence (ACCAUGG) surrounds the start codon and supports translation initiation (Kozak, 1987). AUG is overwhelmingly the most common start codon, however, non-AUG translation start codons, including CUG, GUG, UUG, can also initiate translation (Ingolia, Lareau and Weissman, 2011; Brar *et al.*, 2012; Arribere and Gilbert, 2013; Sendoel *et al.*, 2017; Na *et al.*, 2018; Renz, Valdivia Francia and Sendoel, 2020). The Kozak consensus may be more influential with non-AUG start codons, stabilizing the PIC interaction near the start codon (Na *et al.*, 2018).

Once the initiation complex has formed, with the Met-tRNA$_i$ in the 80S ribosome peptidyl (P) site, another tRNA is selected and bound to the subsequent codon, in the ribosomal aminoacyl (A) site. The ribosome orchestrates the methionyl group transfer from the P site Met-tRNA$_i$ to the A site tRNA, forming the first peptide bond. The ribosome moves along the mRNA transferring the peptidyl-tRNA from the A site, into the P site, known as translocation, before another tRNA then base pairs with the mRNA codon in the A site, continuing protein production (Dever, Dinman and Green, 2018). Elongation factors support the repetitive amino acid-tRNA binding, peptide bond formation, and translocation steps (Dever, Dinman and Green, 2018).

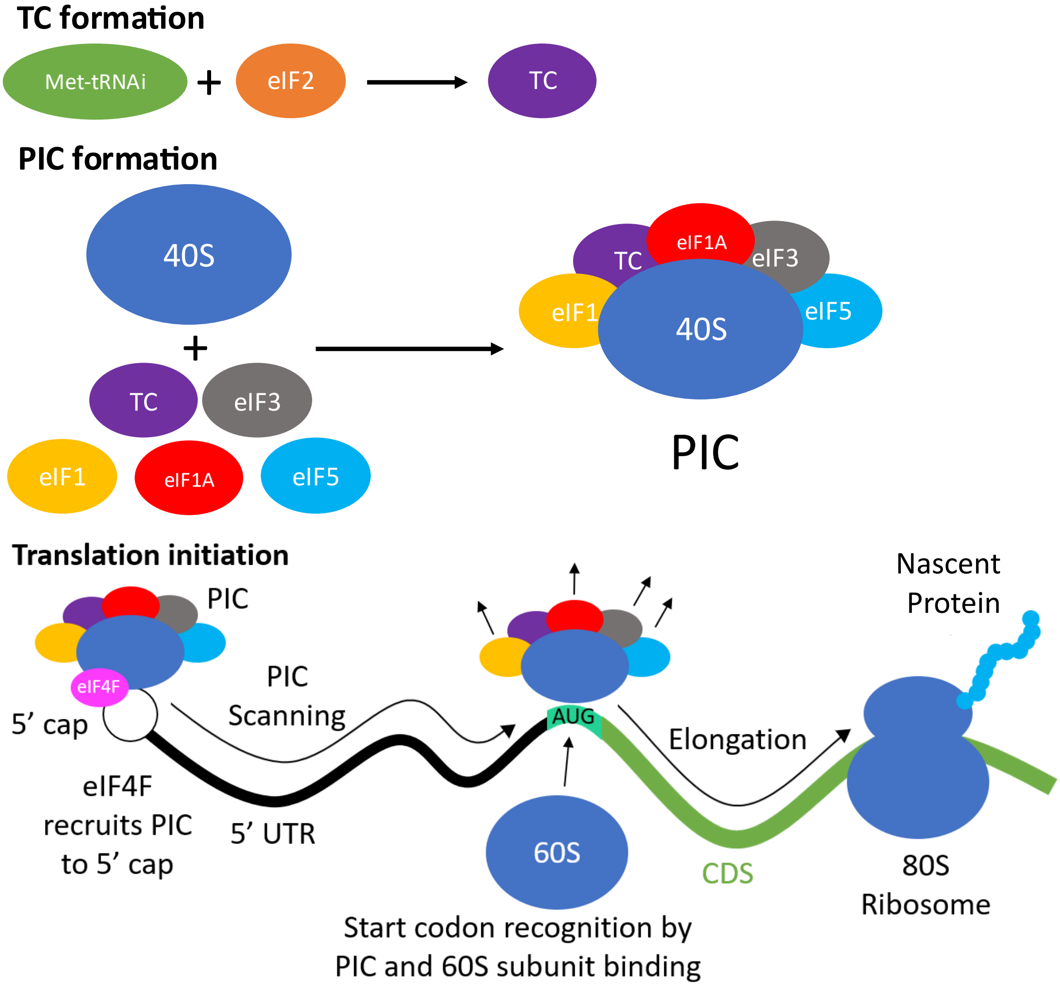

Figure 1.1: Model of translation initiation. *The ternary complex (TC) forms from methionyl-initiator tRNA (Met-tRNAi) and eukaryotic initiation factor 2 (eIF2). The 43S preinitiation complex (PIC) forms by combining the 40S ribosomal subunit with the TC and eIFs 1, 1A, 3, and 5. The PIC binds to the mRNA 5' cap, through interaction with the eIF4F complex and scans the 5' untranslated region (UTR), in a 5' to 3' direction, until a start codon (AUG) is found. Once the start codon is found many initiation factors release from the PIC, allowing the 60S ribosomal subunit to combine to form the 80S ribosome ready to translate the coding sequence (CDS).*

Termination occurs at one of the three stop codons (UAA, UAG, and UGA) (Mccaughan *et al.*, 1995) (Figure 1.6). When a stop codon enters the A-site of the ribosome, a protein release factor binds into the A site, leading to termination of protein synthesis (Hellen, 2018). Recognition of a stop codon is likely to involve multiple processes, including conformational changes in the ribosomal A site and eukaryotic release factor 1 (eRF1) motifs alongside eRF3 driven changes (Fan-Minogue *et al.*, 2008; Wong *et al.*, 2012; Kryuchkova *et al.*, 2013). eRF1 and eRF3 help to mediate termination (Alkalaeva *et al.*, 2006). eRF1 uses an amino-terminal domain to recognise the ribosome A site stop codon (Bertram *et al.*, 2000), and a middle domain to promote protein release, using a GGQ motif (Frolova *et al.*, 1999). The eRF1 carboxy (C)-terminal domain allows binding to eRF3 and ATP-binding cassette sub-family E member 1 (ABCE1) (Song *et al.*, 2000; Mantsyzov *et al.*, 2010). eRF1 can independently allow protein release, however, eRF3 increases this function (Alkalaeva *et al.*, 2006). The mechanism of protein release requires further investigation, however, an extended eRF1 conformation may use the GGQ motif to assist cleavage between the P-site peptidyl-tRNA and nascent protein (Preis *et al.*, 2014; Brown *et al.*, 2015; Muhs *et al.*, 2015).

After termination ribosomes are recycled, ready to translate again. Interactions with additional proteins dissociates tRNA and mRNA from ribosomes before ribosomes are broken down into subunits, known as ribosome recycling (Hellen, 2018). ABCE1 induces recycling of post-termination ribosomes, vacant 80S ribosomes, and stalled ribosomal elongation complexes, following identification by Hbs1/Pelota (Pisarev *et al.*, 2010; Franckenberg, Becker and Beckmann, 2012; Jackson, Hellen and Pestova, 2012). Recycling of post-termination complexes by ABCE1 also requires eRF1 in the A site (Pisarev *et al.*, 2010). ABCE1 has twin nucleotide-binding domains (NBDs) with two nucleotide-binding sites found in an open state while nucleotide-free or bound to ADP (Barthelme *et al.*, 2007, 2011; Karcher, Schele and Hopfner, 2008), transitioning to a closed state when ATP binds (Heuer *et al.*, 2017). Cycling of ATP Binding Cassette (ABC) proteins between closed and open states is suggested to cause conformational changes, leading to structural changes in associated domains or macromolecules (Rees, Johnson and Lewinson, 2009). This cycling could cause ribosomal splitting through ABCE1 conformational changes destabilising ribosomal inter-subunit bridges (Hellen, 2018). However, the recycling mechanism requires further understanding of potential conformational changes and the role of ATP in ABCE1 NBDs (Hellen, 2018).

1.5 The Closed Loop Model and 5’-3’ Interactions

The 3’ UTR is hardly mentioned in the translation process described previously; however, translation is complex. 3’ UTR association with translation involves mRNA structure and 3’ UTR interactions with RNA binding proteins. RNA and protein interactions between the 5′ and 3′ ends form a ‘closed loop’ of mRNA, proposed to be involved in determining mRNA translation efficiency and decay (Gallie, 1991; Wells *et al.*, 1998; Kahvejian *et al.*, 2005; Amrani, Sachs and Jacobson, 2006; Amrani *et al.*, 2008; Chen and Shyu, 2011). The ‘closed loop model’ interactions include: 5′ cap binding to the cap binding protein eIF4E, eIF4E interacting with eIF4G, eIF4G binding to the poly(A) binding protein (PABP), and PABP identifying the 3′ poly(A) tail (Figure 1.2) (Vicens, Kieft and Rissland, 2018). The closed loop allows close proximity of 5′ – 3′ ends, promoting transfer of regulatory information from the 3’ end, into regulatory effects at the 5′ end, called 5′–3′ communication (Vicens, Kieft and Rissland, 2018). The close proximity of 5’ and 3’ ends could promote re-utilization of terminating ribosomes, moving from the 3’ to 5’ end (Thompson and Gilbert, 2017; Vicens, Kieft and Rissland, 2018; Fakim and Fabian, 2019; Pelletier and Sonenberg, 2019). Generally, this hypothesis is accepted, however, it has never been proven experimentally, due to difficulties in monitoring individual ribosomes (Alekhina *et al.*, 2020).

5′–3′ communication is important in controlling mRNA function and fate. The poly(A) tail is important in mRNA translation and decay, where poly(A) tail length is associated with greater cap-dependent translation initiation (Vicens, Kieft and Rissland, 2018). During early embryonic development, maternally derived mRNAs have shortened poly(A) tails, which on fertilization, activate their translation through poly(A) tail lengthening (Gebauer *et al.*, 1994; Barkoff, Ballantyne and Wickens, 1998). Shortening of the poly(A) tail decreases translational efficiency and targets mRNAs for decay through 5’ decapping, allowing exonucleic degradation (Muhlrad, Decker and Parker, 1994; Yamashita *et al.*, 2005). mRNA decay begins with deadenylation, where 3’ exonuclease removes the poly(A) tail, allowing hydrolytic removal of the 5′ cap (Decker and Parker, 1993; Couttet *et al.*, 1997; Bönisch *et al.*, 2007). eIF4E is suggested to prevent cap hydrolysis (Schwartz and Parker, 2000), meaning stabilization of eIF4E, through the closed loop, may help to prevent cap hydrolysis.

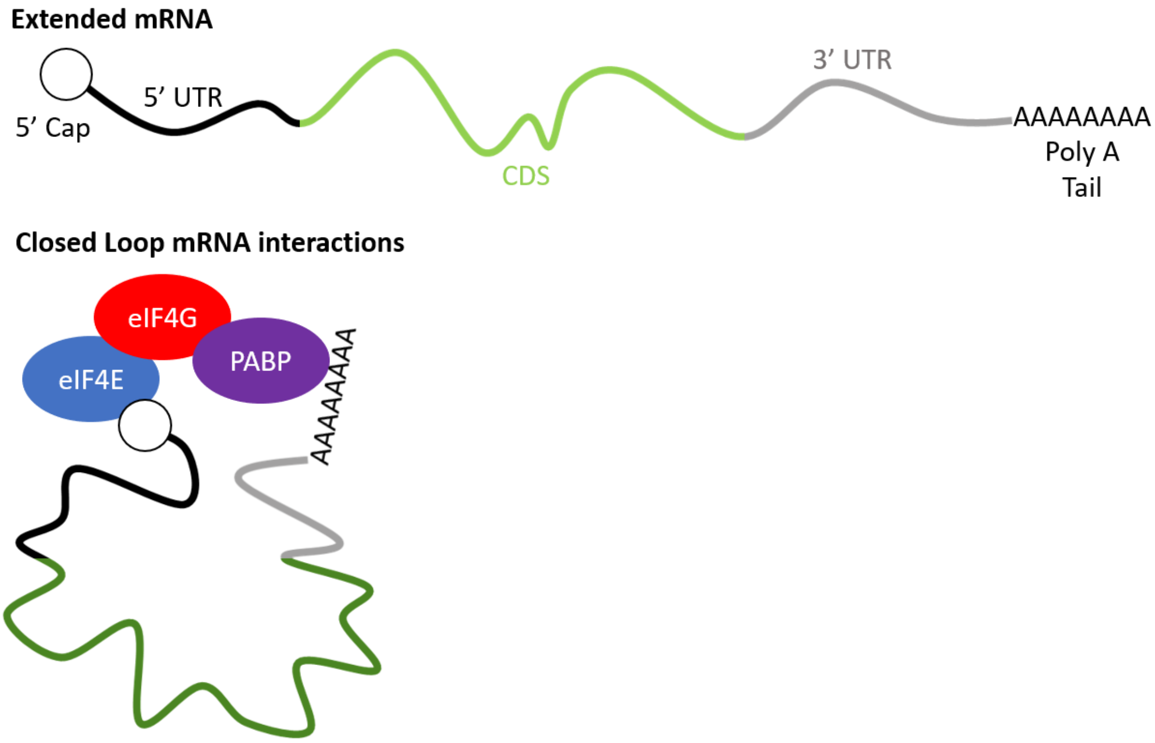

Figure 1.2: Representation of extended mRNA structure and interactions forming the closed loop model. *The extended structure highlight mRNA regions and shows separated 5' and 3' ends. The closed loop model brings 5' and 3' mRNA ends into close proximity through a series of interaction including: 5′ cap binding to the cap binding protein eukaryotic initiation factor 4E (eIF4E), eIF4E interacting with eIF4G, eIF4G binding to the poly(A) binding protein (PABP), and PABP identifying the 3′ poly(A) tail.*

Although widely accepted, there are some considerations with the closed loop model. Although potentially important, it remains unknown if the RNA and protein interactions enables communication between the ends (Vicens, Kieft and Rissland, 2018). Yeast cells can survive disruption of the PABP–eIF4G interaction (Kessler and Sachs, 1998; Park *et al.*, 2011), and an eIF4G RNA recognition motif can bind 3′ UTRs, forming a loop without the PABP–eIF4G interaction (Park *et al.*, 2011). Also unexplained is whether closed loop interactions could occur between different mRNAs (Vicens, Kieft and Rissland, 2018). Typical displays of mRNA should change, instead of horizontal lines with distant ends, mRNA structure should be included. RNA is structured due to every nucleotide being able to pair with others, alongside a flexible RNA backbone and multiple possible tertiary interactions (Vicens, Kieft and Rissland, 2018). RNA structures could support 5′–3′ communication without closed loop interactions (Vicens, Kieft and Rissland, 2018). However, protein binding, helicase activity and ribosome transit means RNA conformation is modified in vivo (Lingelbach and Dobberstein, 1988; Takyar, Hickerson and Noller, 2005; Guo and Bartel, 2016). These processes lead to the removal of secondary structure, increasing the distance between the mRNA ends (Lai *et al.*, 2018). No quantitative value has been assigned to the closeness of the ends or duration of this closeness, to allow the closed loop to function (Vicens, Kieft and Rissland, 2018). The closed loop could exist transiently, where the ends come together and separate multiple times while the mRNA is present (Kluge, Götze and Wahle, 2020). RNAs could require different 5′–3′ proximity, and protein binding to RNA could significantly change the conformation at times, meaning the closed loop structure could vary depending on mRNAs and regulator involvement (Vicens, Kieft and Rissland, 2018). Despite considerations, the closed loop structure seems to be the probable structure in translation, increasing 3' UTR importance.

1.6 Translational Regulation

The complexity of translation provides multiple points where regulation can occur. Translation rate is affected by the number of ribosomes present on the mRNA, which can be influenced by the population of active mRNAs, CDS length, initiation rate, and elongation rate. The elongation rate is usually faster than the initiation rate, which is generally considered the rate-limiting step (Palmiter, 1975; Morisaki *et al.*, 2016; Wu *et al.*, 2016; Yan *et al.*, 2016). However, the ribosome need to move on from the mRNA initiation region to allow another ribosome to initiate, meaning when elongation rates are very low, initiation rates can be affected (Hershey, Sonenberg and Mathews, 2019). Reduced elongation rates can be caused by phosphorylation of elongation factors or rare codon usage (Dever, Dinman and Green, 2018; Proud, 2019). Translational control mechanisms usually involve initiation, which when regulated, quickly alters the rate of protein synthesis compared to transcriptional regulation. This is explained by the duration of transcription, RNA processing, and transport to specific cellular regions potentially required for transcriptional upregulation of gene expression (Hershey, Sonenberg and Mathews, 2019). mRNAs can affect translation initiation through their sequences or secondary structures which influence interactions with translational machinery. Several factors, including proteins and small RNAs, bind to specific mRNAs to prevent or promote ribosome recruitment (Breaker, 2018; Duchaine and Fabian, 2019).

Initiation factors are essential to translation initiation, and when these factors undergo post-translational modifications, such as phosphorylation (Merrick and Pavitt, 2018; Wek, 2018; Proud, 2019), their activity can be affected, which influences initiation rates. eIF2 interacts with Met-tRNA$_i$ during translation, however, eIF2 phosphorylation prevents this interaction, allowing regulation of initiation (Merrick and Pavitt, 2018; Wek, 2018). eIF4E supports mRNA selection and mRNA scanning using its RNA helicase activity. eIF4E-binding proteins (4E-BPs) downregulate eIF4E (Merrick and Pavitt, 2018; Proud, 2019; Robichaud *et al.*, 2019) leading to mixed effects on translation due to the varying requirements for eIF4E activity. There are a range of regulatory elements found in UTRs that can regulate the translation of the transcript. uORFs, found in the 5' UTR, are discussed in more detail in section 1.6.3. Section 1.6.5 contains further detail about miRNAs, which are 3' UTR regulatory elements that can regulate transcript translation.

1.6.1 Translational Regulation in Disease and Cell Stress

The importance of translational regulation is highlighted by its implication in growing numbers of diseases, including immunodeficiency (Piccirillo *et al.*, 2014; Lucas *et al.*, 2016), metabolic disorders (Morita *et al.*, 2013), neurological disorders (Buffington, Huang and Costa-Mattioli, 2014), cancers, and viral infections. Several major signalling pathways, including mammalian or mechanistic target of rapamycin (mTOR), mitogen activated protein kinases (MAPKs) and integrated stress response (ISR), regulate translation by converging on the initiation step, allowing response to varied external and internal conditions (Wek, 2018; Proud, 2019). Disorders associated with translation can involve deregulated tRNA synthesis or function, ribosomopathies, deregulation of the ISR pathway, dysfunction of specific regulatory elements, and deregulation of the mTOR pathway (Tahmasebi *et al.*, 2018). Transcriptional and translational targets of the ISR and the mTOR pathways have crucial roles in tRNA, mitochondrial, and ribosomal biogenesis (Iadevaia *et al.*, 2012).

1.6.2 Short Open Reading Frames (sORFs)

The protein coding capacity of current gene annotations and reference genomes is underestimated. There is increasing discovery and characterisation of 'unannotated proteins' (Slavoff *et al.*, 2013; Bazzini *et al.*, 2014; Lu *et al.*, 2019; van Heesch *et al.*, 2019; Chen *et al.*, 2020; Ouspenskaia *et al.*, 2020; Ruiz Cuevas *et al.*, 2021), alongside the use of several omics-based technologies, especially ribosome profiling (RP) (Ingolia, Lareau and Weissman, 2011). Many small translated ORFs have been found in UTRs, through ribosomal and proteomic profiling in multiple species (Slavoff *et al.*, 2013; Bazzini *et al.*, 2014; Stern-Ginossar and Ingolia, 2015; Calviello *et al.*, 2016; Couso and Patraquim, 2017; Makarewich and Olson, 2017; Brunet *et al.*, 2018). Although RP can find potential short ORFs (sORFs), this is not necessarily enough evidence for translation (Aspden *et al.*, 2014; Bazzini *et al.*, 2014; Ji *et al.*, 2015; Calviello *et al.*, 2016; van Heesch *et al.*, 2019; Weaver *et al.*, 2019; Leong *et al.*, 2022). Translated sORFs can encode functional peptides, such as myoregulin, a conserved regulatory peptide in humans and mice, that influences muscle performance (Anderson *et al.*, 2015). However, sORF translation is generally considered to have regulatory function (Barbosa, Peixeiro and Romão, 2013; Couso and Patraquim, 2017). The presence of sORFs in transcripts does not guarantee biological function (Guttman *et al.*, 2013), and sORFs may be randomly generated in genomes (Couso and Patraquim, 2017). Ribosome profiling data from varied species, suggests widespread sORF ribosomal association, with thousands of long noncoding RNAs (lncRNAs) containing sORFs (Ingolia, Lareau and Weissman, 2011; Guttman *et al.*, 2013; Aspden *et al.*, 2014; Bazzini *et al.*, 2014; Couso and Patraquim, 2017). Regulation by sORFs may involve their presence up or downstream of CDSs, with detrimental effects on translation or induction of NMD, leading to strong selection pressure against, and elimination of uORFs and dORFs (Iacono, Mignone and Pesole, 2005; Neafsey and Galagan, 2007; Johnstone, Bazzini and Giraldez, 2016;

Ruiz-Orera and Albà, 2019). There is a huge number of potential sORFs, however, further understanding is needed into which are translated.

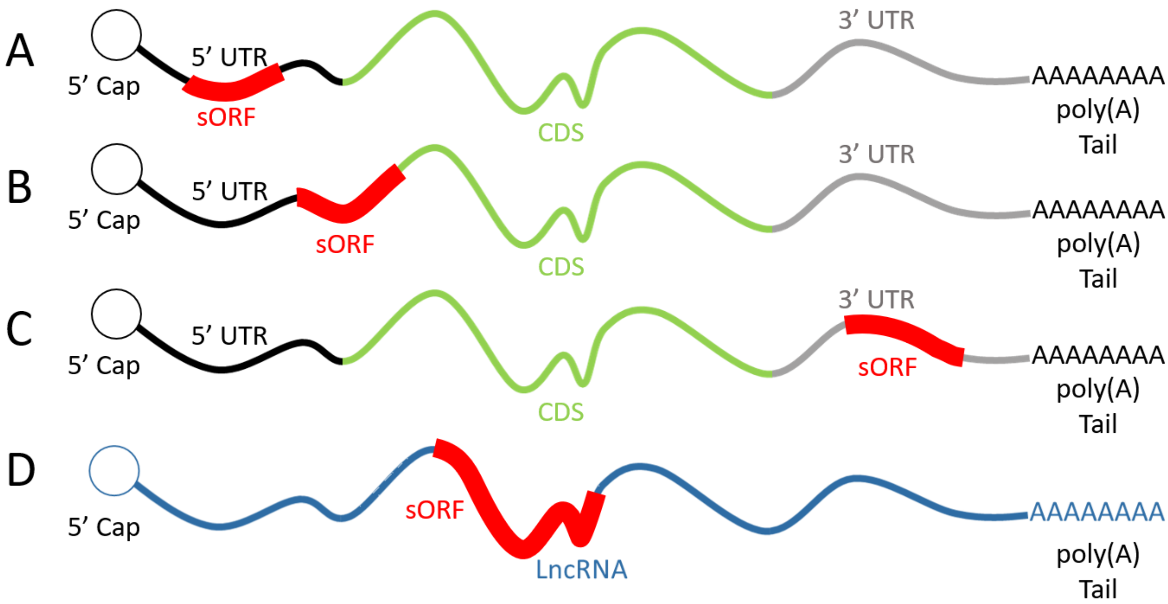

Figure 1.3: Common locations of short open reading frames (sORFs). *A and B – 5' untranslated region (UTR) sORFs, known as upstream open reading frames (uORFs). Found entirely within 5' UTRs (A) or overlapping the coding sequence (CDS) (B). C – 3' UTR sORFs, known as downstream open reading frames (dORFs). D – Long noncoding RNA (lncRNA) sORFs.*

Noncoding ORFs can be: sORFs within 5′ or 3′ UTRs, uORFs or dORFs respectively, ORFs encoded within noncoding RNAs, often lncRNAs (Martinez *et al.*, 2020; Zhang *et al.*, 2021), out of frame overlapping ORFs relative to the canonical CDS, and in-frame overlapping ORFs, producing truncated or extended protein variants (Figure 1.3) (Harding *et al.*, 2000; Jin *et al.*, 2003; Krishna M. Vattem and Wek, 2004; Johnstone, Bazzini and Giraldez, 2016; Chen *et al.*, 2020; Wu *et al.*, 2020b; Leong *et al.*, 2022). Prediction of protein-encoding ORFs often depends on rigorous principles including, monocistronic transcripts (Brunet *et al.*, 2018), AUG start codons (Brůna *et al.*, 2021), no overlapping ORFs (Brůna, Lomsadze and Borodovsky, 2020; Wright, Molloy and Jaschke, 2021) and over 300 nucleotides long (Delcourt *et al.*, 2018). However, these principles are not requirements, such as using a CUG start codon to initiate MYC protein translation (Hann *et al.*, 1988; Kearse and Wilusz, 2017). Some specific sORFs in noncoding RNA transcripts and 5′ UTRs have important functions, including substantial roles in immunology (Niu *et al.*, 2020), cancer (S. Wu *et al.*, 2020), and metabolism (J. Lee *et al.*, 2015; Chugunova *et al.*, 2019).

Aside from sORFs, ORFs can overlap out-of-frame with annotated CDSs, or they can be found inside the annotated CDS. The *RPL36* gene was found to have an overlapping out-of-frame ORF, alt-RPL36, which has the canonical RPL36 nested inside it (Cao *et al.*, 2021). This overlapping ORF produces a larger protein with a different role in PI3K-AKT-mTOR signalling, by initiating translation upstream of *RPL36*, using a GUG start codon (Cao *et al.*, 2021). The G-protein-coupled receptor bradykinin B2 receptor's canonical CDS contains an ORF encoding a 157 amino acid (aa) peptide (Gagnon *et al.*, 2021). This ORF peptide can directly regulate bradykinin B2 receptor activity, and has variable expression profiles in clinically important tissues, such as breast cancer (Gagnon *et al.*, 2021). In-frame overlapping ORFs exist as truncations, extensions, or isoform variants of canonical CDSs, however, identifying overlapping ORFs through ribosome profiling can be difficult (Wright *et al.*, 2022). *FGF2*, a gene controlling cell proliferation, angiogenesis and differentiation, has at least four upstream and in-frame CUG start codons, producing variants that localise to the nucleus, in contrast to the cytoplasmic or secreted canonical protein (Takahashi *et al.*, 2005). *MRPL18* initiates at a downstream and in-frame CUG codon under heat shock conditions, producing a truncated variant, without the N-terminal mitochondrial targeting signal, which is incorporated into cytoplasmic, instead of mitochondrial, ribosomes (Zhang *et al.*, 2015).

1.6.3 Upstream Open Reading Frames (uORFs)

uORFs are well defined translational regulators which are small ORFs found within 5' UTRs, or overlapping CDSs, and uORF translation typically reduces the translational efficiency of the CDS (Mueller and Hinnebusch, 1986; Werner *et al.*, 1987; Krishna M Vattem and Wek, 2004; Brar *et al.*, 2012; Von Arnim, Jia and Vaughn, 2014; Wethmar *et al.*, 2014; Chew, Pauli and Schier, 2016; Johnstone, Bazzini and Giraldez, 2016; Renz, Valdivia Francia and Sendoel, 2020), although in *ATG4* uORFs increase translation in stressful conditions, which is discussed later. RP suggests around half of mammalian genes possess uORFs with translational potential (Lee *et al.*, 2012) and an abundance of non-AUG start codons (Lee *et al.*, 2012; Kearse and Wilusz, 2017), including CUG, GUG, UUG (Ingolia, Lareau and Weissman, 2011; Brar *et al.*, 2012; Arribere and Gilbert, 2013; Sendoel *et al.*, 2017; Na *et al.*, 2018; Renz, Valdivia Francia and Sendoel, 2020). uORF start codon selection is thought to depend on eIFs, meaning eIF mutations and post-translational modifications could influence uORF recognition and translation. This may involve eIFs stoichiometry; in *Saccharomyces cerevisiae,* the levels of eIF1 and eIF1AX influence the selection of conventional AUG or CUG start codons and increased eIF5 and eIF5B levels favour the selection of CUG (Barth-Baus *et al.*, 2013).uORFs can be conserved, often with little sequence similarity, between species (Chew, Pauli and Schier, 2016; Johnstone, Bazzini and Giraldez, 2016; Dumesic *et al.*, 2019).

uORFs function is suggested to occur through engagement of ribosomes away from CDSs. Suggested mechanisms include stalling ribosome scanning and preventing or altering initiation, termination and reinitiation of the ribosome (Somers, Pöyry and Willis, 2013; Couso and Patraquim, 2017; Silva, Fernandes and Romão, 2017). Using the 'leaky scanning' mechanism, the scanning PIC finds a preferable sequence at the uORF and either initiates translation, or scans through and translates the downstream, annotated CDS (Kozak, 2002; Barbosa, Peixeiro and Romão, 2013). More than half of genes with uORFs encode transcript isoforms with and without the uORF, meaning transcriptional regulation, isoform selection, and alternative splicing can influence uORF presence (Pelechano, Wei and Steinmetz, 2013). uORFs can also regulate mRNA decay and transcript half-lives (Jia *et al.*, 2020), and can improve translational output from the downstream CDS (Andreev *et al.*, 2015; Starck *et al.*, 2016). uORFs help to control downstream CDS translation, notably during global translational changes, such as cellular stress (Andreev *et al.*, 2015; Starck *et al.*, 2016; Rodriguez *et al.*, 2019).

Local conditions can influence uORF translation, allowing adaption to these conditions through translational regulation of the CDS (Renz, Valdivia Francia and Sendoel, 2020). However, disease conditions, such as cancer, can also influence uORF regulation (Young and Wek, 2016; Sendoel *et al.*, 2017). Some diseases are influenced by mutations which impair uORF function (Barbosa, Peixeiro and Romão, 2013; Somers, Pöyry and Willis, 2013). Hereditary thrombocytosis, with increased platelets in the peripheral blood and heightened thrombosis risk, can be caused by increased *TPO* translational efficiency due to multiple mutations disrupting a uORF (Cazzola and Skoda, 2000). eIF2A-dependent translation becomes essential in specific cellular contexts, such as the ISR (Starck *et al.*, 2016) or cancer-initiating cells (Sendoel *et al.*, 2017), where eIF2α is phosphorylated, downregulating canonical translation (Renz, Valdivia Francia and Sendoel, 2020). CUG start codons may involve eIF2A binding and delivery of Leu-tRNA to the 40S ribosomal subunit to more efficiently initiate translation (Starck *et al.*, 2012). uORFs are suggested to use CUG start codons, meaning uORFs may have increased translation when canonical eIF2-dependent translation is reduced (Renz, Valdivia Francia and Sendoel, 2020). uORFs raise good questions about whether dORFs could behave in a similar way to uORFs, including the influence of local conditions and cell stress on dORF translation and function.

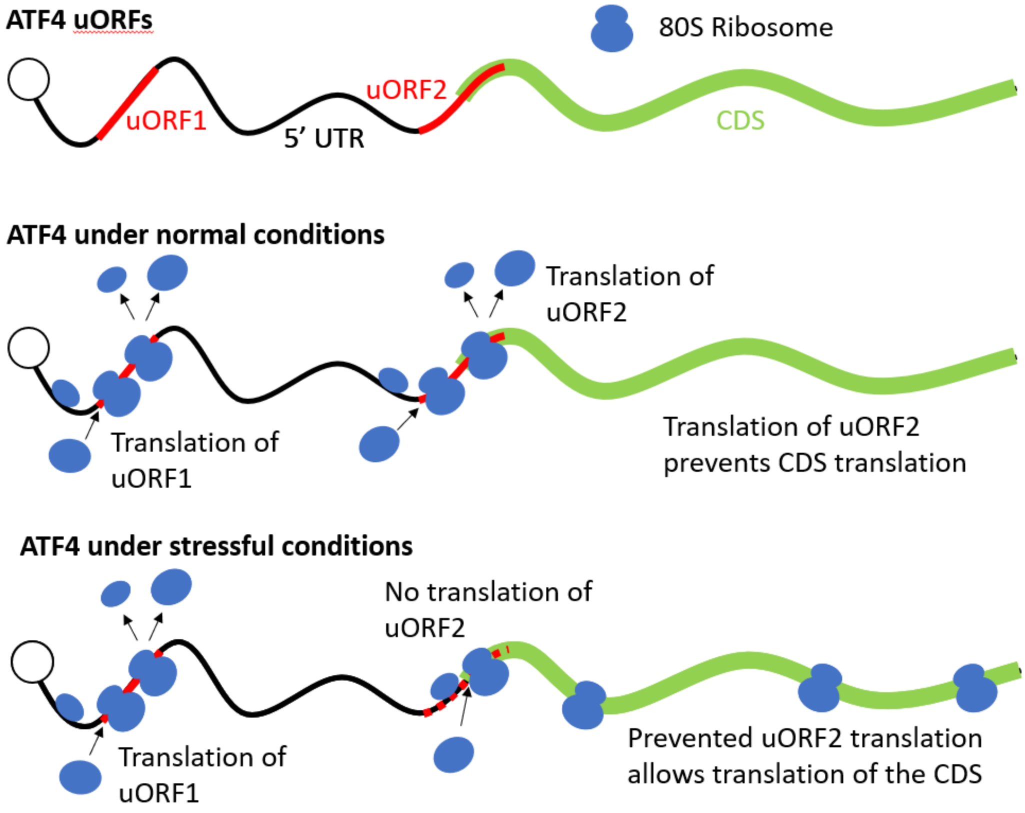
ATF4 uORFs
80S Ribosome
uORF1
5' UTR
uORF2
CDS
ATF4 under normal conditions
Translation of uORF2
Translation of uORF1
Translation of uORF2 prevents CDS translation
ATF4 under stressful conditions
No translation of uORF2
Translation of uORF1
Prevented uORF2 translation allows translation of the CDS

Figure 1.4: ATF4 upstream open reading frames (uORFs) and their function under normal and stressful conditions. *ATF4 mRNA contains two 5' untranslated region (5' UTR) uORFs, uORF1 is entirely within the 5' UTR, and uORF2 begins in the 5' UTR and overlaps the coding sequence (CDS) start codon. Under normal conditions uORF1 and uORF2 are translated, uORF2 translation prevents recognition of the CDS start codon, preventing CDS translation. Under stressful conditions translation initiation at uORF2 can be skipped, allowing translation to initiate at the CDS start codon, translating the CDS under these conditions.*

Several developmental signalling pathways, including Sonic hedgehog SHH, Wnt, Phosphatidylinositol 3-kinase (PI3K), MAPK and Hippo, have components that are translationally regulated by uORFs, such as disruption of the *Ptch1* uORF which can disrupt neurogenesis and reduce hedgehog signalling (Fujii *et al.*, 2017). The *ATF4* gene contains uORFs and has been studied extensively (Krishna M Vattem and Wek, 2004). *ATF4* activates Ihh transcription, which promotes chondrocyte proliferation and differentiation, preventing this activation leads to a dwarfism phenotype in mice (Wang *et al.*, 2009). A range of different stresses can trigger the ISR, resulting in global translational inhibition through eIF2α phosphorylation. Unlike genes without uORFs, *GCN4* and *ATF4* maintain translation under stressful conditions, with the suggested mechanism involving ribosomes bypassing inhibitory uORFs, initiating translation at CDSs in a "leaky scanning" process (Kozak, 2002; Renz, Valdivia Francia and Sendoel, 2020). Mammalian *ATF4* is regulated by two uORFs (Harding *et al.*, 2000), where under normal conditions, uORF1 is efficiently translated and ribosome scanning resumes at uORF2 which overlaps the CDS, repressing *ATF4* expression (Figure 1.4) (Krishna M Vattem and Wek, 2004). Under stressful conditions, eIF2α phosphorylation leads to reduced TC availability which, alongside "leaky scanning", is suggested to bypass uORF2 inhibition, reinitiating translation at the CDS (Figure 1.4) (Krishna M Vattem and Wek, 2004). However, some contradictory RP data has shown increased uORF2 translation upon stress (Renz, Valdivia Francia and Sendoel, 2020). 5' UTR N^6-methyladenine (m^{6}A) modification allows cap-independent alternative translation through direct binding of eIF3 to 5′UTR m^{6}A, which recruits the translation machinery (Meyer *et al.*, 2015). During the heat shock response, 5′ UTRs are methylated and demethylation is halted by the m^{6}A reader, YTHDF2, promoting cap-independent translation (Zhou *et al.*, 2015). m^{6}A was suggested, under stressful conditions, to be involved in *ATF4* uORF translation, *ATF4* uORF2 is methylated, however, m^{6}A on uORF2 is reduced under starvation conditions (Zhou *et al.*, 2018).

Growing numbers of uORFs encode functional peptides (Chen *et al.*, 2020). Although difficult to find, due to short half-lives, uORF-encoded peptides can often regulate CDS translation (Andrews and Rothnagel, 2014; Zhou *et al.*, 2022b). uORF-encoded peptides can act as ligands for the major histocompatibility complex (MHC) class I, leading to T cell responses (Starck *et al.*, 2016). uORF-encoded peptides and the CDS protein, can interact and cooperate, such as in *MIEF1* where a uORF-encoded peptide can bind the CDS protein, regulating mitochondrial fission (Samandi *et al.*, 2017; Chen *et al.*, 2020), highlighting the complexity of uORF regulation.

1.6.4 Short Encoded Peptides (SEPs)

Translation is no longer restricted to annotated protein coding regions, translation also occurs in sORFs within 5' UTRs, lncRNAs and circular RNAs (circRNAs), found through Mass spectrometry (MS)-based proteomics and RP developments (Ingolia, Lareau and Weissman, 2011; Ingolia *et al.*, 2014; Ji *et al.*, 2015; Calviello *et al.*, 2016; Na *et al.*, 2018; Zhou *et al.*, 2022b). sORF translation could also differ from CDS cap-dependent translation (Zhou *et al.*, 2022b). Short encoded peptides (SEPs) can regulate and allow response to varying cellular environments (Merino-Valverde, Greco and Abad, 2020). Some SEPs have important functions in key physiological processes (Couso and Patraquim, 2017; Renz, Valdivia Francia and Sendoel, 2020), from metabolism (C. Lee *et al.*, 2015; Makarewich *et al.*, 2018; Polycarpou-Schwarz *et al.*, 2018; Stein *et al.*, 2018; Chugunova *et al.*, 2019; Zhang *et al.*, 2020), to regulation of gene expression (Guo *et al.*, 2003; D'Lima *et al.*, 2017; Huang, 2017; Huang *et al.*, 2017) and immune responses (Pueyo *et al.*, 2016; Diao *et al.*, 2019; Bhatta *et al.*, 2020; Niu *et al.*, 2020). The apparent prominence of SEPs in some processes may occur through increased study of sORFs in these areas (Schlesinger and Elsässer, 2022). Although there are few reported functional dORF-encoded peptides, these peptides are found in proteomics and RP studies (Slavoff *et al.*, 2013; Vanderperre *et al.*, 2013; Bazzini *et al.*, 2014; Ji *et al.*, 2015; Ma *et al.*, 2016; Schlesinger and Elsässer, 2022).

Even if sORF translation usually provides regulatory function, peptides are generated. Protein coding sORFs were found in three types of moss (*Physcomitrella patens*) cells (Couso and Patraquim, 2017). Many sORFs and SEPs have tissue-specific translation patterns (Couso and Patraquim, 2017). In the past, many sORF-derived proteins were overlooked or removed from datasets due to assumptions that short proteins, under 100 aa, were likely to be non-functional or artefacts (Basrai, Hieter and Boeke, 1997; Carninci *et al.*, 2005). The small size of SEPs could allow them to act on complex biological systems, by acting as: ligands for receptors, regulatory subunits of protein complexes, signalling molecules, allosteric regulators of enzymes, a source of antigenic peptides, or critical transmembrane components (Wright *et al.*, 2022). Although now known to be important, difficulties relating to the identification of translation of SEPs remain (Martinez *et al.*, 2019). Ribosome profiling can struggle to annotate sORFs effectively, and there are issues with the sensitivity of proteogenomic approaches (Martinez *et al.*, 2019).

SEPs have important functions in the *P. patens* genome, with SEP knockouts decreasing growth rate compared to wild-type plants (Couso and Patraquim, 2017). uORF-encoded proteins have potential diagnostic function, such as uORF-encoded proteins from the cardiac troponin T gene found in the blood (J. Lee *et al.*, 2015). Although this detection did not assist myocardial infarction diagnoses, it highlighted a possible clinical use of SEPs (Renz, Valdivia Francia and Sendoel, 2020). uORF-encoded proteins could act as tumour-specific antigens which could be targeted with immunotherapy (Laumont *et al.*, 2016; Erhard *et al.*, 2018; Chong *et al.*, 2020). The importance of precise protein levels during human development depends on translational regulation, seen with SEPs and uORFs (Renz, Valdivia Francia and Sendoel, 2020). Improvement in RP and MS-based proteomics has increased the number of experimentally confirmed protein encoding sORFs, leading to attempts to categorize sORFs into databases such as SmProt (Hao *et al.*, 2018) and OpenProt (Brunet *et al.*, 2021). Including noncanonical ORFs with strong evidence of function into reference databases, such as GENCODE and RefSeq, will improve accessibility to researchers.

ORF-encoded proteins are being found frequently in transcripts annotated as lncRNAs. This questions the definition of lncRNAs, transcribed RNAs that are not translated, when lncRNAs are found to be translated through RP and MS (van Heesch *et al.*, 2019; Chen *et al.*, 2020; Ouspenskaia *et al.*, 2020; Prensner *et al.*, 2021). One study used genome-scale CRISPR knockout screens to study 613 potentially translated and functional lncRNA noncanonical ORFs, observing that inhibiting ORF translation led to poor growth with around 30% of the lncRNAs (Chen *et al.*, 2020). Potentially some of these lncRNAs are misannotated messenger transcripts. However, some transcripts appear to have independent noncoding and coding functions (Lee *et al.*, 2021). In mice, the Linc-RAM lncRNA promotes myogenesis through interactions with the *MyoD* transcription factor (Yu *et al.*, 2017), whereas myoregulin, a sORF-encoded protein, can regulate SERCA activity (Anderson *et al.*, 2015). sORFs may be functional, but they are less conserved than the canonical CDS (Ruiz-Orera *et al.*, 2018; Fesenko *et al.*, 2019), which could mean they represent the early de novo evolution of genes, proto-genes (Carvunis *et al.*, 2012; Schlesinger and Elsässer, 2022).

Regulation of many cancer-associated processes have been linked to SEPs (Merino-Valverde, Greco and Abad, 2020). Immunotherapy could be developed to make use of mutated SEPs in future cancer therapies (Garcia-Garijo, Fajardo and Gros, 2019; Merino-Valverde, Greco and Abad, 2020). Investigation of non-canonical ORFs revealed their importance in cancer cell survival (Prensner *et al.*, 2021). Expression of peptides encoded in UTRs and ncRNAs can change in different cancers, with varied regulatory activity on cancer progression (Sendoel *et al.*, 2017; Sriram, Bohlen and Teleman, 2018; Lei *et al.*, 2020; Wu *et al.*, 2020a; Bakhti and Latifi-Navid, 2021; He *et al.*, 2021; Zhou *et al.*, 2021; Zhou *et al.*, 2022b). The tumorigenesis signalling pathways are the most common target for regulation by ncRNA-encoded peptides (Zhou *et al.*, 2022b). The regulation of several cancers can be influenced by peptides encoded by circRNAs (Zhou *et al.*, 2022b). An alternative 370 aa β-catenin isoform is generated from circβ-catenin that can activate the Wnt/β-catenin pathway to promote tumour growth in liver cancer by stabilising β-catenin through antagonization of its degradation and phosphorylation stimulated by GSK3β (Liang *et al.*, 2019; Zhou *et al.*, 2022b). Overexpression of circAKT3 reduces glioblastoma cell tumorigenicity, radiation resistance, and proliferation, by reducing PI3K/AKT signal intensity through competition with phosphorylated PDK1 to prevent phosphorylation of AKT (Xia *et al.*, 2019; Zhou *et al.*, 2022b). Triple negative breast cancer tumour growth can be inhibited by ASRPS, a peptide encoded by *LINC00908*, through reduced angiogenesis and *VEGF* expression (Wang *et al.*, 2020; Zhou *et al.*, 2022b). STAT3 phosphorylation is prevented by ASRPS binding leading to the previously described effects, meaning ASRPS is downregulated in triple negative breast cancer (Wang *et al.*, 2020; Zhou *et al.*, 2022b). In colorectal cancer, proliferation can be increased by interactions between ASAP, a peptide encoded by *LINC00467*, and ATP synthase α and γ subunits, promoting mitochondrial oxygen consumption and ATP synthase activity (Ge *et al.*, 2021; Zhou *et al.*, 2022b). The function of uORFs and SEPs highlight the potential regulation that translated dORFs could be involved in.

1.6.5 Translational Regulators Associated with 3’ UTRs

If dORFs are translated in 3’ UTRs this could affect other 3’ UTR translational regulators. When considering 3’ UTR translational regulators it is important to consider miRNAs. miRNAs are small noncoding RNAs, around 22 nucleotides in length, that have an important role in regulating mRNA (Bartel, 2018). miRNAs are transcribed from DNA into primary miRNAs, before being processed into precursor miRNAs and mature miRNAs (O’Brien *et al.*, 2018). miRNAs cause degradation and repression of mRNAs through interactions with 3′ UTRs (Figure 1.5) (Ha and Kim, 2014), and although there is some suggestion and debate around miRNAs binding to other regions, this has not been established. Many biological functions are dependent on miRNAs (Fu *et al.*, 2013), however, they are also implicated in diseases when miRNA expression is affected (Tüfekci *et al.*, 2014; Paul *et al.*, 2018; Bergman, Diament and Tuller, 2021). The expanding understanding of miRNAs shows their involvement in cell-cell communication as signalling molecules when secreted extracellularly (Hayes, Peruzzi and Lawler, 2014; Wang, Chen and Sen, 2016; Huang, 2017).

miRNAs cause mRNA deadenylation, leading to decapping and repression of mRNA translation through binding to 3’ UTRs at specific target sequences (Huntzinger and Izaurralde, 2011; Ipsaro and Joshua-Tor, 2015). Almost all miRNAs bind to 3’UTRs with incomplete complementarity at miRNA response elements (MREs), where there are mismatched pairings often in the centre (Jonas and Izaurralde, 2015). Once bound, miRNAs recruit the miRNA-induced silencing complex (miRISC), which can be as simple as an Argonaute protein, which are involved in RNA silencing, combined with the miRNA guide strand (Kawamata and Tomari, 2010). Then, following the MRE and miRNA interaction, miRISC recruits the GW182 family of proteins, allowing other effector proteins, including the poly(A)-deadenylase complexes, PAN2-PAN3, and CCR4-NOT, to be recruited (Behm-Ansmant *et al.*, 2006; Christie *et al.*, 2013; Jonas and Izaurralde, 2015), leading to mRNA repression and degradation (Jonas and Izaurralde, 2015). miRNAs binding to 3’ UTRs could be interrupted by dORF translation within that region, preventing miRNA activity.

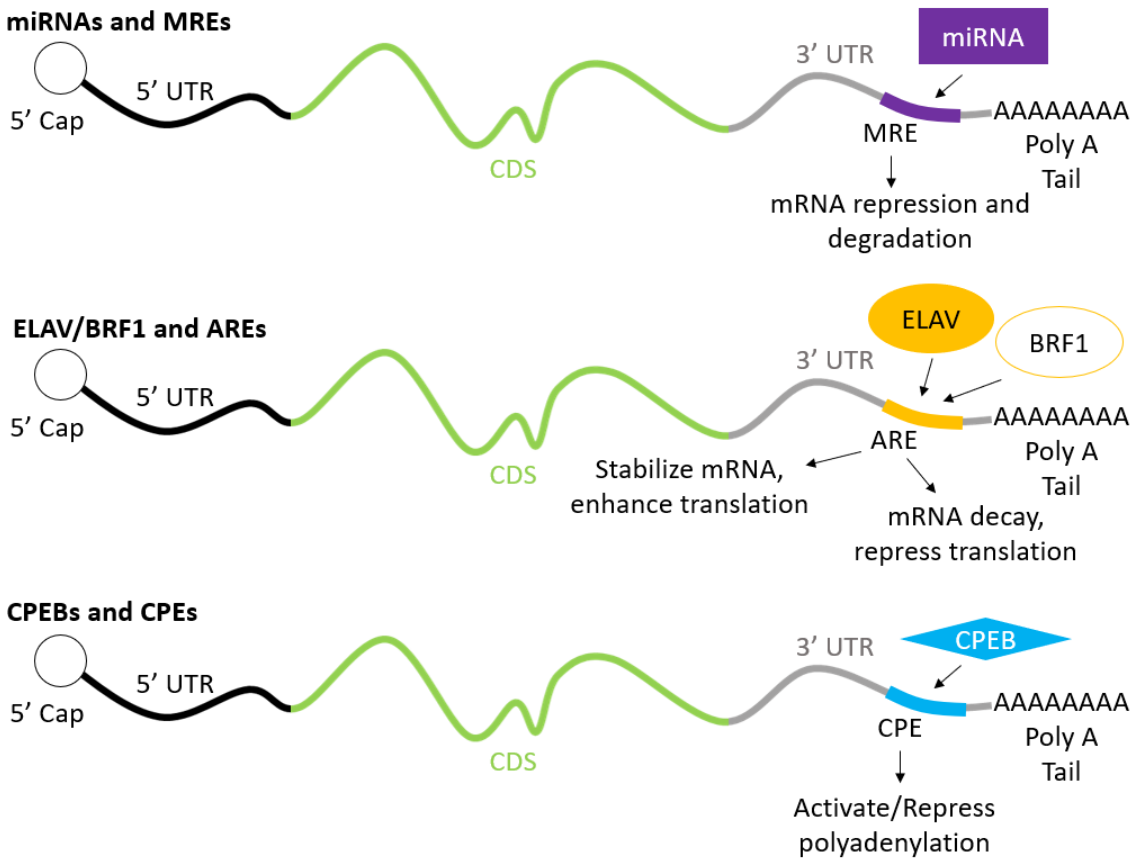

Figure 1.5: Examples of 3' untranslated region (UTR) post-transcriptional regulators. *Micro-RNAs (miRNAs) can bind to 3' UTR miRNA response elements (MREs), leading to mRNA repression and degradation. AU-rich elements (AREs) can be bound by different proteins to cause different regulatory effects. The examples in this figure are the ELAV protein family which bind to AREs to support mRNA stabilization and enhance translation; and the BRF1 protein which binds to an ARE, leading to mRNA decay and repression of translation. Cytoplasmic polyadenylation elements (CPEs) are bound by CPE binding proteins (CPEBs) leading to activation or repression of polyadenylation, allowing translational regulation of specific mRNAs.*

There are other elements in 3' UTRs that can influence translation. AU-rich elements (AREs) can regulate translation when factors bind to 3' UTR AUUUA sequences (Eberhardt *et al.*, 2007). AREs can regulate a variety of genes, such as oncogenes, transcription factors and tumour suppressors (Eberhardt *et al.*, 2007). AREs are important 3' UTR regulatory motifs (Plass, Rasmussen and Krogh, 2017). AREs are commonly seen within transcripts encoding transcriptional and post-transcriptional regulatory factors, allowing formation of auto-regulatory networks (Stoiber *et al.*, 2015). Proteins that bind to ARE elements can enhance or repress translation, ELAV family and NF90 proteins support mRNA stabilization and enhance translation, whereas, TTP, TIAR, TIA1, AUF1, BRF1, KSRP proteins increase mRNA decay and repress translation (Figure 1.5) (Harvey *et al.*, 2018). Another 3' UTR element are

cytoplasmic polyadenylation elements (CPEs) (Fox, Sheets and Wickens, 1989; Hake and Richter, 1994). CPE binding proteins (CPEBs) bind to these uridine-rich CPE sequences and regulate specific mRNA translation by activating and repressing polyadenylation (Figure 1.5) (Fox, Sheets and Wickens, 1989; Hake and Richter, 1994; Ivshina, Lasko and Richter, 2014). In a similar way to
miRNAs these other regulators require protein binding to 3'UTRs which could be disrupted by dORF translation.

1.6.6 Altered 3' UTR Processing in Cancer

Translational regulation is associated with cancer, and this can relate to changes in 3' UTR processing. mRNA transcripts with truncated 3' UTRs, often resulting from alternative polyadenylation (APA), are regularly found in cancer cells which can prevent regulation from 3' UTRs (Mayr and Bartel, 2009). Truncated 3' UTR transcripts can have increased stability and produce increased amounts of proteins, potentially influenced by reduced miRNA repression (Mayr and Bartel, 2009; Singh *et al.*, 2009). During embryonic development (Ji *et al.*, 2009), T cell activation (Sandberg *et al.*, 2008) or neuronal activation (Flavell *et al.*, 2008), gene expression can by influenced by APA-induced 3' UTR length changes (Mayr and Bartel, 2009). APA and truncation of 3' UTRs can affect tumour suppressor genes and oncogenes, meaning APA in cancer could have both suppressive and promotional effects, depending on the genes involved (Mayr and Bartel, 2009).

Stability and translational regulators, including AREs and iron response elements (IREs), are the most frequently observed cis 3' UTR elements (López De Silanes, Paz Quesada and Esteller, 2007). In histone mRNAs, the 3' UTR contains stem-loop determinants which are regulated by the cell cycle (Guhaniyogi and Brewer, 2001; López De Silanes, Paz Quesada and Esteller, 2007). AREs are one of the best characterised specific 3' UTR sequences found in many mRNAs, including those associated with interleukins, cyclins, cell-cycle regulators, cytokines, TNF-α, and oncogenes (López De Silanes, Paz Quesada and Esteller, 2007). Enhanced translation or stabilisation of these mRNAs through 3' UTR length changes can allow overexpression in cancer (López De Silanes, Paz Quesada and Esteller, 2007). IREs and 3' UTR stem loops can act as targets for RNA binding proteins which can regulate transcript translation or stability (López De Silanes, Paz Quesada and

Esteller, 2007). Altered miRNA expression, compared to healthy tissues, is seen regularly in cancer (Lu *et al.*, 2005; López De Silanes, Paz Quesada and Esteller, 2007). 3' UTR changes can affect the regulation of gene expression controlled by this region, by removing regulatory 3' UTR elements, which can promote cancer (López De Silanes, Paz Quesada and Esteller, 2007).

Altered 3' UTR processing can allow differential identification between similar tumour subtypes (Singh *et al.*, 2009). Differential degradation and APA can lead to altered abundancy of different transcript isoforms in a B-cell leukaemia/lymphoma mouse model (Singh *et al.*, 2009). The usual affect is 3' UTR truncation, however, transcripts can also be extended (Singh *et al.*, 2009). Cancer types associated with morphology and cell-cell adhesion tend to have overrepresentation of extended transcripts (Singh *et al.*, 2009). In some cancer cell lines and proliferating cells, there is a preference for comparatively shorter 3' UTRs (Sandberg *et al.*, 2008; Mayr and Bartel, 2009; Singh *et al.*, 2009). Decreased stability, with or without reduced translation, is associated with increased 3' UTR length (Kuersten and Goodwin, 2003; Sandberg *et al.*, 2008; Mayr and Bartel, 2009; Singh *et al.*, 2009). During tumorigenesis, there can be widespread changes in 3' UTR processing (Singh *et al.*, 2009). The terminal portion of 3' UTRs can be affected by APA site selection, leading to inclusion or exclusion of sequences that can affect mRNA stability (Carninci *et al.*, 2006; Singh *et al.*, 2009). mRNA processing was altered in just over 800 genes in mouse models of B-cell leukaemia/lymphoma (Singh *et al.*, 2009).

In addition to the global implications of altered 3' UTR processing in cancer, there are many examples of 3' UTR changes in single genes leading to disease. PD-1 (also known as PDCD1) and its ligand (PD-L1; also known as CD274) help mutated cells to avoid the immune system during the development of some cancers (Kataoka *et al.*, 2016; Kumar and Sharawat, 2018). 3' UTR truncation of *PD-L1* transcripts stabilises these transcripts and increases their abundance, which may support immune evasion and correlates with poor prognosis (Kataoka *et al.*, 2016; Kumar and Sharawat, 2018). 3' UTR truncation may allow PD-L1 to avoid the suppressive effects of miRNA binding (Kumar and Sharawat, 2018). Overexpression of *CSF1*, driven by *CSF1* rearrangement, is common in Tenosynovial giant cell tumours (Ho *et al.*, 2020). These rearrangements usually result in 3' UTR truncation, which could increase *CSF1* expression by reducing the effect of suppressive 3' UTR regulators (Ho *et al.*, 2020). In ovarian cancer, overexpression of HER-2/neu correlates with poor prognosis (Doherty *et al.*, 1999). An alternative HER-2 transcript was found in a SKOV-3 ovarian carcinoma cell line, a model of HER2-driven ovarian cancer (Doherty *et al.*, 1999). APA, instead of gene rearrangement is suggested to generate the alternative *HER-2* transcript (Doherty *et al.*, 1999). This alternative transcript is suggested to be more stable, enhancing *HER-2* expression (Doherty *et al.*, 1999).

1.7 Translation Within the 3' UTR

Translation terminates at the CDS stop codon before the 3' UTR. However, there is general agreement that there are ribosome profiling footprints (RPFs) present within 3' UTRs in different species, including humans (Ingolia *et al.*, 2009; Guydosh and Green, 2014; Ji *et al.*, 2015; Miettinen and Björklund, 2015; Young *et al.*, 2015; Hsu *et al.*, 2016). These 3' UTRs have significantly lower ribosome density compared to CDSs or even 5' UTRs (Ingolia *et al.*, 2009). Although the RPF and ribosomal presence within 3' UTRs is agreed, there is discrepancy when it comes to why the ribosome is there and its function. There is suggestion that 3' UTR associated ribosomes are not translating when they pass along 3' UTRs (Guydosh and Green, 2014; Miettinen and Björklund, 2015). This may be supported by potential changes in 3' UTR ribosome conformation, or composition, and no increased abundance around start or stop codons, which may be seen if translation reinitiation or readthrough was occurring (Miettinen and Björklund, 2015). Whereas there are other reports that 3' UTR translation occurs, evidenced by the distribution of 80S RPFs (Guydosh and Green, 2014; Ji *et al.*, 2015; Young *et al.*, 2015; Hsu *et al.*, 2016). This translation may result from issues with ribosome recycling following stop codon recognition (Young *et al.*, 2015). Another aspect of 3' UTR ribosomal presence is related to the loop structure mRNAs form during translation, through interactions between polyA binding protein and initiation factors (Tarun and Sachs, 1996; Dever and Green, 2012; Hinnebusch and Lorsch, 2012). Translational regulation may be the reason for the observed extensive interactions between 3' UTRs and ribosomes (Miettinen and Björklund, 2015). 3' UTR ribosomes may be recycled back to the 5' end by passing along 3’ UTRs (Brogna and Wen, 2009), transferring important translational machinery components from the 3’ to the 5’ end of mRNA, allowing increased translational efficiency (Hinnebusch and Lorsch, 2012; Miettinen and Björklund, 2015).

Translating 3’ UTR ribosomes may depend on reinitiation within 3’ UTRs (Guydosh and Green, 2014). 3’ UTR reinitiated ribosomes have potentially been found, their presence in any reading frame and detection of 3’ UTR translation products suggests that reinitiation can occur reasonably close to CDS stop codons (Young *et al.*, 2015). A hypothesis for 3’ UTR ribosomal presence describes prevention of usual termination and recycling processes, inhibiting full subunit dissociation, possibly explained by the improper functioning or reduced levels of necessary cofactors, such as recycling factor Rli1 (Guydosh and Green, 2014). This explanation would mean ribosomes may be present on mRNA, following protein release, but in a state where translation is not resumed, instead scanning may occur in 3’ UTRs (Guydosh and Green, 2014). Almost all genes demonstrate 80S ribosomal presence at stop codons and within 3’ UTRs in a Rli1-depleted yeast strain (Young *et al.*, 2015). *Rli1* overexpression can reduce the 3′UTR ribosomes found, suggesting that prevention of terminating ribosome recycling at CDS stop codons is the likely cause of 3’ UTR ribosomal presence. Together this builds evidence that Rli1 has a role in regulating 3′ UTR translation (Young *et al.*, 2015). In yeast, 3’ UTR ribosomes appear to cluster nearer to CDS stop codons than the end of 3’ UTRs (Guydosh and Green, 2014), supporting the concept that these may be post-termination ribosomes (Miettinen and Björklund, 2015). Some specific mRNAs have demonstrated an interaction between ribosomes present in CDSs and 3’ UTRs (Eldad, Yosefzon and Arava, 2008). This could mean these 3’ UTR ribosomes interfere with CDS ribosomes, potentially having some role in elongation and mRNA folding (Miettinen and Björklund, 2015). Hypothetically, the presence of 3’ UTR ribosomes could lead to competition at binding sites or removal of factors bound to mRNA (Qu *et al.*, 2011).

Translational changes, involving decreased initiation of translation, are seen in yeast cells during acute amino acid starvation (Hinnebusch, 1988, 2005). During amino acid starvation in yeast, there was greater ribosome occupancy of 3’ UTRs, although still lower than other regions (Ingolia *et al.*, 2009). RP is widely used in determining ribosomal interactions, but there are considerations. Ribosomal footprints do not guarantee translation, as non-productive binding of single ribosomes to mRNAs and scanning 40S ribosomal subunits can generate footprints (Wilson and Masel, 2011). The short nature of sORFs means that there is a small target for ribosomal binding to produce footprints, meaning it is often more difficult to use to study sORFs (Aspden *et al.*, 2014). The impact of stressful conditions and evidence of 3' UTR translation leaves the potential for translational regulation.

Reinitiation following uORF translation (Kozak, 1984, 1987) is acknowledged as an important post-transcriptional regulator of gene expression, and although previously identified in 5' UTRs rather than 3' UTRs, this reinitiation could help to understand dORF translation. Depending on the progression of ribosomal recycling, various mechanisms can allow reinitiation of translation to take place. Efficient reinitiation often only occurs following translation of sORFs, where the level of reinitiation drops with increasing uORF length (Luukkonen, Tan and Schwartz, 1995; Kozak, 2001). However, reinitiation efficiency is controlled by duration of uORF translation instead of uORF length, leading to the hypothesis that reinitiation requires ribosomal retention of a critical factor throughout elongation and termination and only ribosomes that retain the factor can reinitiate (Kozak, 2001). This may indicate an issue for reinitiation for dORFs as this would follow translation of CDSs not sORFs, possibly reducing reinitiation efficiency. The mechanism behind reinitiation is debated and is not fully understood. There are suggestions that reinitiation relies on initial involvement of eIF4F, probably retained on ribosomes through interaction with eIF3 (Pöyry, Kaminski and Jackson, 2004). There is also suggestion that a variety of initiation factors (eIF1, eIF1A, eIF2, eIF3) are also needed to allow reinitiation (Hinnebusch, 2005; Skabkin *et al.*, 2013). Another suggestion is that reinitiation can also be promoted by eIF2D and MCT1/DENR downstream from uORFs (Skabkin *et al.*, 2013; Schleich *et al.*, 2014, 2017).

Several studies suggest post-termination ribosomes often remain weakly attached to mRNA when ABCE1 mediated splitting of ribosomes does not occur (Skabkin *et al.*, 2013; Young *et al.*, 2015; Zinoviev, Hellen and Pestova, 2015; Mills *et al.*, 2016), allowing ribosomes to slide upstream and downstream from the stop codon, reinitiating at P-site cognate codons. It is hypothesised that dissociation of eRF1 from post-termination complexes means P-site tRNA can enter a P/E hybrid state, disrupting P-site codon–anticodon base pairing, allowing bidirectional movement of the ribosome affected by local mRNA structure (Skabkin *et al.*, 2013). Pseudo-translocation may follow binding of a cognate aa-tRNA to the A-site sense codon, leading to resumption of translation without a start codon (Schwab *et al.*, 2003). Functional gene products are not thought to be produced by this mechanism, probably due to start codon selection being semi-random, but translation of rare peptides from 3′ UTR ORFs lacking AUG codons could use this mechanism (Schwab *et al.*, 2003). Many of these mechanisms are poorly accepted or understood but are useful in indicating ways dORF may be translated.

1.7.1 Stop Codon Readthrough

To translate in 3' UTRs, ribosomes need to either readthrough the stop codon or reinitiate in 3' UTRs. Stop codon readthrough is the continuation of translation into the 3' UTR, instead of terminating at a stop codon, extending the protein (Figure 1.6) (Doronina and Brown, 2006; Namy and Rousset, 2010). Readthrough could allow translation of C-terminally extended proteins or dORFs (Hellen, 2018). Readthrough occurs infrequently when a tRNA with an anticodon similar to the anticodon for the stop codon, prevents stop codon recognition by eRF1, allowing translation to continue into 3' UTRs (Brenner, Stretton and Kaplan, 1965; Dabrowski, Bukowy-Bieryllo and Zietkiewicz, 2015; Hellen, 2018; Rodnina *et al.*, 2020). 3'UTR length and the codon present in the ribosomal P site are thought to influence readthrough efficiency (Mangkalaphiban *et al.*, 2021) ATA, ACA, ACC, CTG, and GAC codons in the P site and shorter 3' UTRs are associated with increased readthrough efficiency (Mangkalaphiban *et al.*, 2021). Readthrough can be an important process through: reduction of the negative effects of premature stop codons (Kopczynski, Raff and Bonner, 1992; Fearon *et al.*, 1994), opposing NMD (Keeling *et al.*, 2004), alteration of protein C-terminus, allowing changes in protein activity (Torabi and Kruglyak, 2012), stability (Namy, Duchateau-Nguyen and Rousset, 2002), and localization (Freitag, Ast and Bölker, 2012). Readthrough occurs in humans and can regulate gene expression and function (Dunn *et al.*, 2013). VEGF-A has an isoform produced by readthrough with the opposite function to the canonical protein, with decreased expression in colon cancer cells, allowing greater tumour progression (Eswarappa *et al.*, 2014; Eswarappa and Fox, 2015). There is evidence within insects and crustaceans that readthrough occurs in hundreds of genes (Jungreis *et al.*, 2016). In *Anopheles gambiae* and *Drosophila melanogaster,* 5-6% of stop codons are readthrough, meaning many biological pathways are impacted (Jungreis *et al.*, 2016).

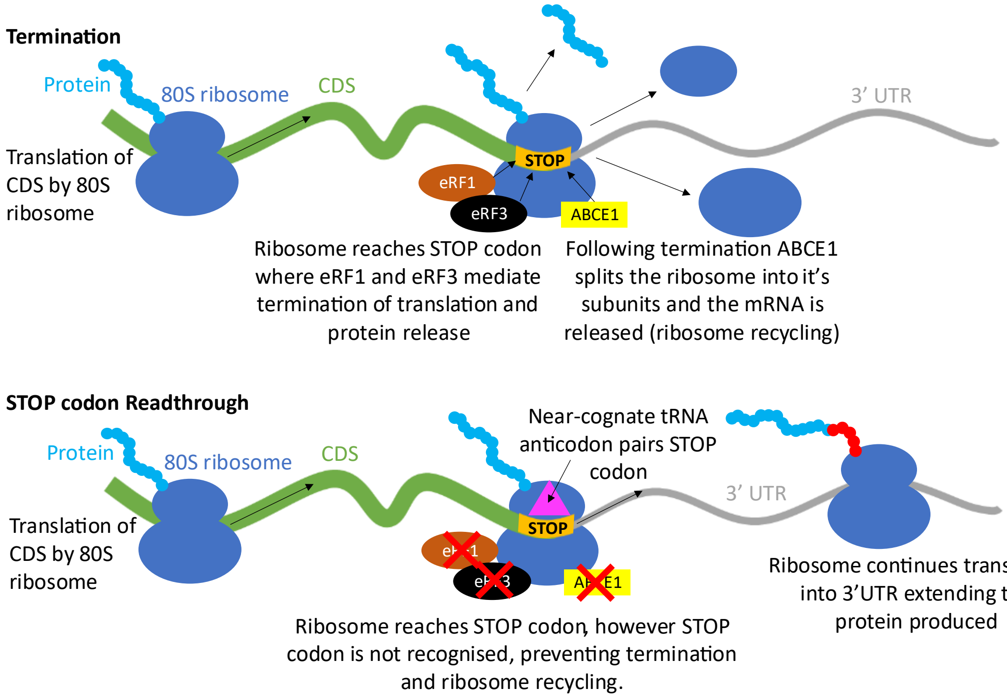
Termination
Protein
80S ribosome
CDS
3' UTR
Translation of CDS by 80S ribosome
STOP
eRF1
eRF3
ABCE1
Ribosome reaches STOP codon where eRF1 and eRF3 mediate termination of translation and protein release
Following termination ABCE1 splits the ribosome into it's subunits and the mRNA is released (ribosome recycling)
STOP codon Readthrough
Protein
80S ribosome
CDS
Near-cognate tRNA anticodon pairs STOP codon
3' UTR
Translation of CDS by 80S ribosome
STOP
eRF1
eRF3
ABCE1
Ribosome reaches STOP codon, however STOP codon is not recognised, preventing termination and ribosome recycling.
Ribosome continues translation into 3'UTR extending the protein produced

Figure 1.6: Representations of termination and stop codon readthrough. *The 80S ribosome translates the coding sequence (CDS) until it reaches a stop codon. Eukaryotic release factor 1 (eRF1) and eRF3 help recognise the stop codon to mediate termination of translation and cause protein release. After termination ABCE1 and eRF1 cause ribosomal recycling, splitting the ribosomal subunits so they can commence translation again, and the mRNA is released. Stop codon readthrough is suggested to occur when a near-cognate tRNA pairs with the stop codon, preventing recognition of the stop codon, meaning termination and ribosome recycling do not occur. The ribosome continues to translate into the 3' UTR extending the protein (Brenner, Stretton and Kaplan, 1965; Doronina and Brown, 2006; Namy and Rousset, 2010; Dabrowski, Bukowy-Bieryllo and Zietkiewicz, 2015; Hellen, 2018; Rodnina et al., 2020).*

Translation termination can be regulated by many factors which can also combine with processes such as NMD (Karousis and Mühlemann, 2019). Regulation of termination can impede translation of subsequent ORFs and modulate readthrough (Beier and Grimm, 2001). Compared to regular stop codons, termination at premature stop codons is suggested to be slower and more inefficient leading to mRNAs undergoing NMD (He and Jacobson, 2015). Reducing termination enhancer activity and increasing negative regulators can prevent termination at premature stop codons. In humans, PABP supports termination through binding to eRF3 (Hoshino *et al.*, 1999; Ivanov *et al.*, 2008, 2016). Prevention of PABP's stimulatory influence, involving increased distance between the PABP-bound 3′-poly(A) tail and premature stop codons, compared to regular stop codons, may impair termination (Hellen, 2018). ABCE1's promotion of termination does not rely on prior dissociation of eRF3 from eRF1, unlike it's recycling activity, and is not caused by stabilisation of eRF1 binding to ribosomes (Shoemaker and Green, 2011). The basis for this activity remains unknown. In eukaryotes, readthrough has been associated with downstream structural elements, such as the *D.melanogaster hdc* stop codon, where a hairpin can function in heterologous mRNAs (Steneberg and Samakovlis, 2001; Jungreis *et al.*, 2011). Readthrough can also be promoted by a downstream GUAC motif (Eswarappa *et al.*, 2014; Loughran *et al.*, 2014).

1.8 Downstream Open Reading Frames (dORFs)

Translation within UTRs has not been explored to the same extent as CDSs. The understanding of uORFs, and now dORFs, undermines the term UTR, with evidence of translation of these regions (Renz, Valdivia Francia and Sendoel, 2020). The field of sORFs and SEPs will likely expand, with suggestion that there has been an underestimation of the number of sORFs conserved in the plant kingdom (Couso and Patraquim, 2017). There are considerably fewer dORFs compared to uORFs, possibly explained by translation termination reducing 3' UTR ribosomal presence (Wright *et al.*, 2022). However, there remain hundreds of ribosome profiling observations of translational activity in 3′ UTR sORFs. Stop codon readthrough could explain this, through an extension of the canonical CDS, generating 3' UTR ribosome footprints (Wu *et al.*, 2020b). Ribosome termination at the canonical CDS followed by reinitiation at dORFs is another explanation, though this would require a novel mechanism. dORFs have been revealed through RP and proteomics, which found small translated 3' UTR ORFs (Bazzini *et al.*, 2014; Ji *et al.*, 2015; Mackowiak *et al.*, 2015; Chen *et al.*, 2020). dORFs have not been systematically characterised and their functions remain unknown, however, recent evidence has suggested a novel dORF regulatory mechanism, whereby dORF translation enhances the translation of the main CDS (Wu *et al.*, 2020b). This regulatory function of dORFs could be an important consideration in the translational regulation field.

Especially when many mRNAs contain dORFs, including in orthologues, which could imply an evolutionary pressure to maintain dORFs (Wu *et al.*, 2020b). dORF regulation could be comparable to uORFs, microRNAs, and m6A-mediated regulation (Meyer *et al.*, 2015; Chew, Pauli and Schier, 2016; Johnstone, Bazzini and Giraldez, 2016), with the function of dORFs appearing to be consistent with both AUG and non-AUG start codons (Wu *et al.*, 2020b). As discussed, uORFs are influenced by stressful conditions, however, dORFs have been suggested to have consistent translation and function under different conditions (Wu *et al.*, 2020b). Although more research is needed to establish whether dORFs may be regulated in a cell-type or condition-dependent manner (Wu *et al.*, 2020b).

Although dORFs may be translated, reinitiation of translation following CDS stop codons appears to be uncommon, with translation efficiency of dORFs shown to be far lower (30-fold on average) than CDSs (Ji *et al.*, 2015). In mice, some dORF sequences are more conserved than the untranslated sequences in close proximity to the dORF (Wu *et al.*, 2020b), with 32% of human dORF peptides showing conservation in mice (Ji *et al.*, 2015). Most peptides produced by UTR translation are not suggested to be functional (Slavoff *et al.*, 2013; Ji *et al.*, 2015). Translated ORFs have been discovered in 3' UTRs of plant species (Hsu *et al.*, 2016), and many previously ignored translated ORFs are involved in regulatory mechanisms or provide bioactive proteins (Orr *et al.*, 2021). It remains unknown if dORFs often encode functional peptides. Some studies, using MS, have detected dORF products (Bazzini *et al.*, 2014; Chong *et al.*, 2020; Ruiz Cuevas *et al.*, 2021), but they are limited and absent of rigorous scrutiny and any characterization of dORF microproteins. One of the few functional dORF microproteins, the dORF in chemotherapeutic drug-resistance gene *ABCB5*, was shown to be expressed and have function with immunogenic activity in melanoma-derived samples (Chong *et al.*, 2020).

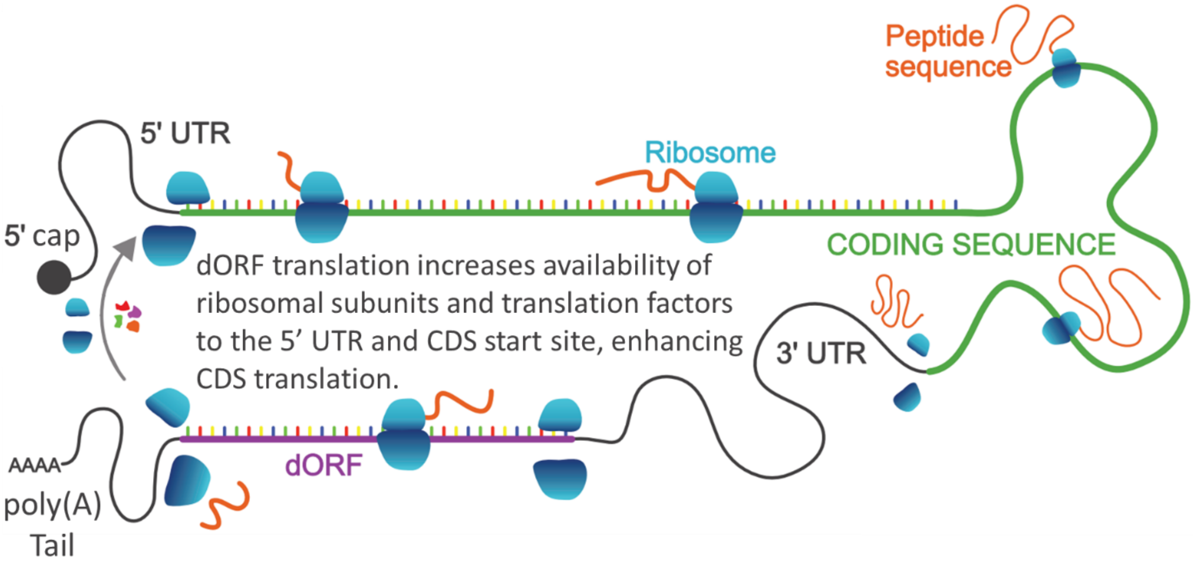
Peptide sequence
5' UTR
Ribosome
5' cap
CODING SEQUENCE
dORF translation increases availability of ribosomal subunits and translation factors to the 5' UTR and CDS start site, enhancing CDS translation.
3' UTR
AAAA
poly(A) Tail
dORF

Figure 1.7: Hypothesised explanation for downstream open reading frame (dORF) function to enhance coding sequence (CDS) translation. *A closed structure of mRNA, with the 5' and 3' ends in close proximity, means that translation of dORFs increases the availability of translation factors and ribosomal subunits to the 5' untranslated region (UTR) and CDS start site, leading to increased CDS translation.*

The mechanism behind dORF regulation remains unclear, it is hypothesised that dORF translation in 3' UTRs makes use of the close proximity of the 3' and 5' UTR to pass translation factors and/or ribosomal subunits from dORFs to 5' UTRs or CDS start sites to increase CDS translation (Figure 1.7)(Wu *et al.*, 2020b). The mechanism could be similar to viral cap independent 3' UTR translation enhancers, which attract either ribosomes, initiation factors or both (Nicholson and White, 2011; Simon and Miller, 2013; Wu *et al.*, 2020b). Induced recruitment of eIF4G to 3' UTRs (Paek *et al.*, 2015) and methyltransferase *METTL3* m6A modification of 3' UTRs can promote oncogenesis by physically interacting with eIF3h to circularise mRNA, enhancing CDS translation (Choe *et al.*, 2018). This circularisation and the looped model, with 5' and 3' UTR ends in close proximity, may mean that attracting translation factors to dORFs supports CDS translation (Figure 1.7) (Choe *et al.*, 2018). The hypothesised dORF function suggests it is dORF translation that enhances CDS translation not an encoded peptide (Wu *et al.*, 2020b), however, two peptides encoded by dORFs have been associated with cell proliferation (Chen *et al.*, 2020). Further support of the translation factor recruitment mechanism comes from the evidence that the number of dORFs, rather than the dORF length, correlates with the regulatory effect (Wu *et al.*, 2020b). Ribosomal recruitment, rather than ribosome readthrough from the CDS, is suggested to be part of the dORF translation mechanism, and 3’ UTR sequences upstream of dORFs may be involved (Wu *et al.*, 2020b).

Upstream of start codons, internal ribosome entry sites (IRESs) can also initiate translation without need for the 5’ cap with the help of initiation factors (Jackson, 2013; Nobuta *et al.*, 2020). Some human proteins require IRES initiated translation for their production (Nobuta *et al.*, 2020). The mechanism and structure of IRESs are varied with differing needs for IRES trans-acting factors and initiation factors (Nobuta *et al.*, 2020). It is worth noting that IRESs have been found in 3’ UTRs (Weingarten-Gabbay *et al.*, 2016; Nobuta *et al.*, 2020). There may be global avoidance of 3’ UTR translation due to the often negative effects of proteins with extended C-termini (Nobuta *et al.*, 2020). In yeast, suppression of quality control factors allows reinitiation of translation in 3’ UTRs (Guydosh and Green, 2014; Nobuta *et al.*, 2020), where dORF translation in four 3’ UTRs occurred, three of which used non-AUG start codons (Nobuta *et al.*, 2020). A mammalian translation construct determined that, with unstructured regions surrounding the stop codon, reinitiation could happen at AUGs up or downstream of the stop codon with eIF2, eIF3, eIF1 eIF1A, and Met-tRNAi, preventing the 40S subunit from detaching from mRNA, allowing scanning in either direction (Skabkin *et al.*, 2013; Nobuta *et al.*, 2020). IRES activity was found to initiate dORF translation in the *GCH1* 3’ UTR, with support from eIF4G, eIF2, eIF3, eIF4A and eIF4B (Nobuta *et al.*, 2020). This IRES may be targeted by eIF4G which recruits the other factors in complex with the 40S subunit as the 43S preinitiation complex (Nobuta *et al.*, 2020). In yeast, polyU is the preferred target for eIF4G, however, in mammals the eIF4G binding sequence remains unknown (Zinshteyn, Rojas-Duran and Gilbert, 2017; Nobuta *et al.*, 2020). One hypothesis describes eIF4G and eIF4A/B rearranging the 40S subunit to allow reattachment to mRNA (Nobuta *et al.*, 2020). There remains much to be understood about dORFs, their function, and the mechanisms behind their function.

1.9 Aim and Objectives

Translational regulation is important in controlling protein production in healthy states to meet the needs of local conditions, and it is also implicated in an increasing number of disease states, from cancers to immune or neurological disorders. This makes understanding translational regulation essential to understanding human health and disease, and can not only improve understanding but may provide new therapeutic opportunities. The suggested function of dORFs as translational regulators and the little known about these regulators makes them good candidates to explore (Wu *et al.*, 2020b). This study aims to investigate translational regulation through bioinformatic analysis, with a particular focus on dORFs and their role in translational regulation. The first objective is to explore the translation and function, as translational regulators, of 'Wu dORFs' (which are the dORFs identified by Wu *et al.* (2020b)), with investigation into whether this function changes in cancer. The second objective is to look into the Wu dORFs composition, compared to the 3' UTR and CDS, and search for other potential dORFs and assess their abundance and potential translational regulation. The third objective is to investigate dORF conservation across species and whether ribosomal association with dORFs and 3' UTRs is influenced by different cell types, disease states and conditions.

The established function of uORFs in influencing translation of mRNAs could help to understand dORFs, alongside consideration of 3' UTR ribosomal presence and potential translation which may occur. The global impact of cancers and difficulties associated with treatments make this an important starting point when considering dORFs, especially considering the impact that cancer can have on uORFs. This study will explore the translation and function, as translational regulators, of 'Wu dORFs', with investigation into whether this function changes in cancer. The understanding of these Wu dORFs will be expanded by looking into their composition, particularly compared to 3' UTRs and CDSs. This study will then move beyond the Wu dORFs and will search for other dORFs. The abundance and potential translational regulation function of these additional dORFs will then be investigated. If dORFs function as translational regulators, it seems likely that there will be conservation between species. This is supported by conservation of parts of 3' UTR and suggestion that uORFs and Wu dORF presence may be conserved across species.

This study will investigate dORF conservation across species, particularly compared to conservation of 3' UTRs, and will investigate whether ribosomal association with dORFs and 3' UTRs is influenced by different cell types, disease states and conditions.

In this study a bioinformatic approach has been used. Bioinformatics is a broad field combining biological data with computer and data science (Bartlett, Penders and Lewis, 2017; Attwood *et al.*, 2019). The expansion of available biological data, through technological developments has made bioinformatics an indispensable part of biological science (Bartlett, Penders and Lewis, 2017; Attwood *et al.*, 2019). This approach allows publicly available datasets to be reanalysed to explore dORFs, saving time and resources and maximising the value of existing data. This also allows study of dORFs in a much larger range of cell types and conditions than would be possible if all data were to be generated *de novo*. A bioinformatic approach is suited to the conceptual nature of this study; with little known about dORFs, larger analyses looking in a genome-wide context can be done. This made use of a discovery type approach and initial analysis looking into function of dORFs as a group of regulators rather than individual dORFs. Databases and bioinformatic tools are also important in allowing conservation analysis of a large number of dORFs in a wide variety of species. Together a bioinformatic approach allows identification of candidate dORFs with evidence of function for future laboratory-based studies. This approach, using available datasets with bioinformatic tools and scripts makes the analysis included in this study highly reproducible.

Chapter 2: Materials and Methods

This research used bioinformatic analysis to explore translational regulation, with a focus on dORFs and their potential function as one such regulator. Translational regulation was investigated in Wu dORFs initially, which are the dORFs presented by Wu *et al.* (2020b) with potential function as translational regulators. Ribosomal association with, and possible translation of, 3’ UTRs was also considered. Analysis expanded to include a wider range of dORFs. The abundance of dORFs in 3’ UTRs and the composition of Wu dORFs, especially compared to 3’ UTRs and CDSs was investigated. The possibility of dORF conservation across species was also explored, in addition to further investigation into ribosomal association with dORFs and 3’ UTRs in a variety of different local conditions, diseases, cell or tissue types.

All Python scripts, ending .py, are original and are available in a GitHub repository (available at: https://github.com/joetomlinson/Joe_Tomlinson_Thesis_Python_Scripts) . Where it is indicated that the scripts have been run this is with Rocky Linux (version - 8.20) operating system with Python (version – 3.9.13).

2.1 Wu dORFs used in bioinformatic analyses

The dORFs detailed in the original publication, highlighting the novel regulatory function of these dORFs, were used in the analysis, providing 2152 ‘Wu dORFs’ in 1406 genes (Wu *et al.*, 2020b). These Wu dORFs possessed AUG and non-AUG start codons and were suggested to be translated using ribosome profiling read alignments and the method and equation described in Bazzini *et al.* (2014) (Wu *et al.*, 2020b).

2.2 Publicly available datasets used in analyses

All datasets are publicly available from the Sequence Read Archive (SRA) (available at https://www.ncbi.nlm.nih.gov/sra) with the project accessions provided in this section (Leinonen, Sugawara and Shumway, 2011). Paired RNA sequence (RNAseq) and ribosome profiling (RP) datasets generated from the same sample are included in Table 2.1. The datasets were selected because they were publicly available, there was evidence of publication using them, and they were from human tissue, or cells, that allowed comparison between healthy and cancerous states. There was a very limited number of datasets that fitted these criteria, meaning there was little choice over which datasets could be used. The use of kidney tissue was useful as it followed on from the work done by Wu *et al.* (2020b) which involved human embryonic kidney (HEK) cells. The depth of these datasets is included in the data provided in Tables 8.1 and 8.2 in Appendix 1. In project PRJNA256316 there were two RP datasets and one RNAseq dataset available for each sample, the RP datasets were denoted with '.1' and '.2' following the dataset annotation e.g. NT1.1.

Table 2.1: Sequence Read Archive (SRA) projects with Paired RNAseq and ribosome profiling (RP) datasets used in subsequent analysis. *This table includes where the samples are obtained from for the datasets in each project and the nomenclature of the datasets in this research.*

SRA Project ID	Cell/Tissue Type and Treatment/Disease State	Dataset Nomenclature
PRJNA256316	Human tissue samples from healthy kidney	NT
	Human tissue samples from renal cell carcinoma	TT
PRJNA532400	Unmodified A549 pulmonary adenocarcinomic human alveolar basal epithelial cells	A549 Ctrl
	A549 cells with shRNA induced FKBP10 knockdown	A549 KD
PRJNA880902	RKO cells under fed (400 uM arginine) conditions	RKO Fed
	RKO cells under starved (12.5 uM arginine) conditions	RKO starved

In addition to the above RP datasets, additional RP datasets from various cells and tissue types in a range of conditions, shown in Table 2.2, were used for the analysis described in section 2.10. These datasets were selected because they had evidence of publication from the datasets, they were publicly available, and they were taken from human cells, or tissues, from healthy or diseased states with a variety of modifications and treatments to provide a wide range of comparison for ribosomal association with dORFs and 3' UTRs. The depth of these datasets is provided in Supplementary Table 1 available at https://github.com/joetomlinson/Joe_Tomlinson_Thesis_Supplementary_Table.

Table 2.2: Sequence Read Archive (SRA) projects with ribosome profiling datasets used in subsequent analysis. *This table includes details about where the samples were obtained from to generate the RP datasets.*

SRA Project ID	Cell/Tissue Type and Treatment/Disease State
PRJNA768399	Human oral mucosal tissue taken from tumorous and paracancerous tissue around oral cavity carcinomas (OCCs)
PRJNA982716	BJ-Ras-ER cells in either proliferative or senescent state treated with either dimethyl sulfoxide (DMSO) or N1-guanyl-1,7-diaminoheptane (GC7), an eIF5A hypusination inhibitor
PRJNA795419	HeLa cells transfected with either control or U1 AMO
PRJEB43705	iPSC-derived neurons with and without TDP-43 knockdown
PRJNA768478	HeLa cells in either mitotic or interphase states either wildtype cells or with DENR knockout
PRJNA756018	Human primary cells (VSMC, HUVEC, Hepatocytes, HCAEC, HA EC, Fibroblast, ES) and tissues (adipose and brain)
PRJNA674567	Control or DDX3-depleted SAS cells
PRJNA591767	Glioblastoma cell lines, U251 and U343, exposed to ionizing radiation with measurement at three time points, 0, 1 and 24 hours
PRJNA369742	Huh7 cells treated with 1.5 µM PF-06446846, a translational inhibitor, or vehicle.
PRJNA415033	Wildtype and DHX36 knockout HEK293 cells
PRJNA369552	HEK293 cells with either cycloheximide or harringtonine treatment
PRJNA238879	HEK293T cells 30 minutes following treatment with and without arsenite (40 µM)
PRJNA599943	HeLa cells with and without siRNA-mediated ABCE1 knockdown
PRJNA858047	THP-1 cells treated with vehicle control (0.1% ethanol/PBS), lipopolysaccharide (LPS) and LPS with dexamethasone for three hours
PRJNA822939	Human corneal epithelial cells exposed to mild osmotic stress (500 mOsm) for one and six hours with datasets including torin1 (mTOR inhibitor) or MeAIB (SNAT2 inhibitor) added for the final treatment hour of six-hour mild osmotic stress
PRJNA418238	APC deficient or restored SW480 cells with either control or eIF2B5 knockdown
PRJNA406823	HEK293 cells treated with either thapsigargin (THAP) or DMSO

2.3 Galaxy Platform

Galaxy is a web-based platform for bioinformatic analysis accessible at https://usegalaxy.org/ (Afgan *et al.*, 2018). Galaxy provides a range of different bioinformatic tools from one platform which can be easily modified in Galaxy giving good user control over functions. The tools can also be piped together in combinations, reducing repeated manual inputs, creating workflows. Using Galaxy does not require coding languages, making the bioinformatic tools more accessible to more researchers. The web-based nature of Galaxy also provides a large amount of cloud storage for data. It is also worth noting that the analysis described in subsequent sections does not have to be completed through Galaxy as the tools described can be used in isolation as standalone software.

2.4 Translational Regulation Analysis

2.4.1 Dataset preparation and alignment

Galaxy workflows were developed for RNAseq and RP datasets from the PRJNA256316, PRJNA532400, and PRJNA880902 SRA projects. These workflows can be run within Galaxy by modifying the SRR accession provided for each RP (Figure 2.1) or RNAseq (Figure 2.2) dataset. The workflow for RP datasets is shown in Figure 2.1, beginning with the Download and Extract Reads in FASTQ (Galaxy Version 3.1.1+galaxy1) and Extract dataset (Galaxy Version 1.0.1) tools to upload the dataset using default settings and the SRR accession (Leinonen, Sugawara and Shumway, 2011). Then the FastQC tool (Galaxy Version 0.74+galaxy1) was run with default settings on the dataset as an initial quality control step (Andrews, 2010). The Trim Galore! tool (Galaxy Version 0.6.7+galaxy0) was then used three times in succession with quality trimming set to 30 and the minimum length set at 25 for each run, with each run including a different adapter sequence with overlap for adapters set as 5, for illumina, nextera and small RNA adapters, and other defaults (Krueger, 2021). This removed any adapter sequences that were present and trimmed the reads based on the quality scores and set a minimum length following this processing. The Trimmomatic tool (Galaxy Version 0.39+galaxy2) was then used with crop length set to 35 and other default settings to trim all reads to a maximum length of 35 bases (Bolger, Lohse and Usadel, 2014). The reason to restrict RP read lengths, in this analysis and across the project, is driven by the lengths of ribosomal footprints, generally agreed to be around 30 bases in length with some small variation (Martinez *et al.*, 2019; Liu *et al.*, 2020; François *et al.*, 2021; Schott *et al.*, 2021). To allow for this variation in RP read length, RP reads were restricted to between 25 and 35 bases. Then the FastQC tool was used again with defaults to provide another quality control step to check processing had been successful (Andrews, 2010). The RP reads were then aligned to the Galaxy in-built human hg38 genome using the HISAT2 tool (Galaxy Version 2.2.1+galaxy1) with unstranded and single-end reads settings and other defaults (Kim *et al.*, 2019). A prerequisite for this workflow is the annotated human hg38 genome uploaded in gff format from GENCODE, available at https://www.gencodegenes.org/human/ (Frankish *et al.*, 2019) to use with the htseq-count tool (Galaxy Version 2.0.5+galaxy0) (Anders, Pyl and Huber, 2015), with feature type set as 'CDS', attribute ID set as 'transcript_ID', stranded set as no, and other default settings, to quantify the HISAT2 alignments to produce count tables for the number of RP reads aligned to the CDS of transcripts.

The Galaxy workflow is very similar for the RNAseq datasets (Figure 2.2), however the Trimmomatic step is removed due to the RNAseq reads not being restricted to the size of ribosomal footprints. The other change is an addition of two further htseq-count runs with altered feature types to also include 'five_prime_UTR' or 'three_prime_UTR', otherwise htseq-count settings remain the same, meaning a count table is produced for the RP reads aligned to the CDS, 5' UTR and 3' UTR of transcripts (Anders, Pyl and Huber, 2015).

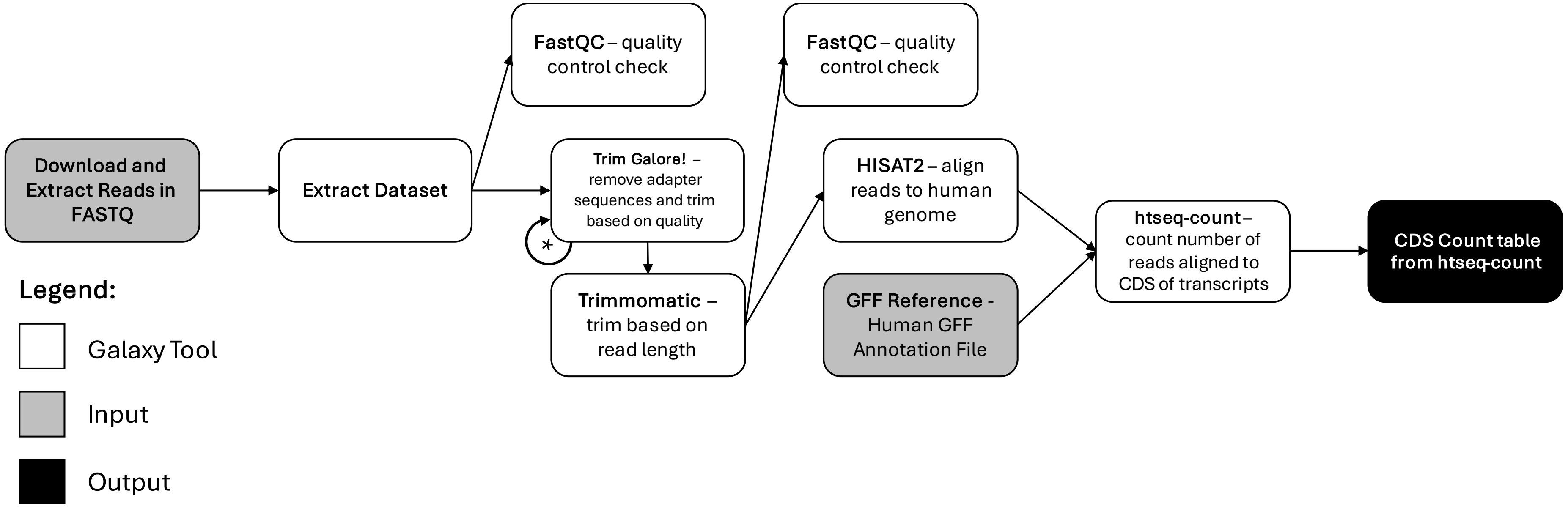
FastQC – quality control check
FastQC – quality control check
Download and Extract Reads in FASTQ
Extract Dataset
Trim Galore! – remove adapter sequences and trim based on quality
*
HISAT2 – align reads to human genome
htseq-count – count number of reads aligned to CDS of transcripts
CDS Count table from htseq-count
Trimmomatic – trim based on read length
GFF Reference - Human GFF Annotation File
Legend:
Galaxy Tool
Input
Output

Figure 2.1: Workflow for ribosome profiling (RP) dataset processing, alignment to human genome, and quantification of read alignment to transcripts. *This workflow was created and run through the Galaxy bioinformatics platform for RP datasets from PRJNA256316, PRJNA532400, and PRJNA880902 Sequence Read Archive (SRA) projects. The workflow highlights the key tools and a brief description of their functions. * The Trim Galore! tool is run three times in succession to ensure different adapter sequences are removed.*

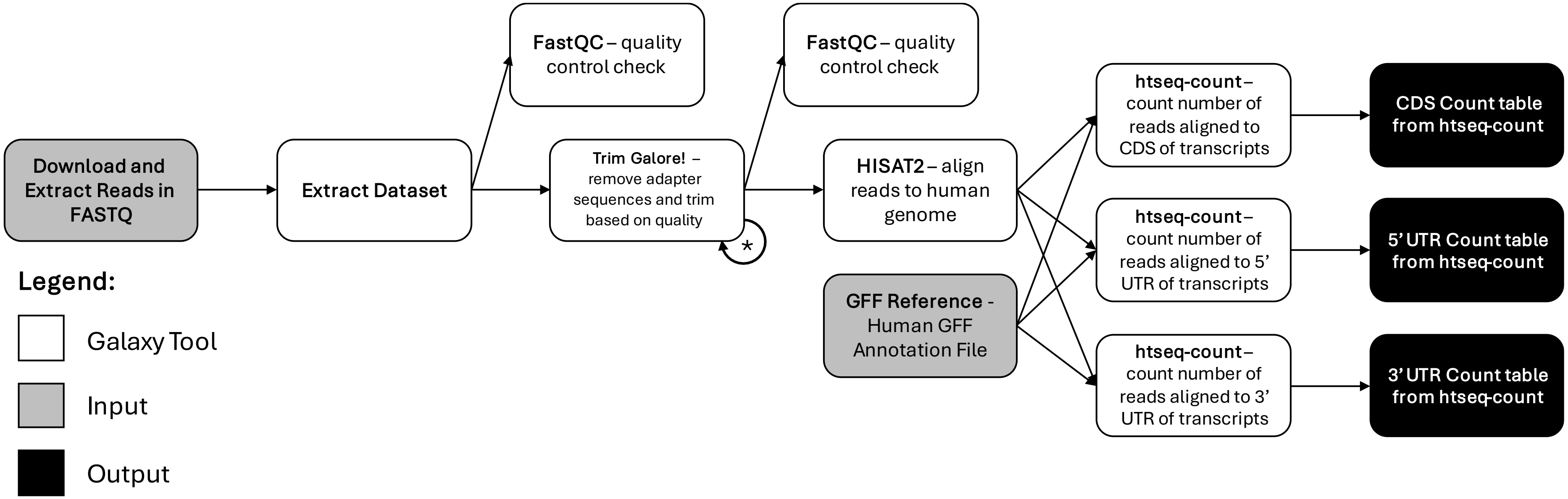
FastQC – quality control check
FastQC – quality control check
htseq-count– count number of reads aligned to CDS of transcripts
CDS Count table from htseq-count
Download and Extract Reads in FASTQ
Extract Dataset
Trim Galore! – remove adapter sequences and trim based on quality
*
HISAT2 – align reads to human genome
htseq-count– count number of reads aligned to 5' UTR of transcripts
5' UTR Count table from htseq-count
GFF Reference - Human GFF Annotation File
htseq-count– count number of reads aligned to 3' UTR of transcripts
3' UTR Count table from htseq-count
Legend:
Galaxy Tool
Input
Output

Figure 2.2: Workflow for RNAseq dataset processing, alignment to human genome, and quantification of read alignment to transcripts. *This workflow was created and run through the Galaxy bioinformatics platform for RNAseq datasets from PRJNA256316, PRJNA532400, and PRJNA880902 Sequence Read Archive (SRA) projects. The workflow highlights the key tools and a brief description of their functions. * The Trim Galore! tool is run three times in succession to ensure different adapter sequences are removed.*

In Galaxy the FastQC data from each RNAseq or RP dataset for each of the three SRA projects were pulled together using the MultiQC tool (Galaxy Version 1.24.1+galaxy0) by selecting the datasets to include in the analysis with inclusion of data for plots and other default settings (Ewels *et al.*, 2016). Tables 8.1 and 8.3 in Appendix 1 include a summary of the quality control data for each RNAseq dataset before, and after processing with the Galaxy workflow and the overall percentage alignment to the human genome provided by the HISAT2 output. The same summaries are also available for the RP datasets in Tables 8.2 and 8.4 in Appendix 1.

2.4.2 Differential expression of transcripts from RNAseq to RP datasets

To investigate the regulatory function of Wu dORFs a model for the translational regulation of transcripts was used. A similar method was used in the original paper reporting dORFs (Wu *et al.*, 2020b), using the change in expression of a transcript from the RNAseq dataset compared to the RP dataset. The RNAseq and RP dataset transcript expression was represented as a Reads Per Kilobase of transcript, per Million mapped reads (RPKM) normalized expression value. In RNAseq datasets the expression is calculated across the whole transcript, whereas RP dataset expression is calculated for the CDS. Then the log2 fold change in expression from RNAseq to RP datasets was calculated. Before the RPKM value can be calculated the number of reads aligned to each CDS, or transcript, was quantified for the RNAseq and RP datasets. The count tables for RNAseq datasets for the 5’ UTR, 3’ UTR and CDS were added together to generate a count across the whole transcript. To avoid the issue with zero values and infinite log2 fold changes, any transcript with zero reads aligned to the CDS, for RP datasets, or the transcript, for RNAseq datasets, were discarded. To allow the RPKM value to be calculated, all human transcript versions with RefSeq mRNA IDs were obtained using Ensembl Biomart with the transcript and CDS lengths (Harrison *et al.*, 2024). Only transcripts with RefSeq mRNA IDs are used. In RNAseq datasets the RPKM value for each transcript was calculated by initially dividing the number of reads aligned to the transcript by the total number of reads aligned to all the transcripts in the dataset divided by one million, then dividing this value by the transcript length in kilobases. The only difference for RP datasets is using reads aligned to the CDS and the CDS length in kilobases. This provides an expression value for each transcript version for each dataset. The log2 fold change in RPKM value from the paired RNAseq to RP dataset provided the differential expression of the transcript as a model of the translational regulation of the transcript. Then the transcripts were grouped into transcripts with and without Wu dORFs and comparisons were made between the groups and samples.

2.4.3 RP dataset 3’ UTR RPKM correlation with CDS RPKM and differential expression

The above method was altered slightly to calculate an RPKM expression value for the 3’ UTR of a transcript in the RP datasets. This provided a potential measure of the ribosomal association with the 3’ UTR and by extension, possibly the translation occurring. The previous method is changed at the htseq-count step. The htseq-count tool is run on the HISAT2 RP dataset alignment files, with the feature type, changed to ‘three_prime_utr’ from ‘CDS’, counting the number of reads aligned to the 3’ UTR of a transcript (Anders, Pyl and Huber, 2015). Ensembl Biomart is used to get all human RefSeq mRNA ID 3’ UTR sequences which were uploaded into Galaxy (Harrison *et al.*, 2024). The FASTA-to-Tabular tool (Galaxy Version 1.1.1) with default settings, converted the 3’ UTRs into a table with transcripts and sequences in columns which was used with the Search in textfiles (grep) tool (Galaxy Version 9.3+galaxy1) to find sequences which did not match ‘sequence unavailable’ (Grüning *et al.*, 2016), before converting the tabular output back to FASTA using the Tabular-to-FASTA tool (Galaxy Version 1.1.1) with headings as column 1 and sequences in column 2 (Afgan *et al.*, 2018). The 3’ UTR sequence lengths were gathered using the Compute sequence length tool (Galaxy Version 1.0.4) (Blankenberg *et al.*, 2010). Then the calculation of the RPKM remains the same, however, it used the number of reads aligned to the 3’ UTR, instead of the CDS, and the 3’ UTR length in kilobases for transcripts with reads aligned where the 3’ UTR length was obtainable. The expression values for the 3’ UTR in the RP datasets were used in a correlation analysis against the differential expression values and the CDS RP expression to look for possible relationships between the two.

2.5 Investigating ribosomal presence and possible translation in the 3’ UTR in PRJNA256316 RP datasets

The hypothesized mechanism for dORF function is suggested to involve translation of the dORF, leading to investigation of Wu dORF and 3’ UTR ribosome association and possible translation (Wu *et al.*, 2020b). To investigate ribosomal association, and possible translation, of Wu dORFs and 3’ UTRs, the PRJNA256316 RP datasets were used, by aligning the RP reads against the regions of interest. To compare with regions of known translation, there was also consideration of RP read alignment to CDS regions.

2.5.1 RP dataset preparation

The RP datasets required slightly different processing for the below analysis. It was important to ensure that processing did not alter the 5’ end of the RP read. The RP datasets previously imported with the Download and Extract Reads tools were used again (Leinonen, Sugawara and Shumway, 2011). Adapter sequences were removed using the Trim Galore! tool set with defaults, which does not affect the 5’ end of the reads (Krueger, 2021). Then following adapter removal, the Trimmomatic tool was used for quality and length trimming (Bolger, Lohse and Usadel, 2014). The settings for this tool included: Sliding Window Trimming across 4 bases at quality 30, Cut bases off end at quality 30, MINLEN set at 28, CROP set to 30, and all others as defaults (Bolger, Lohse and Usadel, 2014). These settings discarded any reads with a quality score of less than thirty averaged across each four bases, then from the 3’ end bases were removed if the quality score is less than thirty, before the reads were trimmed to thirty bases at the 3’ end as the maximum length and any reads shorter than twenty eight bases were removed (Bolger, Lohse and Usadel, 2014). Following the processing, the RP datasets were checked again with the FastQC tool (Andrews, 2010).

2.5.2 Aligning RP reads to dORFs, 3’ UTR and CDS

RefSeq mRNA ID 3’ UTR sequences were obtained from Ensembl Biomart as described in section 2.4.3 (Harrison *et al.*, 2024). The methods described in this section were carried out on all the PRJNA256316 RP datasets, summarized in Figure

2.3. Any duplicated 3' UTR sequences in other transcripts were removed to prevent RP reads aligning to the same site in identical 3' UTR transcripts. In Galaxy, the NCBI BLAST+ blastn tool (Galaxy Version 2.14.1+galaxy2) was used to align the RP reads to the 3' UTR sequences (Altschul *et al.*, 1997; Camacho *et al.*, 2009; Cock *et al.*, 2015). This tool searched a nucleotide database, made from the 3' UTR sequences using the NCBI BLAST+ makeblastdb tool (Galaxy Version 2.14.1+galaxy2) (Altschul *et al.*, 1997; Camacho *et al.*, 2009; Cock *et al.*, 2015). Before aligning, the FASTQ to FASTA tool (Galaxy Version 1.0.2+galaxy2) converted the RP dataset format into FASTA (Blankenberg *et al.*, 2010). The NCBI BLAST+ blastn tool was run with the RP read sequences in FASTA format as the query sequence, the 3'UTR sequences as the subject database, percentage identity and coverage cutoff set at 100, DUST deactivated, forward strand alignments, the extended column output and all other default settings (Altschul *et al.*, 1997; Camacho *et al.*, 2009; Cock *et al.*, 2015). This meant that RP reads must align completely within the 3' UTR sequence to be reported, meaning any overlap across regions was discarded. The table of alignments for each RP dataset was then downloaded and processed further in Microsoft Excel. Initially any RP reads aligned to multiple 3' UTRs or locations were discarded. This left RP reads entirely aligned to one 3' UTR transcript in one location (Figure 2.3). Then these RP reads were aligned following the same method against CDS and 5' UTR sequences with RefSeq mRNA IDs also obtained and processed as described for the 3' UTR. Then any 3' UTR aligned RP reads aligned entirely to these regions were also discarded, leaving RP reads which were aligned fully to the 3' UTR and with confidence in this alignment location (Figure 2.3).

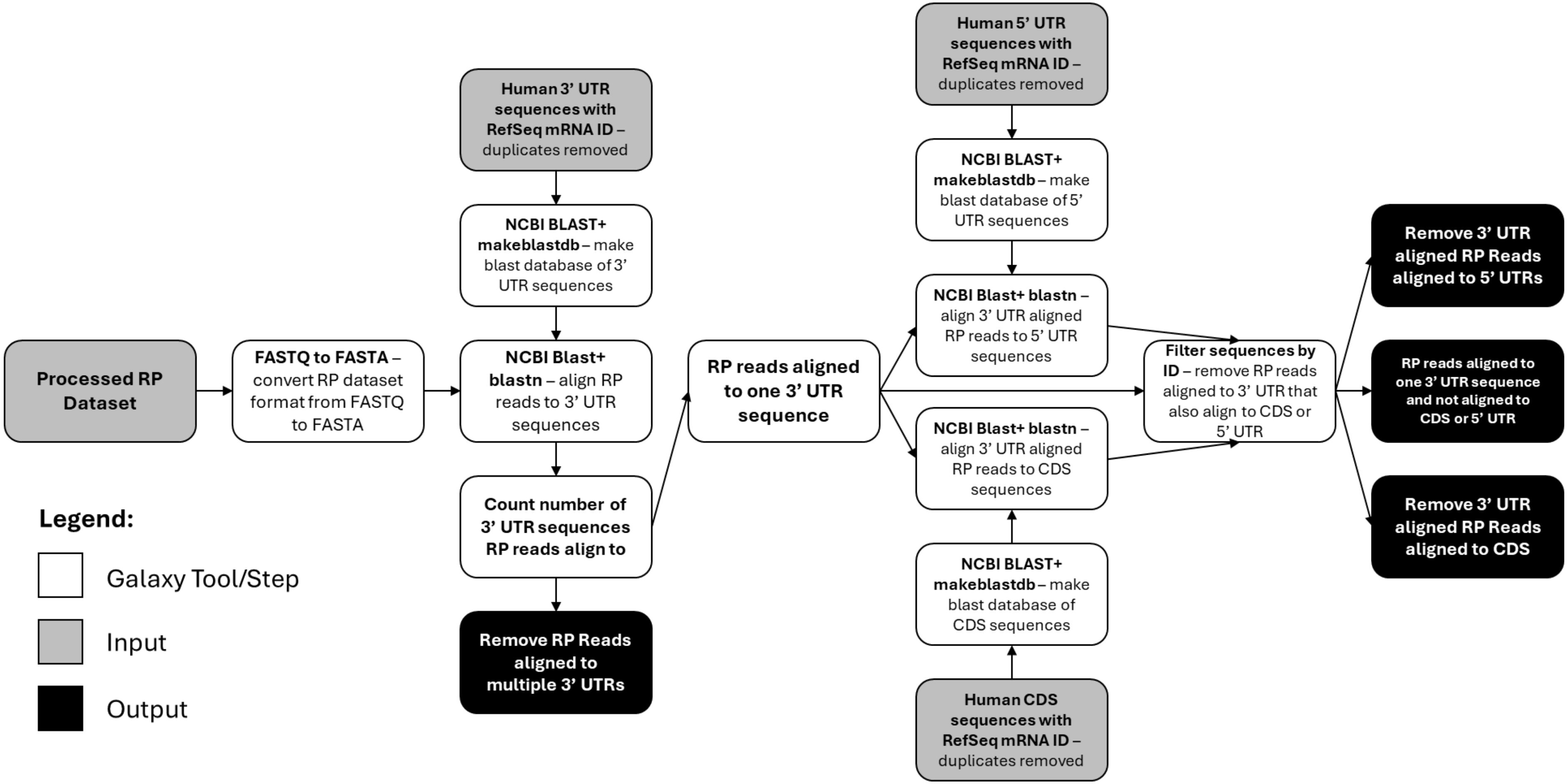
Human 3’ UTR sequences with RefSeq mRNA ID – duplicates removed
NCBI BLAST+ makeblastdb – make blast database of 3’ UTR sequences
Processed RP Dataset
FASTQ to FASTA – convert RP dataset format from FASTQ to FASTA
NCBI Blast+ blastn – align RP reads to 3’ UTR sequences
Count number of 3’ UTR sequences RP reads align to
Remove RP Reads aligned to multiple 3’ UTRs
RP reads aligned to one 3’ UTR sequence
Human 5’ UTR sequences with RefSeq mRNA ID – duplicates removed
NCBI BLAST+ makeblastdb – make blast database of 5’ UTR sequences
NCBI Blast+ blastn – align 3’ UTR aligned RP reads to 5’ UTR sequences
NCBI Blast+ blastn – align 3’ UTR aligned RP reads to CDS sequences
NCBI BLAST+ makeblastdb – make blast database of CDS sequences
Human CDS sequences with RefSeq mRNA ID – duplicates removed
Filter sequences by ID – remove RP reads aligned to 3’ UTR that also align to CDS or 5’ UTR
Remove 3’ UTR aligned RP Reads aligned to 5’ UTRs
RP reads aligned to one 3’ UTR sequence and not aligned to CDS or 5’ UTR
Remove 3’ UTR aligned RP Reads aligned to CDS
Legend:
Galaxy Tool/Step
Input
Output

Figure 2.3: Workflow to align ribosome profiling (RP) reads to 3’ untranslated region (UTR) sequences with RefSeq mRNA IDs. *This workflow was created and run through the Galaxy platform. The processed RP dataset was converted into FASTA format and aligned to human RefSeq mRNA ID 3’ UTRs using the NCBI BLAST+ blastn tool. If the entire RP read is not aligned, or the RP read is aligned to multiple 3’ UTRs, the coding sequence (CDS) or 5’ UTRs, they are removed to leave RP reads fully aligned to RefSeq mRNA IDs 3’ UTRs. The workflow highlights the key tools and a brief description of their functions.*

The 3’ UTR aligned RP reads were then checked for alignment within a dORF (Figure 2.4). Before running alignment between the RP reads aligned to the 3’ UTR and dORF sequences, the dORF sequence databases were created in Galaxy using the NCBI BLAST+ makeblastdb tool (Altschul *et al.*, 1997; Camacho *et al.*, 2009; Cock *et al.*, 2015). This was done with the Wu dORFs (Wu *et al.*, 2020b), and also another database of ‘Potential dORF’ sequences found in the RefSeq mRNA ID 3’ UTR sequences using the getorf tool (Galaxy Version 5.0.0.1) in Galaxy (Figure 2.4) (P Rice, Longden and Bleasby, 2000; Blankenberg *et al.*, 2007). This tool was run with the following settings: maximum ORF length 300, standard code with alternative start codons, output nucleic sequence between start and stop codon, no ORFs in reverse complement and all other default settings, generating a list of dORFs with length between 30-300 bases, known as ‘Potential dORFs’ (P Rice, Longden and Bleasby, 2000; Blankenberg *et al.*, 2007). In Galaxy, the RP reads aligned to the 3’ UTR were then used with the NCBI BLAST+ blastn tool using the same settings described previously and were run against the different databases of Wu dORFs and ‘Potential dORFs’ (Altschul *et al.*, 1997; Camacho *et al.*, 2009; Cock *et al.*, 2015). The output was a table of RP reads aligned to the 3’ UTR which are also shown to align fully to either Wu or ‘Potential dORFs’, allowing the 3’ UTR aligned RP reads to be grouped as aligned within a Wu dORF, ‘Potential dORF’ or remaining 3’ UTR sequence (Figure 2.4). This alignment only considers reads fully aligned within these dORFs to ensure maximum stringency of the alignment. However, this could mean that some RP reads aligned partially to dORFs are not annotated as such.

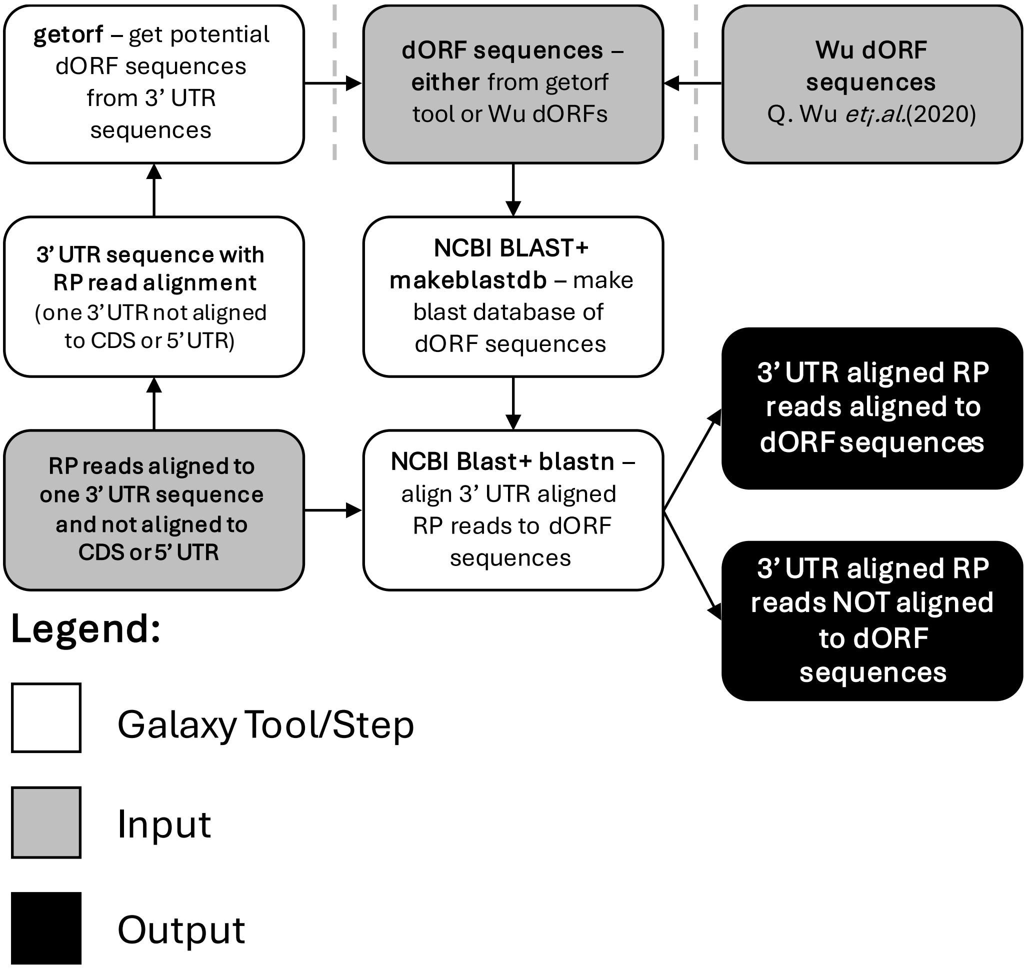
getorf – get potential dORF sequences from 3' UTR sequences
dORF sequences – either from getorf tool or Wu dORFs
Wu dORF sequences Q. Wu etj.al.(2020)
3' UTR sequence with RP read alignment (one 3' UTR not aligned to CDS or 5' UTR)
NCBI BLAST+ makeblastdb – make blast database of dORF sequences
3' UTR aligned RP reads aligned to dORF sequences
RP reads aligned to one 3' UTR sequence and not aligned to CDS or 5' UTR
NCBI Blast+ blastn – align 3' UTR aligned RP reads to dORF sequences
3' UTR aligned RP reads NOT aligned to dORF sequences
Legend:
Galaxy Tool/Step
Input
Output

Figure 2.4: Workflow to align 3' untranslated region (UTR) aligned ribosomal profiling (RP) reads to downstream open reading frames (dORFs). *This workflow was created and run on the Galaxy platform. Wu dORFs, presented by Wu et al. (2020b), or 'Potential dORFs', found using getORF tool, were aligned against the 3' UTR aligned RP reads using the NCBI BLAST+ blastn tool to reveal RP reads that also fully aligned to dORFs. The workflow highlights the key tools and a brief description of their functions.*

2.5.3 RP read density analysis

The RP read density of all the RP reads, and the subsets of RP reads, described in section 3.6, in the CDSs, 3' UTRs and Wu dORFs was assessed to consider their ribosomal association. The RP read density reported the total RP reads aligned, divided by the total length of either the CDSs, 3' UTRs or Wu dORFs in transcripts with RP alignments. This value was adjusted for different RP dataset sizes by dividing the RP read density by the total reads in the dataset then multiplying this by 20 000 000, a value similar to the original size of the datasets.

2.6 Shortlist Candidate Wu dORFs

Several proteomics databases containing short proteins were used to search for Wu dORF proteins validated by MS. Short proteins validated with MS were obtained from OpenProt (Leblanc *et al.*, 2024), SmProt (Y. Li *et al.*, 2021) and MetamORF (Choteau *et al.*, 2021) databases. In Galaxy, these short proteins were converted into a protein blast database using the NCBI BLAST+ makeblastdb tool (Altschul *et al.*, 1997; Camacho *et al.*, 2009; Cock *et al.*, 2015). The Wu dORFs were uploaded to Galaxy and the transeq tool (Galaxy Version 5.0.0) was used with code to use set as standard (with alternative initiation codons) and '*' or X characters for the stop codon were removed to leave the protein sequences of the Wu dORFs (Afgan *et al.*, 2018). The Wu dORF proteins were run against the MS validated protein database with NCBI BLAST+ blastp tool (Galaxy Version 2.14.1+galaxy2) with coverage set to 100% and other default settings (Altschul *et al.*, 1997; Camacho *et al.*, 2009; Cock *et al.*, 2015). Those Wu dORFs with MS validation were referred to as MSVW dORFs (MS validated Wu dORFs).

2.7 Proportion of 3' UTR covered by 'Potential dORFs'

When considering the proportion of 3' UTR aligned RP reads that align within dORF sequences, it is important to consider what proportion of 3' UTRs the dORFs cover. 'Potential dORFs' refers to any dORF sequences appearing in RefSeq annotated 3' UTRs. The coverage of 3' UTR by 'Potential dORFs' would show the abundance of these dORFs within 3' UTRs. The output from the getorf tool, run previously and described in section 2.5.2, presents the transcript, dORF number, dORF start location within the 3' UTR, dORF stop within the 3' UTR, and the sequence. A spreadsheet was produced from these outputs with the transcript and the 'Potential dORF' start and stop locations. This spreadsheet is then used with the dORF_Coverage_of_3UTR.py Python script which reported the total coverage of the 3' UTR by 'Potential dORFs' within each transcript (Figure 2.5). The Python script accounted for overlapping dORFs and, rather than adding the dORF lengths together, calculated the start and stop of overlapped regions of dORFs (Figure 2.5). The output from the Python script was used to calculate the percentage of the transcript length covered by 'Potential dORFs', using the lengths of RefSeq mRNA ID 3' UTR sequences generated in section 2.4.3. This percentage coverage data can then be restricted to different transcript groups, for example the transcripts with RP read alignment in the PRJNA256316 RP datasets.

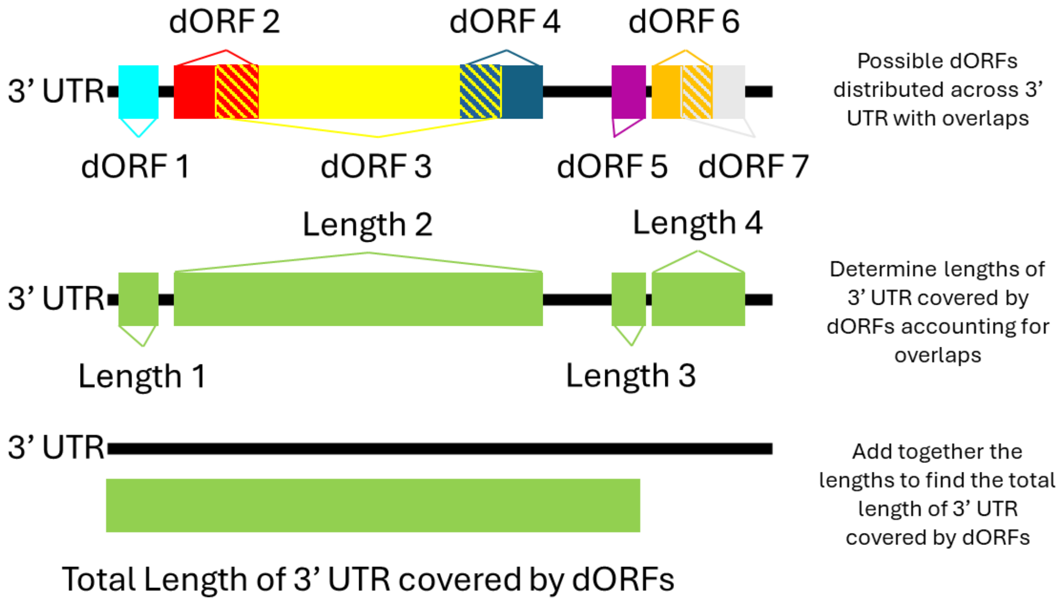
dORF 2
dORF 4
dORF 6
3' UTR
Possible dORFs distributed across 3' UTR with overlaps
dORF 1
dORF 3
dORF 5
dORF 7
Length 2
Length 4
3' UTR
Determine lengths of 3' UTR covered by dORFs accounting for overlaps
Length 1
Length 3
3' UTR
Add together the lengths to find the total length of 3' UTR covered by dORFs
Total Length of 3' UTR covered by dORFs

Figure 2.5: Representation of how the dORF_Coverage_of_3UTR.py Python script determines the total length of a 3' untranslated region (UTR) sequence covered by potential downstream open reading frames (dORFs). *The script determines the length of the 3' UTR sequence that is covered by dORFs, regardless of whether dORFs overlap.*

2.8 Nucleotide, Dinucleotide and Trinucleotide composition of 3' UTR, CDS and Wu dORFs

In Galaxy, the RefSeq mRNA ID 3' UTR sequences, with duplicates removed, were used with the shuffleseq tool (Galaxy Version 5.0.0.1) with default settings, changing position of bases, but maintaining the length and sequence composition (P Rice, Longden and Bleasby, 2000; Blankenberg *et al.*, 2007). This was repeated twice more, leaving the genomic 3' UTR sequences and three shuffled repeats. The nucleotide, dinucleotide, and trinucleotide frequency in the genomic and shuffled RefSeq mRNA ID 3' UTR sequences, Wu dORF-containing RefSeq 3' UTR sequences, Wu dORF sequences, and RefSeq mRNA ID CDSs was calculated using the compseq tool (Galaxy Version 5.0.0.1) on each set of sequences in Galaxy with word sizes of 1, 2 and 3, respectively (P Rice, Longden and Bleasby, 2000; Blankenberg *et al.*, 2007). To find the codon frequency in the RefSeq mRNA ID CDS and Wu dORF sequences the compseq tool was used, with a word size of 3 in frame 1 (P Rice, Longden and Bleasby, 2000; Blankenberg *et al.*, 2007). The comparisons of codon frequency excluded stop codons, due to large difference in size and number of sequences being considered with only one stop codon per sequence. This analysis allowed comparisons of specific dinucleotides, trinucleotides and codons across the different sequences, such as start codons.

2.8.1 CDS Stop codons preceding Wu dORFs

This analysis compared the stop codon frequency seen across human CDSs and CDSs that precede 3' UTRs that contain Wu dORFs. In Galaxy, the compseq tool was used, with a word size of 3 in frame 1 to find the codon frequency of CDSs that precede 3' UTRs that contain Wu dORFs (P Rice, Longden and Bleasby, 2000; Blankenberg *et al.*, 2007). The human stop codon usage was obtained from the Codon Statistics Database: A Database of Codon Usage Bias accessible from http://codonstatsdb.unr.edu (Subramanian *et al.*, 2022).

2.9 AUG dORF and 3' UTR Conservation Analysis

This research investigated the conservation of all potential human dORFs starting with an AUG start codon (AUG dORFs) with comparison to their 3' UTR across a variety of species. Initially the NCBI nucleotide database was used to gather all human RefSeq annotated transcripts by filtering for 'human' and 'RefSeq', then downloading 199205 sequences in FASTA format (Sayers *et al.*, 2022). These sequences were then uploaded into Galaxy and run with the FASTA-to-Tabular tool with default settings to create a table with the transcript ID in column 1 and the sequence in column 2 (Afgan *et al.*, 2018).

2.9.1 Python Scripts used in dORF and 3' UTR conservation analysis

Running the Human_RefSeq_3UTR_and_AUG_dORF_Sequences.py Python script generated a dataframe with the available human 3' UTR sequences containing dORFs, with AUG start codons, within these 3' UTRs by using the transcript IDs collected above and the NCBI nucleotide database (Figure 2.6) (Sayers *et al.*, 2022). The script also collected 3' UTR sequences which didn't contain these dORFs and also transcript IDs which could not collect 3' UTR sequences. The script makes use of the pandas module (version – 2.2.0) to process and generate results dataframes (McKinney, 2011) to collect the data relating to the NCBI nucleotide database entry for each transcript (Sayers *et al.*, 2022). The data from these entries were collected using the Entrez module from bioPython (version – 1.83) (Figure 2.6) (Cock *et al.*, 2009). The 3' UTR sequences were gathered using the full transcript sequence and CDS location within the transcript to gather the transcript sequence following the CDS end where possible. Another important module was the regular expression, or re, module (version – 2.2.1), used to gather the dORF sequences with AUG start codons and gather data about these dORFs within the 3' UTR. Another shorter Python script, Human_dORF_ID_remove_duplicate_3UTRs.py, was then used on the dataframe of AUG dORF-containing human RefSeq mRNA ID 3' UTRs to add a unique dORF ID and remove duplicate 3' UTR sequences (Figure 2.6).

Then to gather homologous genes to investigate dORF and 3' UTR conservation across other species the NCBI HomoloGene database was used (Sayers *et al.*, 2022). This database contained the homolog data for species shown in Table 2.3. This database was downloaded in XML format to be used to collect the homolog genes of the human genes containing AUG dORFs in other species. The Homolog_Gene_Lists_and_nucID_in_species.py Python script was used to do this and collected lists of homolog genes for each species (Figure 2.6).

In each species the homolog gene lists were used to collect transcript IDs to use with the NCBI nucleotide database (Sayers *et al.*, 2022) to gather AUG dORF-containing RefSeq 3' UTR sequences in homolog genes to the human genes with AUG dORFs in the 3' UTR. To do this a similar Python script to Human_RefSeq_3UTR_and_AUG_dORF_Sequences.py was used for each species using similar tools and commands (Figure 2.6). Running Ptroglodytes_Get_homolog_3UTRs_and_dORFs.py, and other Python scripts for each species, example is for chimpanzee, collected homolog 3' UTR sequences to the human genes with AUG dORF-containing 3' UTRs. These 3' UTRs from other species were split by the Python scripts into those that contained AUG dORFs and those which did not (Figure 2.6). These Python scripts also included a step to remove duplicate 3' UTRs and assign a dORF ID.

Table 2.3: Species included in the NCBI HomoloGene database. *The table includes the species, common name, and taxonomy ID for each of the included species.*

Species	**Common Name**	**Taxonomy ID**
Homo sapiens	Human	9606
Pan troglodytes	Chimpanzee	9598
Macaca mulatta	Rhesus Monkey	9544
Canis lupus familiaris	Dog	9615
Bos taurus	Cow	9913
Mus musculus	Mouse	10090
Rattus norvegicus	Rat	10116
Gallus gallus	Chicken	9031
Xenopus tropicalis	Western Clawed Frog	8364
Danio rerio	Zebrafish	7955
Drosophila melanogaster	Fruit Fly	7227
Anopheles gambiae	Malaria Mosquito	7165
Caenorhabditis elegans	Nematode	6239
Saccharomyces cerevisiae	Budding Yeast	4932
Kluyveromyces lactis	Ascomycetes	28985
Eremothecium gossypii	Ascomycetes	33169
Schizosaccharomyces pombe	Fission Yeast	4896
Magnaporthe oryzae	Rice Blast Fungus	318829
Neurospora crassa	Ascomycetes	5141
Arabidopsis thaliana	Thale Cress	3702
Oryza sativa	Rice	4530

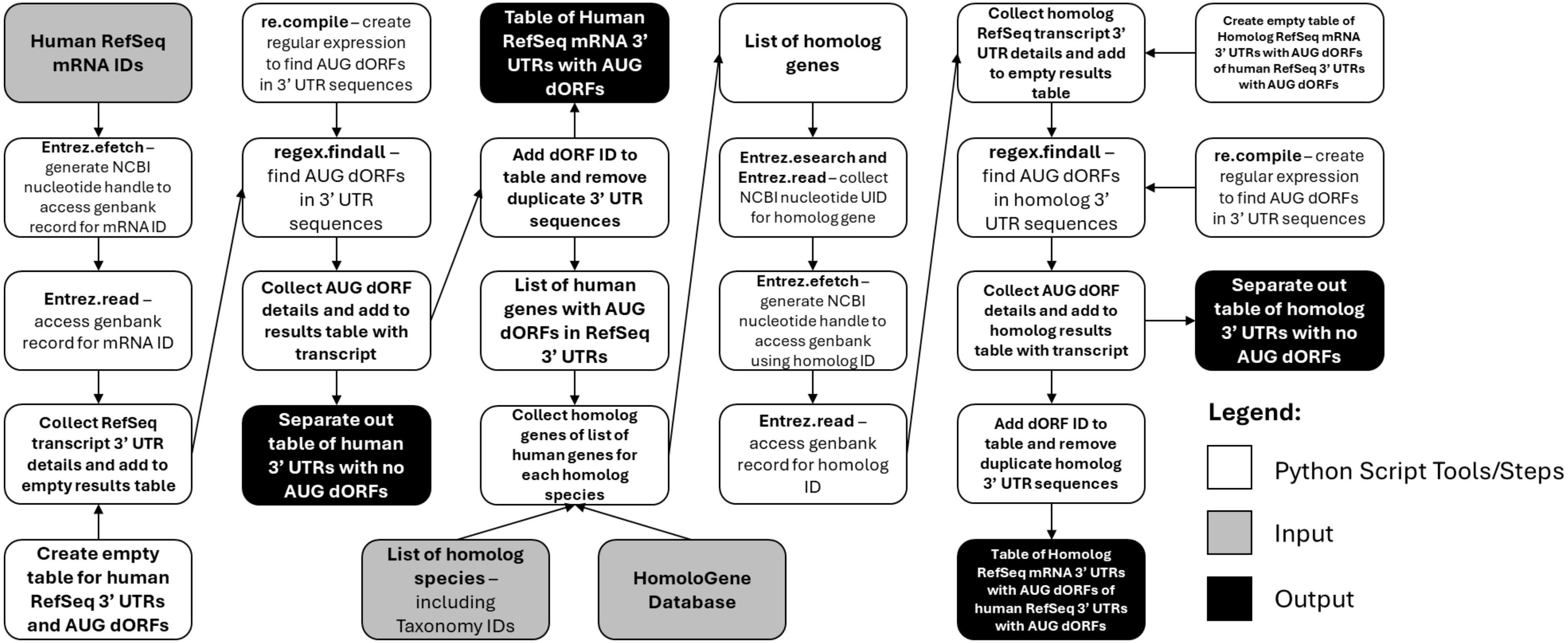
Human RefSeq mRNA IDs
Entrez.efetch – generate NCBI nucleotide handle to access genbank record for mRNA ID
Entrez.read – access genbank record for mRNA ID
Collect RefSeq transcript 3' UTR details and add to empty results table
Create empty table for human RefSeq 3' UTRs and AUG dORFs
re.compile – create regular expression to find AUG dORFs in 3' UTR sequences
regex.findall – find AUG dORFs in 3' UTR sequences
Collect AUG dORF details and add to results table with transcript
Separate out table of human 3' UTRs with no AUG dORFs
Table of Human RefSeq mRNA 3' UTRs with AUG dORFs
Add dORF ID to table and remove duplicate 3' UTR sequences
List of human genes with AUG dORFs in RefSeq 3' UTRs
Collect homolog genes of list of human genes for each homolog species
List of homolog species – including Taxonomy IDs
HomoloGene Database
List of homolog genes
Entrez.esearch and Entrez.read – collect NCBI nucleotide UID for homolog gene
Entrez.efetch – generate NCBI nucleotide handle to access genbank using homolog ID
Entrez.read – access genbank record for homolog ID
Collect homolog RefSeq transcript 3' UTR details and add to empty results table
regex.findall – find AUG dORFs in homolog 3' UTR sequences
Collect AUG dORF details and add to homolog results table with transcript
Add dORF ID to table and remove duplicate homolog 3' UTR sequences
Table of Homolog RefSeq mRNA 3' UTRs with AUG dORFs of human RefSeq 3' UTRs with AUG dORFs
Create empty table of Homolog RefSeq mRNA 3' UTRs with AUG dORFs of human RefSeq 3' UTRs with AUG dORFs
re.compile – create regular expression to find AUG dORFs in 3' UTR sequences
Separate out table of homolog 3' UTRs with no AUG dORFs
Legend:
Python Script Tools/Steps
Input
Output

Figure 2.6: Workflow describing how the Python scripts gather RefSeq human and homolog 3' untranslated region (UTR) sequences with downstream open reading frames (dORFs) starting with an AUG start codon. *Python Scripts find human AUG dORF containing RefSeq mRNA ID 3' UTRs and then identify RefSeq homolog genes in other species using the HomoloGene database and for each species gather RefSeq homolog 3' UTRs with AUG dORFs. The workflow highlights the key tools, or commands, and a brief description of their functions.*

To investigate conservation of the AUG dORFs and 3' UTR sequences between humans and each homolog species, each human AUG dORF was compared against the homolog species AUG dORFs where the human AUG dORF gene matched the homolog gene in the other species (Figure 2.7). The most similar homolog species AUG dORF to the human AUG dORF was then used to compare these two dORF sequences, considering the relative location in the 3' UTR, the lengths and the sequence similarity. In addition, the 3' UTR sequences in the human and homolog species for this dORF comparison were also compared considering lengths and sequences. This was carried out for every human AUG dORF in every homolog species with matching homolog genes with AUG dORFs described previously. To do the similarity comparison between sequences the Smith-Waterman algorithm was used (Smith and Waterman, 1981). Running the Ptroglodytes_dORF_Conservation_Analysis.py Python script, and other Python scripts for each species (example is for chimpanzee), generated results dataframes for each species with comparisons of dORF and 3' UTR similarity (Figure 2.7). These scripts make use of the operating system, or os, module to access the command line from within the Python scripts, and some of the previously described tools and commands, also making use of the EMBOSS WATER tool (version – 6.6.0.0) to run the Smith-Waterman local alignment between the sequences (Smith and Waterman, 1981; Peter Rice, Longden and Bleasby, 2000). The subsequent results dataframes for each species were then run with Python scripts, such as Ptroglodytes_Filter_dORF_Conservation_ResultsDF.py for each species (Chimpanzee example given). This script removed any duplicate human AUG dORF results, keeping the comparison with greatest similarity. It also filtered the results to ensure that the similarity comparison for dORF and 3' UTRs between the human and homolog species had an alignment length at least as long as the shorter of the two aligned sequences (Figure 2.7). Results in Appendix 2 were from modified filtering scripts, allowing the dORF alignment length to be 90% and 80% of the shortest dORF being aligned.

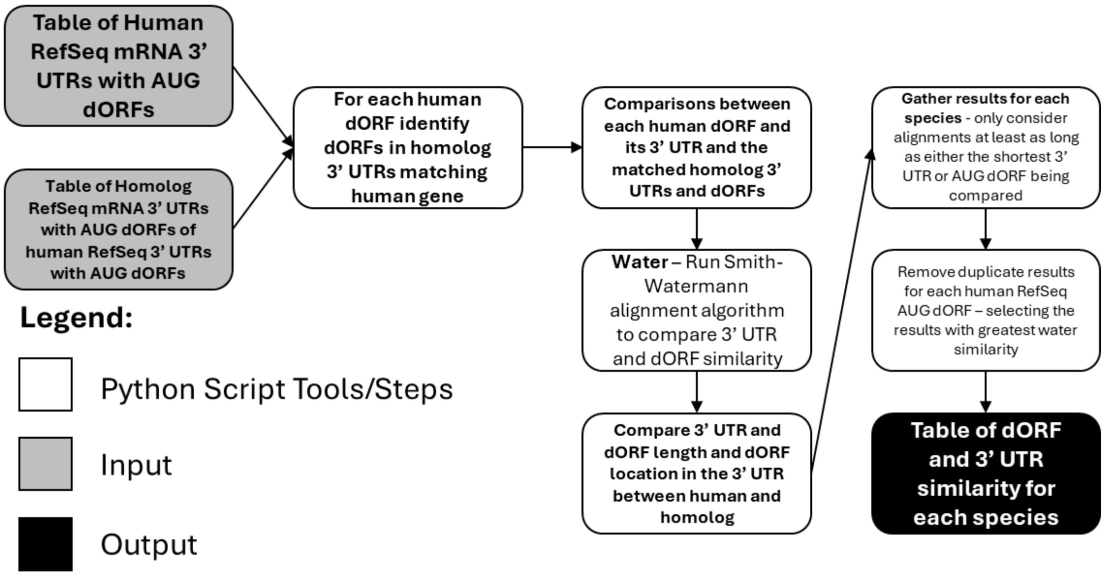
Table of Human RefSeq mRNA 3' UTRs with AUG dORFs
Table of Homolog RefSeq mRNA 3' UTRs with AUG dORFs of human RefSeq 3' UTRs with AUG dORFs
Legend:
Python Script Tools/Steps
Input
Output
For each human dORF identify dORFs in homolog 3' UTRs matching human gene
Comparisons between each human dORF and its 3' UTR and the matched homolog 3' UTRs and dORFs
Water – Run Smith-Watermann alignment algorithm to compare 3' UTR and dORF similarity
Compare 3' UTR and dORF length and dORF location in the 3' UTR between human and homolog
Gather results for each species - only consider alignments at least as long as either the shortest 3' UTR or AUG dORF being compared
Remove duplicate results for each human RefSeq AUG dORF – selecting the results with greatest water similarity
Table of dORF and 3' UTR similarity for each species

Figure 2.7: Workflow describing how Python scripts carry out conservation analysis to compare human 3' untranslated region (UTR) and downstream open reading frame (dORF) similarity against homolog 3' UTR and dORFs. *The workflow shows that the Smith-Waterman alignment algorithm is used to run this similarity analysis and highlights the key tools, or commands, and a brief description of their functions.*

2.9.2 Developing control sequences and validating conservation analysis

To help validate findings from the previous section, the results dataframes for each species were used to develop various control and validating sequences (Figure 2.9) to carry out further analysis using Python scripts. Each results dataframe comparing human and a homolog species were run with a Python script such as Ptroglodytes_Control_100_downstream_Conservation_Analysis.py (chimpanzee example) for each species. Where possible these scripts generated additional similarity results comparing a control section of the human and homolog 3' UTRs, that was the same length as the original AUG dORFs, 100 bases downstream of the AUG dORFs originally compared for each dORF comparison (Figures 2.8 and 2.9). The similarity analysis of these downstream control sequences was the same as the previous section, with restriction to only consider alignments at least as short as the shorter of the two control sequences. Very similar Python scripts, differing in control sequence location, for each species were also run to compare control sequences 200 and 500 bases downstream of the dORFs and also 100, 200, and 500 bases upstream of the dORF in the human or homolog species (Figure 2.8). These scripts explored conservation of the 3' UTR sequence surrounding the AUG dORFs.

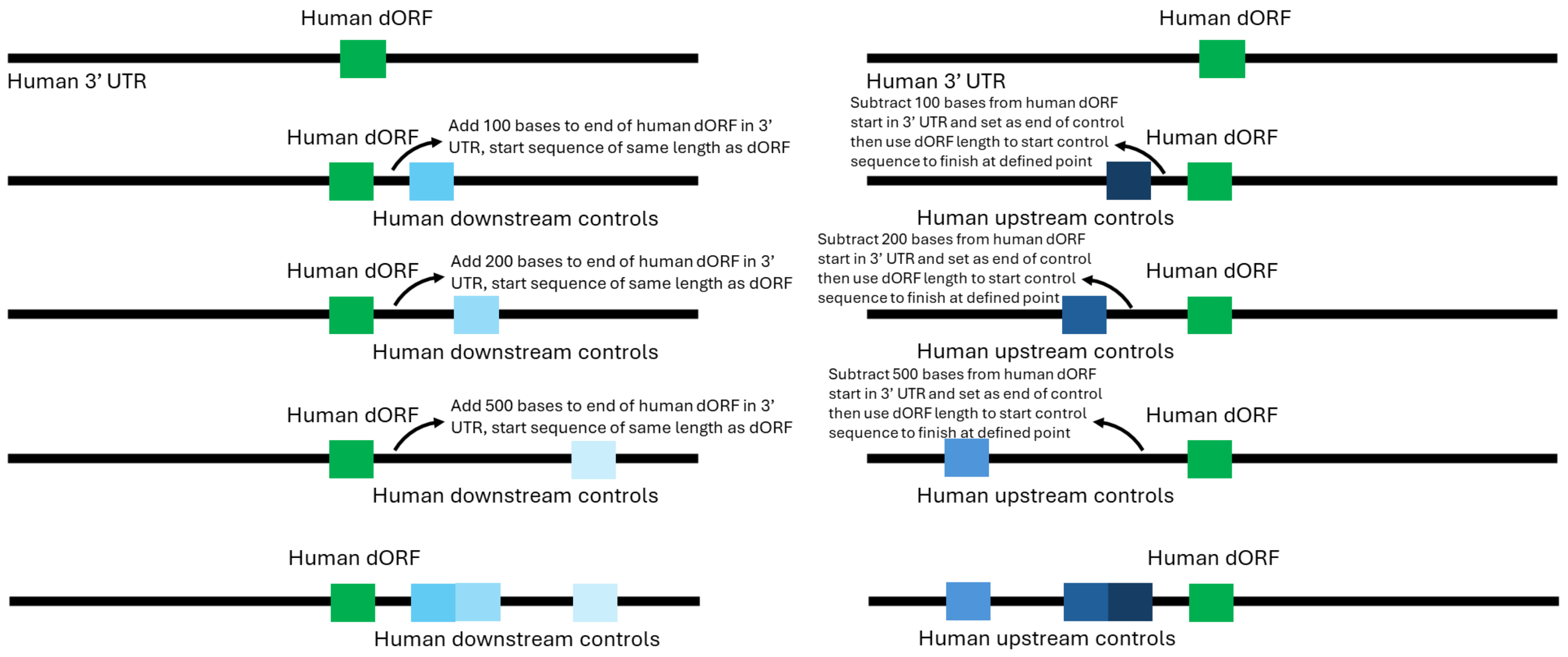
Human dORF
Human 3' UTR
Human dORF
Add 100 bases to end of human dORF in 3' UTR, start sequence of same length as dORF
Human downstream controls
Human dORF
Add 200 bases to end of human dORF in 3' UTR, start sequence of same length as dORF
Human downstream controls
Human dORF
Add 500 bases to end of human dORF in 3' UTR, start sequence of same length as dORF
Human downstream controls
Human dORF
Human downstream controls
Human dORF
Human 3' UTR
Subtract 100 bases from human dORF start in 3' UTR and set as end of control then use dORF length to start control sequence to finish at defined point
Human dORF
Human upstream controls
Subtract 200 bases from human dORF start in 3' UTR and set as end of control then use dORF length to start control sequence to finish at defined point
Human dORF
Human upstream controls
Subtract 500 bases from human dORF start in 3' UTR and set as end of control then use dORF length to start control sequence to finish at defined point
Human dORF
Human upstream controls
Human dORF
Human upstream controls

Figure 2.8: Diagram to represent how Python scripts generate control sequences upstream and downstream of downstream open reading frames (dORFs). *The control sequences were the same length as the AUG dORF starting 100, 200, and 500 bases following the dORF stop codon or ending 100, 200 or 500 bases before the dORF start codon within the 3' untranslated region (UTR). The black line indicates the 3' UTR, the green box indicates the location of the dORF, in blue are the various control sequences generated in the 3' UTR up and downstream of the dORF. The varying shades of blue show each of the different control sequence locations.*

In addition, control sequences were also generated from the middle of the human 3' UTRs, investigating whether other parts of the 3'UTR, of similar length to AUG dORFs or aligned regions, could also be conserved in other species (Figure 2.9). The results dataframes comparing human and homolog species were used with Python scripts, such as Ptroglodytes_Control_3UTR_middle_Conservation_Analysis.py (chimpanzee example) for each species,. These scripts used the results dataframe and removed duplicate human 3' UTRs in the dataframe keeping the most similar dORF comparison for each 3' UTR. Then for each human 3' UTR, it took a control sequence starting at the centre of the 3' UTR that was the same length as the original AUG dORF that was compared. This control sequence was run with the Smith-Waterman algorithm to locally align this control sequence against the whole homolog 3' UTR that had been compared originally for the AUG dORF similarity analysis with the human 3' UTR (Figure 2.9) (Smith and Waterman, 1981). In these Python scripts only results where the alignment length was as long as the control sequence were included.

A final validation step was to remove the influence of the AUG start codons from the dORF similarity comparison, as these codons were present in all dORFs compared, potentially increasing the similarity of the sequences (Figure 2.9). The results dataframes comparing human and homolog species were used with Python scripts, such as Ptroglodytes_Control_Remove_AUG_Conservation_Analysis.py (chimpanzee example) for each species. These scripts took the compared human and homolog AUG dORF sequences and removed the AUG start codons before comparing the dORF similarity again in the same manner as described previously (Figure 2.9).

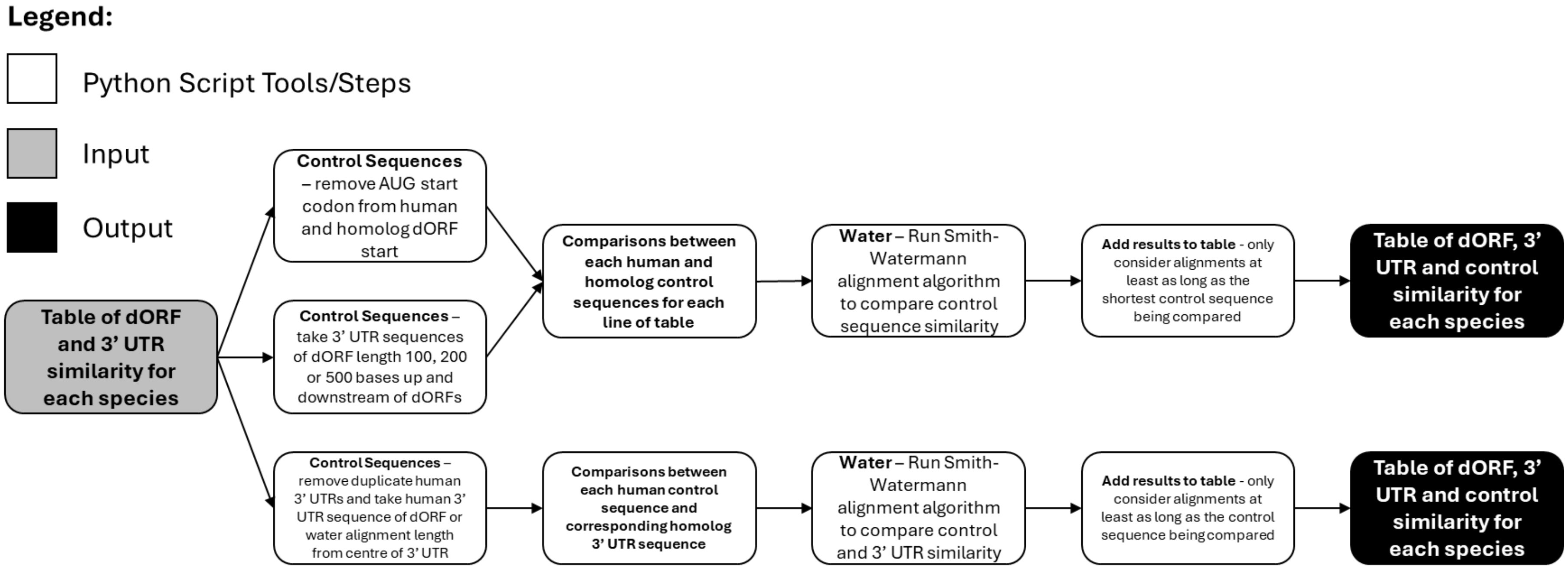
Legend:
Python Script Tools/Steps
Input
Output
Table of dORF and 3' UTR similarity for each species
Control Sequences – remove AUG start codon from human and homolog dORF start
Control Sequences – take 3' UTR sequences of dORF length 100, 200 or 500 bases up and downstream of dORFs
Control Sequences – remove duplicate human 3' UTRs and take human 3' UTR sequence of dORF or water alignment length from centre of 3' UTR
Comparisons between each human and homolog control sequences for each line of table
Comparisons between each human control sequence and corresponding homolog 3' UTR sequence
Water – Run Smith-Watermann alignment algorithm to compare control sequence similarity
Water – Run Smith-Watermann alignment algorithm to compare control and 3' UTR similarity
Add results to table - only consider alignments at least as long as the shortest control sequence being compared
Add results to table - only consider alignments at least as long as the control sequence being compared
Table of dORF, 3' UTR and control similarity for each species
Table of dORF, 3' UTR and control similarity for each species

Figure 2.9: Workflow describing how Python scripts generate, and run conservation analysis on, different control sequences for human, and/or homolog, downstream open reading frames (dORFs). *Python scripts used the Smith-Waterman alignment algorithm to investigate the conservation, or similarity, of these control sequences to compare to the dORFs and 3' untranslated regions (UTRs). The workflow highlights the key tools, or commands, and a brief description of their functions.*

2.9.3 Ontology analysis of the Highly Conserved dORF shortlist (HC dORFs)

Several web-based gene ontology analysis tools were used on the list of genes containing HC dORFs, a shortlist of highly conserved AUG dORFs (Highly conserved dORFs), searching for enriched ontologies. The DAVID functional annotation tool was used by uploading the RefSeq RNA IDs of the HC dORFs and then running functional annotation with all possible results values included in the outputs and other default settings (Sherman *et al.*, 2022). The g:Profiler g:GOSt functional profiling tool was also used with the gene list with default settings except advanced settings set to include all results and the significance threshold set to be Bonferroni correction (Raudvere *et al.*, 2019). The final tool was the Gene Ontology Enrichment analysis and visualization tool (GOrilla) run with the same gene list and the ontology set as function and other default settings (Eden *et al.*, 2009).

2.10 Influence of different cell types, disease states and conditions on RP alignments to dORF shortlists and associated 3' UTRs and Genes

This analysis investigated ribosomal association with the two groups of shortlisted dORFs, MSVW dORFs (MS validated Wu dORFs) and HC dORFs (Highly conserved dORFs) and the 3' UTRs they appear in, relative to the ribosomal association with the gene, in various RP datasets with different cell, or tissue, types and cellular conditions or disease states.

2.10.1 Datasets and sequences used in analysis

The publicly available SRA RP datasets used in this analysis were listed in section 2.2 with a brief description of the cell, or tissue, type and the disease state, treatment, or modification.

Two groups of dORFs were used in this analysis, the MSVW dORFs (section 2.6), and the HC dORFs. The genomic location of these dORFs was obtained using Integrative Genomics Viewer (IGV) (version - 2.9.4) with the human hg38 genome loaded (Robinson *et al.*, 2011). Within IGV the BLAT tool was used to align each dORF sequence, reporting the genomic location of the dORFs (Kent, 2002). The same method was used to get the genomic locations of the 3’ UTR sequences containing the dORFs. The genomic location of the dORF-containing genes were obtained using the NCBI Gene database (Sayers *et al.*, 2022). In addition, three housekeeping genes were included as a control which should have relatively consistent ribosomal association across datasets, very little 3’ UTR ribosomal association, and none of the above dORFs present. The housekeeping genes were *GAPDH*, *ACTB* and *TUBB*. Their gene locations were taken from the NCBI Gene database MANE select transcript, and the CDS location of MANE transcripts was used to determine the 3’ UTR sequences (Sayers *et al.*, 2022). The same method described previously was used to find the 3’ UTR genomic location.

The way the tools in the Python scripts described in this section determine and report where RP reads aligned is based on whether a single base of the read aligns within the genomic location defined. To have confidence that the translated codon of the RP read is present within the defined genomic region, the genomic locations described above were modified. Removing the first and last 20 bases from the genomic location meant that even if the first or last base of an RP read was found to be overlapping the genomic region defined it would be highly likely that the translated codon would fall within the original genomic location. This was done for all genomic locations described previously except where the dORF length was 40 bases or shorter as this would leave no region to look for alignments. Instead for these dORFs the first and last 14 bases were removed to allow dORFs of 30 bases to be aligned to.

2.10.2 Python Scripts for RP dataset alignment analysis

This analysis gathered and processed all of the RP datasets, trimming reads based on the quality, adapter content and lengths. Then these processed RP datasets were aligned against the human genome, before the alignments were used to determine and quantify how many RP reads in each dataset were aligned to the dORFs, 3’ UTRs and genes, described in the section 2.10.1, using the modified genomic locations provided (Figure 2.10). To do this analysis a set of Python scripts were used, such as, RP_Alignments_MSVW_dORFs_and_3UTRs_Part1.py, with the RP datasets split into eleven parts to decrease the duration of the Python scripts (Part one example provided). This was run using the modified genomic locations for the housekeeping genes and the MSVW dORF shortlist. Another set of Python scripts, such as RP_Alignments_HC_dORFs_and_3UTRs_Part1.py, ran the same analysis using the modified genomic locations for the HC dORF shortlist instead.

These scripts made used of Python commands and bioinformatic tools, some of which have been used previously on the Galaxy platform. The human grch38 genome was downloaded from ensembl for use with these scripts (Harrison *et al.*, 2024). A summary of the steps and tools involved with these Python scripts are provided in Figure 2.10. A list of the RP datasets was imported from the SRA using the fasterq-dump tool and the dataset accessions (version - 3.1.0) (Leinonen, Sugawara and Shumway, 2011). An initial quality control step was carried out using the FastQC tool (version - 0.12.1) with default settings on the datasets (Andrews, 2010). Then the RP dataset reads were trimmed using the Trim Galore! tool (version – 0.6.10) initially run to remove illumina adapters with overlap for adapter set as 5, and trim reads based on a minimum quality score of 30, minimum length of 25 bases, and remove ‘N’ bases from the start or end of reads (Krueger, 2021). This initial trim was then repeated with the same settings but to remove the nextera adapters. A third trim used the same settings but removed the small RNA adapter and also filtered out any reads still containing any ‘N’ bases. The final trimming step trimmed all remaining reads to a maximum length of 35 bases (Figure 2.10). Then the FastQC tool was used again with default settings, as another quality control check (Andrews, 2010). The HISAT2 tool (version – 2.2.1) was used to align the RP reads to the human genome, the settings for this tool were to ensure that any mismatches or ambiguous bases in the alignment were not permitted (Kim *et al.*, 2019). The output BAM alignment file was then run with both SAMtools (version – 1.19.2) sort and index tools to allow the SAMtools view tool to be used to filter the BAM alignment file to the modified genomic locations of the genes, 3’ UTRs or dORFs (Li *et al.*, 2009). These filtered alignment files were then used with RSeQC bam_stats tool (version – 2.6.4) to quantify the number of reads aligned in the alignment file in a text file (Wang, Wang and Li, 2012). The final step was to gather the number of RP reads aligned to each genomic location of interest and then add this data for each RP dataset to a results dataframe (Figure 2.10).

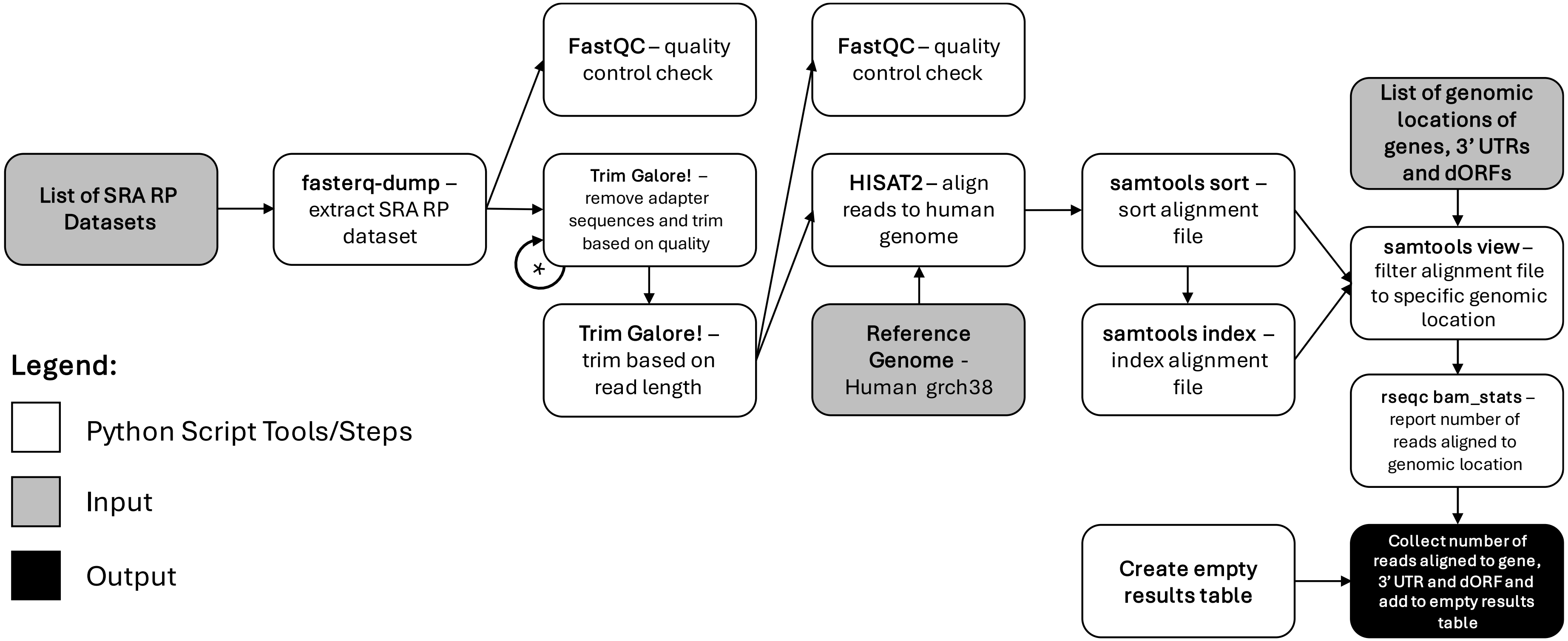
List of SRA RP Datasets
fasterq-dump – extract SRA RP dataset
FastQC – quality control check
Trim Galore! – remove adapter sequences and trim based on quality
*
Trim Galore! – trim based on read length
FastQC – quality control check
HISAT2 – align reads to human genome
Reference Genome - Human grch38
samtools sort – sort alignment file
samtools index – index alignment file
List of genomic locations of genes, 3' UTRs and dORFs
samtools view – filter alignment file to specific genomic location
rseqc bam_stats – report number of reads aligned to genomic location
Create empty results table
Collect number of reads aligned to gene, 3' UTR and dORF and add to empty results table
Legend:
Python Script Tools/Steps
Input
Output

Figure 2.10: Workflow describing how Python scripts identify the number of ribosome profiling (RP) reads, across a list of RP datasets, that are aligned to a list of genomic locations for genes, 3' untranslated regions (UTRs) and downstream open reading frames (dORFs). *The workflow highlights the key tools, or commands, and a brief description of their functions. * The Trim Galore! tool is run three times in succession to ensure different adapter sequences are removed. The RP datasets are extracted, processed and aligned against the human genome, then these alignments are filtered for specified genomic locations and the reads aligned to these were reported.*

2.10.3 Processing data generated by Python scripts

The brief summary of the FastQC quality control data for each RP dataset before and following the trimming steps described previously are provided alongside the overall human genome alignment for each dataset from the HISAT2 tool output, in Supplementary Table 1 available at https://github.com/joetomlinson/Joe_Tomlinson_Thesis_Supplementary_Table.

The results dataframes with RP read alignments were opened and the overall alignment percentage of each RP dataset to the human genome was included alongside the RP reads aligning to the genomic regions of interest. For each RP dataset the overall project SRA accession and treatment group or cell, or tissue, type was assigned. Datasets were excluded from further analysis if, across the datasets for a project, the overall alignment percentage to the human genome was consistently below 50%. To allow comparisons between RP datasets and to look at 3' UTR and dORF enrichment of RP read alignment compared to the gene, the number of RP reads aligned to the 3' UTR or dORF per 1000 RP reads aligned to the gene was calculated. In addition, the same calculation was done for the 3' UTRs alone, with the number of RP reads aligned to the dORF within the 3' UTR removed. This analysis investigated whether the preference for 3' UTR or dORF aligned RP reads changed with different conditions, rather than only considering total RP reads in the 3' UTR or dORFs in different datasets.

2.11 Statistical Analysis

All statistical analysis was carried out using GraphPad Prism (version 10.2.3). Various data types can be analysed using a range of statistical tests in GraphPad Prism (version 10.2.3). All data to be analysed was uploaded into GraphPad Prism (version 10.2.3) and appropriate statistical tests were carried out. Where applicable, parametric, rather than nonparametric, tests were used. Parametric tests assume a normal distribution of data, and this was supported by the data summaries provided by GraphPad Prism (version 10.2.3) when analysing the data. To reduce the increase in Type I errors, false positives, caused by carrying out multiple statistical comparisons, the Holm-Šídák method was used to adjust P values. The Holm-Šídák method was selected because this is an effective method commonly used to adjust P values that has a reduced chance of increasing Type II Errors, false negatives, compared to other adjustment methods. Selection of statistical tests was based on the use of these tests in the literature, the data being analysed, and the recommendations and details from within GraphPad Prism (version 10.2.3).

To make comparisons between more than two means the One-Way ANOVA multiple comparisons test with Šídák adjusted P values was used (Sections 3.2, 3.5, 3.7, 4.3, 4.4, 5.2, 5.3, 5.5 and Appendix 3). In instances where only two means were compared, an unpaired t test was used and the subsequent P values were then adjusted for multiple comparisons using the Holm-Šídák method (Section 3.6). Statistical analysis of the correlations used the Pearson correlation coefficient (Sections 3.3, 3.7, 4.3). The data displayed in the heat maps in section 4.4 were analysed using Chi-square goodness of fit tests. The Chi-square goodness of fit test was used when comparing the differences between expected and observed distributions.

Chapter 3: Translational Regulation by dORFs

3.1 Introduction

Understanding translational regulators is important due to their role in healthy and disease states. Downstream open reading frames (dORFs) are a recently proposed translational regulator, suggested to function through their translation in the 3' UTR (Wu *et al.*, 2020b). Bioinformatic approaches exploring translational regulation and dORFs have the potential to improve understanding of human health and diseases, which could support the development of treatments. Much like uORFs (see section 1.6.3), evidence of dORF translation could be found through 3' UTR ribosomal association (Ingolia *et al.*, 2009; Guydosh and Green, 2014; Ji *et al.*, 2015; Miettinen and Björklund, 2015; Young *et al.*, 2015; Hsu *et al.*, 2016). Oncogenesis is associated with dysfunction of translational regulators, such as uORFs (Young and Wek, 2016; Sendoel *et al.*, 2017), but also 3' UTR regulators, where altered 3' UTR processing in cancer can influence the activity of these regulators (Mayr and Bartel, 2009; Singh *et al.*, 2009). 3' UTR ribosomal association can be influenced by cellular conditions (Ingolia *et al.*, 2009), with implications for cancer cells. dORFs could represent a new target for cancer therapy. The hypothesis for this chapter is that ribosomes will associate with 3' UTRs and dORFs, with dORF-containing transcripts having increased association. In cancer it is anticipated that there may be changes in both ribosomal association with dORFs and the 3' UTR, and dORF activity.

This chapter contains the results generated through bioinformatic analysis of the dORFs described in Wu, Q. *et al.* (2020), referred to throughout this thesis as 'Wu dORFs'. The methods described in section 2.4 were used to run analysis of translational regulation on transcripts with and without Wu dORFs using paired RNAseq and ribosome profiling (RP) datasets. This led to examination of correlations between translational regulation or ribosomal presence in the coding sequence (CDS) with ribosomal presence in the 3' UTR, investigating the influence 3' UTR ribosomal association has on that of the CDS and the translational regulation of transcripts. The translational regulation analysis was used in combination with analysis of the tumour transcriptome due to its potential impact on Wu dORFs.

Analysis then turned to ribosomal presence and RP read density in different transcript regions, including Wu dORFs, through the methods described in section 2.5. The aim of this analysis was to investigate dORF translation and differences in ribosomal association in healthy and tumour tissue. Finally, methods from section 2.6 were used to identify translated Wu dORFs and gather data from previous analyses to generate a shortlist of Wu dORFs for further analysis.

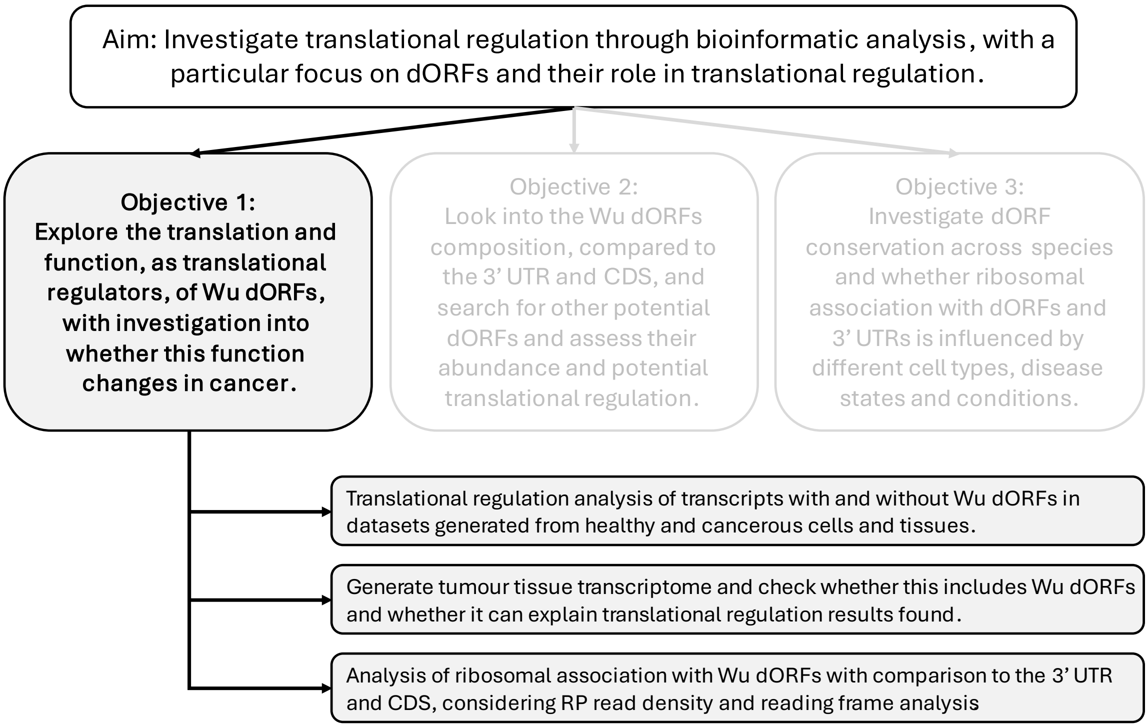
Aim: Investigate translational regulation through bioinformatic analysis, with a particular focus on dORFs and their role in translational regulation.
Objective 1: Explore the translation and function, as translational regulators, of Wu dORFs, with investigation into whether this function changes in cancer.
Objective 2: Look into the Wu dORFs composition, compared to the 3' UTR and CDS, and search for other potential dORFs and assess their abundance and potential translational regulation.
Objective 3: Investigate dORF conservation across species and whether ribosomal association with dORFs and 3' UTRs is influenced by different cell types, disease states and conditions.
Translational regulation analysis of transcripts with and without Wu dORFs in datasets generated from healthy and cancerous cells and tissues.
Generate tumour tissue transcriptome and check whether this includes Wu dORFs and whether it can explain translational regulation results found.
Analysis of ribosomal association with Wu dORFs with comparison to the 3' UTR and CDS, considering RP read density and reading frame analysis

Figure 3.1: Summary of approaches used in section 3 to meet the chapter objectives and overall study aim. *dORF – downstream open reading frame, 3' UTR – 3' untranslated region, CDS – coding sequence, RP – ribosome profiling.*

3.2 In healthy kidney tissue, transcripts containing Wu dORFs are translationally upregulated

Initially, the translational regulation of Wu dORFs was investigated in healthy human kidney and kidney cancer datasets to validate dORF activity and to understand if dORF activity may change in cancer; there is currently no reported evidence of dORF function in cancer. The different datasets generated from healthy kidney and kidney tumour samples are referred to as 'NT' for healthy tissue and 'TT' for tumour tissue and the different decimals (e.g NT1.1 or NT1.2) refer to which RP dataset was used, as discussed in section 2.2. The data presented used the differential expression of a transcript, or transcript CDS, from paired RNAseq and RP datasets to model the potential translational regulation, using methods described in section 2.4.2. In healthy kidney, transcripts that contain Wu dORFs show increased relative translation compared with those lacking Wu dORFs (Figure 3.2). This increase is seen in a change in the mean differential expression of transcripts with and without Wu dORFs. The mean in transcripts with Wu dORFs is increased by 1.380, (95% confidence interval for difference 1.653 to 1.107) 1.436 (1.705 to 1.167), 1.605 (1.880 to 1.330), 1.629 (1.901 to 1.357) compared to transcripts without. The shifted distribution shown by the box and whisker plots also supports these changes (Figure 3.2). All of these differences in means are statistically significant, meaning the P value reported was less than 0.05, with all P values reported to be less than 0.0001 when running a One-Way ANOVA multiple comparisons test with Šídák adjusted P values. There is a large distribution of results, and some transcripts, regardless of dORF presence, had large positive or negative differential expression values (Figure 3.2). The transcript groups are different in sizes, with transcripts without Wu dORFs around ten times more prevalent. This analysis excludes transcripts with no expression in either RNAseq or RP datasets to avoid infinite Log2 fold changes. There is also no restriction on low expression in either RNAseq or RP datasets, meaning some of the transcripts may be poorly expressed and could lack biological relevance.

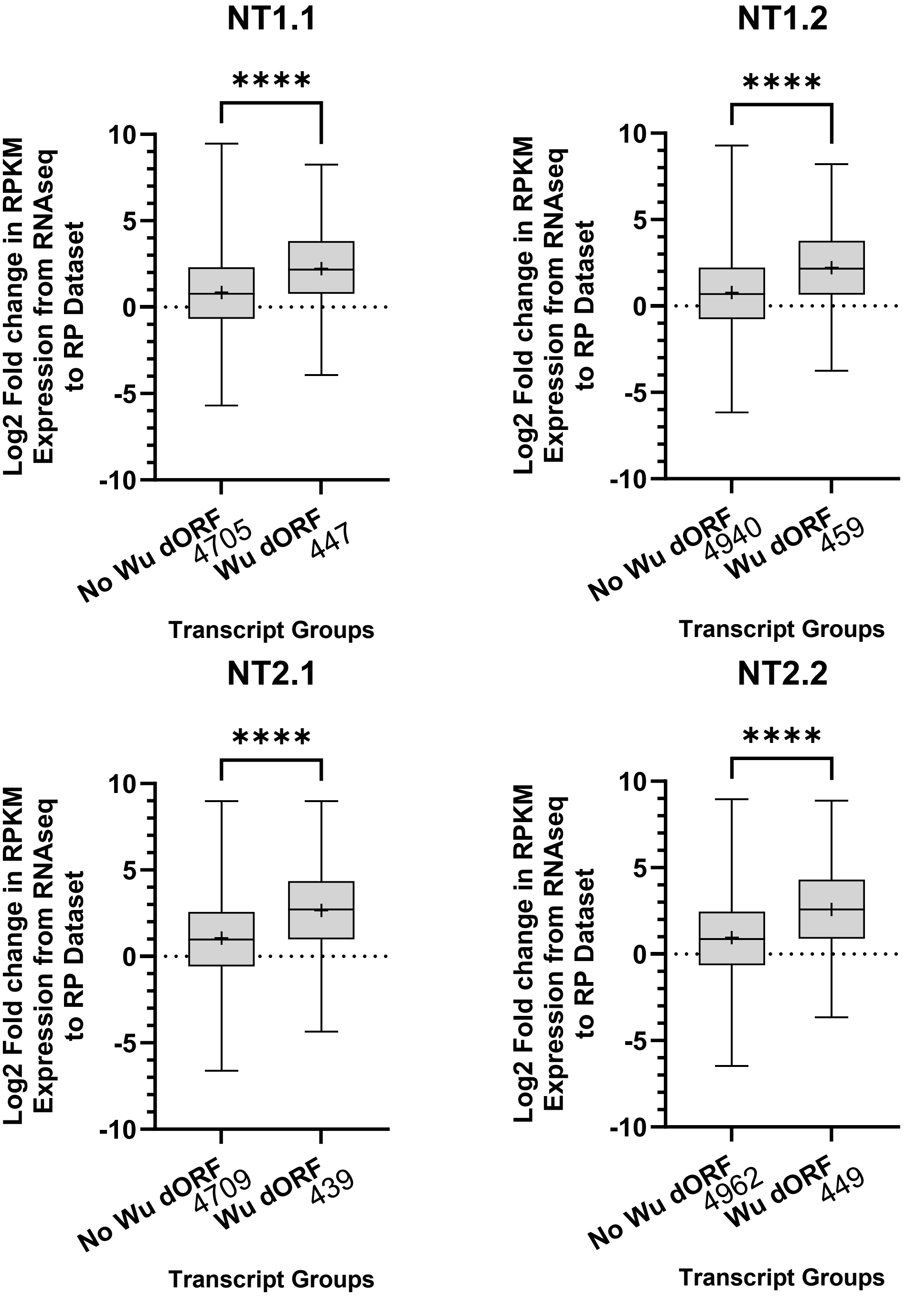

Figure 3.2: RefSeq transcripts with Wu downstream open reading frames (dORFs) are translationally upregulated in healthy kidney samples (NT) from PRJNA256316. *Relative transcript translation calculated as Log2 fold change in expression from RNAseq to RP dataset for each transcript. Box and whisker plots for each group of transcripts, the horizontal line within the box marks the median, the whiskers denote the maximum and minimum values, and the box marks the 25th to the 75th percentile of the values distribution, with the mean marked with '+'. One-Way ANOVA multiple comparisons test with Šídák adjusted P values was used to compare the means. Each chart title describes which RP dataset was used in the analysis. Transcript group sizes included in the x axis labels. **** - P<0.0001.*

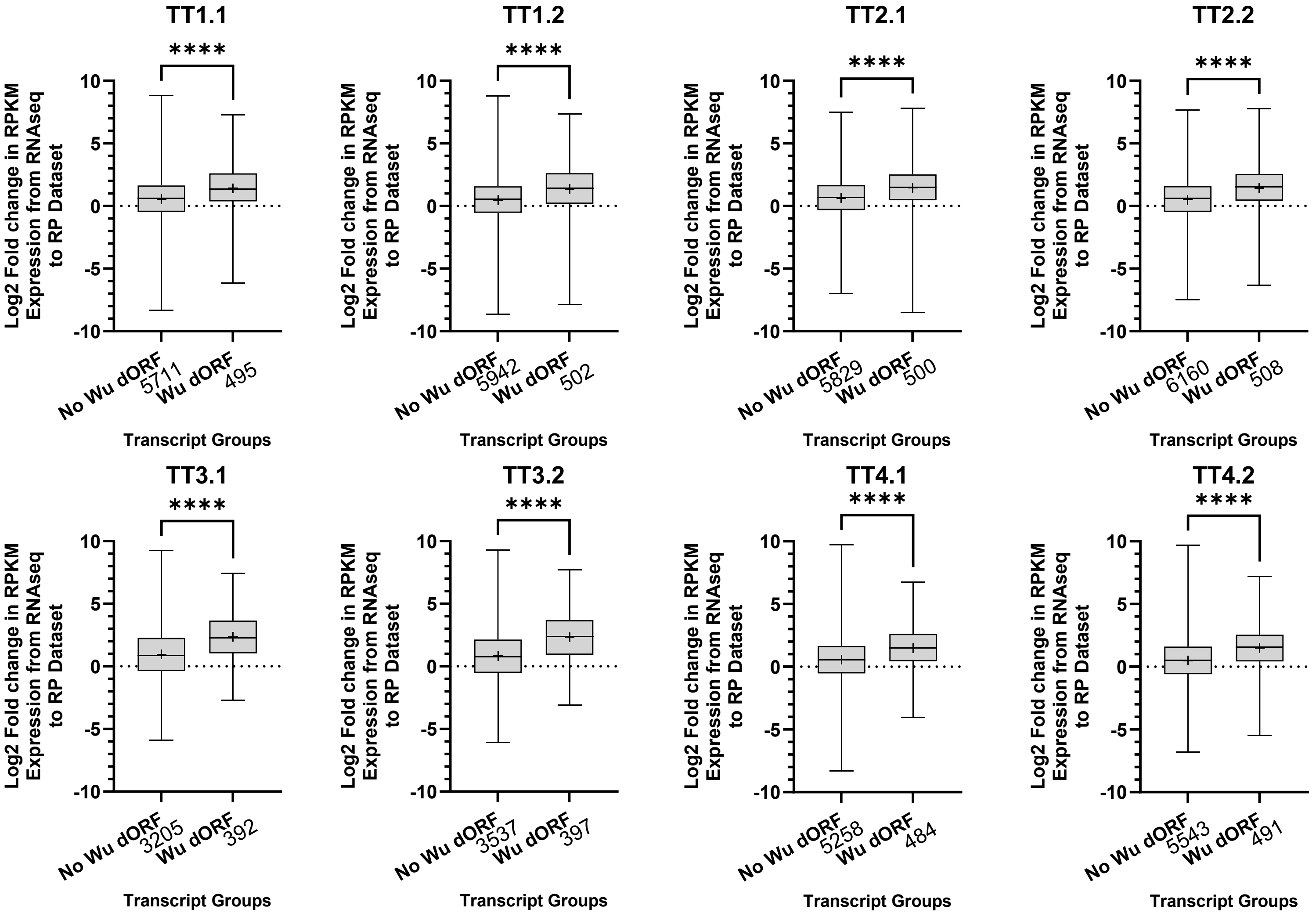

Figure 3.3: RefSeq mRNA ID transcripts with Wu downstream open reading frames (dORFs) appear to be translationally upregulated in kidney tumour samples (TT) from PRJNA256316. *Relative transcript translation calculated as Log2 fold change in expression from RNAseq to RP dataset for each transcript. Box and whisker plots for each group of transcripts, the horizontal line within the box marks the median, the whiskers denote the maximum and minimum values, and the box marks the 25th to the 75th percentile of the values distribution, with the mean marked with '+'. One-Way ANOVA multiple comparisons test with Šídák adjusted P values was used to compare the means. Each chart title describes which RP dataset was used in the analysis. Transcript group sizes included in the x axis labels. **** - P<0.0001.*

The same translational regulation analysis was performed on datasets from human kidney tumour samples, allowing comparison of Wu dORF function between healthy and tumour tissue to investigate changes in dORF activity. The tumour datasets show an increased relative translation in transcripts with Wu dORFs, compared to those without, with shifts in the box and whisker plots (Figure 3.3). The mean differential expression differences when comparing transcripts with, and without, Wu dORFs were 0.8610 (95% confidence interval for difference 1.119 to 0.6026), 0.8602 (1.116 to 0.6039), 0.8389 (1.096 to 0.5819), 0.9435 (1.198 to 0.6889), 1.397 (1.692 to 1.102), 1.511 (1.803 to 1.219), 0.9369 (1.199 to 0.6750), 1.007 (1.267 to 0.7478) all P<0.0001, but this increase is smaller than that seen in the healthy tissue, except in datasets generated from kidney tumour sample 3. The distribution of the box and whisker plots for the tumour tissue is reduced, condensing the plots, compared to the healthy tissue (Figures 3.2 and 3.3). There are also more transcripts considered in the kidney tumour dataset analysis, apart from kidney tumour sample 3 datasets. The data suggests potential translational upregulation of transcripts containing Wu dORFs in healthy tissue, much like the regulation described for dORFs (Wu *et al.*, 2020b). The potential function of Wu dORFs is consistent in the tumour tissue datasets, however, the activity is reduced. Questions are also raised over the tumour tissue datasets generated from tumour sample 3 which will be described in more detail in section 3.4.

3.3 Investigating the relationship between ribosomal presence in the 3' UTR and differential expression

The presence of Wu dORFs appears to affect differential expression of a transcript from RNAseq to RP datasets, implying translational regulation. The proposed function of dORFs relies on dORF translation (Wu *et al.*, 2020b). To look beyond dORFs specifically, the next section considers the impact of ribosomal presence in the 3' UTR on the relative translation of a transcript, using methods described in section 2.4.3. If dORFs are translated, ribosomes would be found in 3' UTRs (Wu *et al.*, 2020b). The hypothesis is that greatest translational upregulation occurs in mRNAs with the most active dORFs, and these are likely to be mRNAs with higher 3' UTR ribosome association. This analysis used the same human healthy and tumour kidney datasets. To maximise the number of transcripts considered in the correlation analysis there were no minimum RP expression parameters set, meaning that transcripts with low ribosome occupancy are included. It was observed that whether using healthy or tumour datasets there is a very small positive correlation between the Log2 fold change in expression from RNAseq to RP datasets and the Log2 3' UTR RP expression for transcripts (Figure 3.4). In the healthy datasets (NT) all Pearson r values, used to quantify the correlation, are statistically significant ($P<0.001$ and $P=0.0054$). In the tumour tissue (TT), half of the correlations are statistically significant ($P<0.05$) but the Pearson r values indicate there is likely no biologically meaningful correlation. The data points are very spread across all datasets (Figure 3.4). Even at increasing 3' UTR ribosome occupancy, there are datapoints suggesting potential translational downregulation of the transcript (Figure 3.4). The correlation for tumour tissue sample 3 datasets was more similar to the healthy datasets rather than the other tumour datasets, also seen in Figures 3.2 and 3.3, alongside considerably fewer points considered in the correlation (Figure 3.4), also seen in the translational regulation analysis previously (Figure 3.3). Fewer transcripts have 3' UTR expression in the tumour tissue sample 3 RP datasets. The very small correlation in the healthy tissue datasets (NT) reduced further in the tumour tissue (TT) datasets, except for human kidney tumour sample 3 datasets (Figure 3.4). Although there are more data points with most of the tumour tissue datasets, the data point distribution is tighter (Figure 3.4). These differences between the healthy and tumour tissue datasets are in addition to the differences observed with the previous translational regulation analysis. Overall, there is a small correlation between translational regulation and ribosomal association with the 3' UTR in healthy datasets, which appears to be lost in tumour datasets.

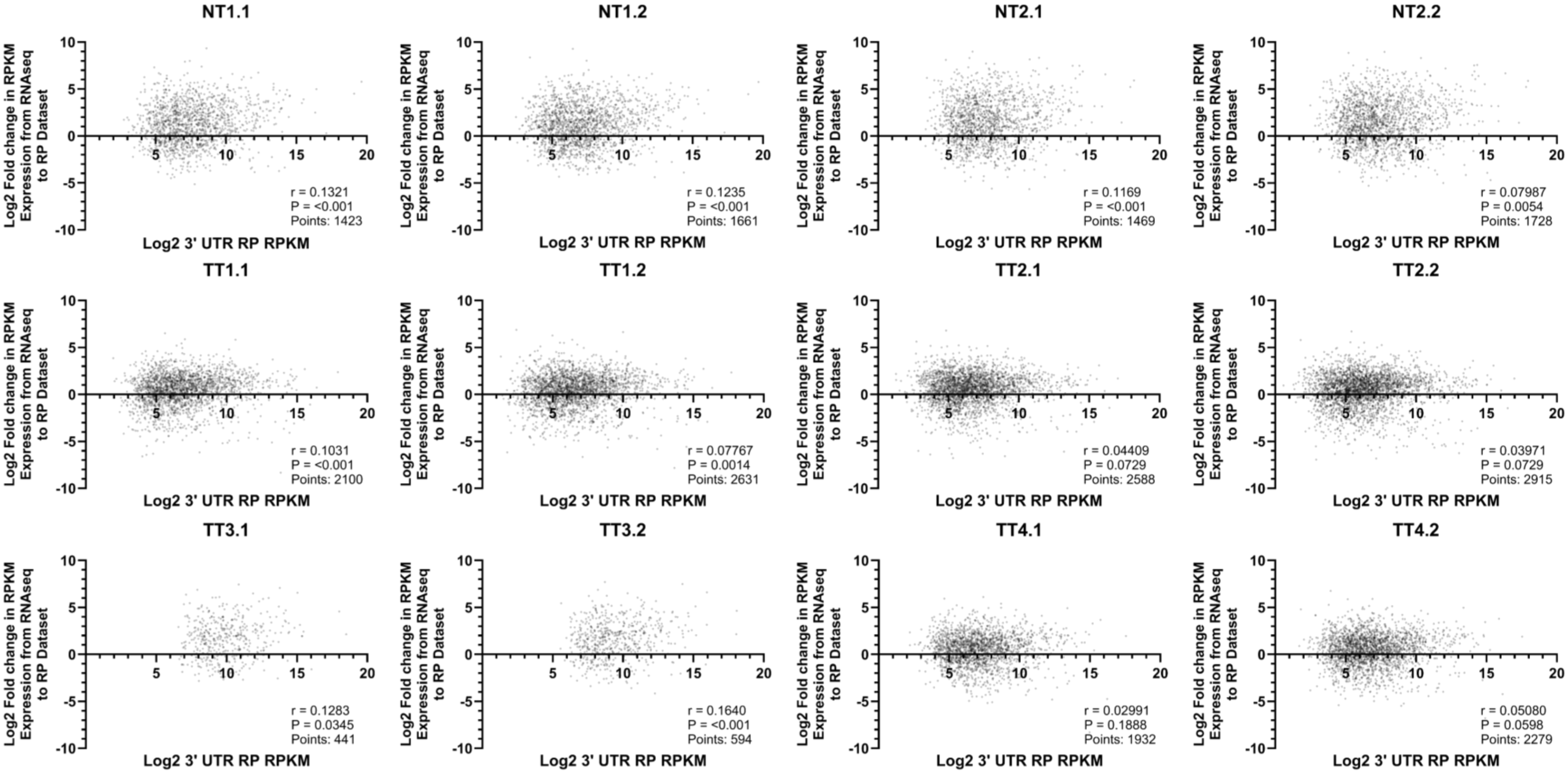

Figure 3.4: Correlation between translational regulation and 3' untranslated region (UTR) expression in ribosome profiling (RP) datasets for RefSeq mRNA ID transcripts in PRJNA256316 datasets. *Relative transcript translation calculated as Log2 fold change in expression from RNAseq to RP dataset for each transcript. 3' UTR expression was calculated as Log2 values. The correlation was quantified using the Pearson correlation coefficient. Each chart title describes which RP dataset was used in the analysis e.g. NT1.1. The Pearson r coefficient, P value and number of data points are included on the charts. P values adjusted for multiple comparisons using Holm-Šídák method.*

Following on from the previous analysis, the same correlation analysis was applied to transcripts containing Wu dORFs. The translational upregulation of these transcripts suggested in Figures 3.2 and 3.3 could imply that Wu dORFs are translated, meaning increased ribosomal association with those 3' UTRs would be expected. This analysis investigated whether transcripts with Wu dORFs that had increased 3' UTR ribosomal association have increased relative translation, supporting the hypothesised mechanism of the Wu dORFs (Wu *et al.*, 2020b). However, there is likely no biologically meaningful correlation between the potential translational regulation and ribosomal association in the 3' UTR for transcripts with Wu dORFs across the healthy and tumour RP datasets (Figure 3.5). This is indicated by the small Pearson r values with all correlations having no statistical significance (P>0.05) (Figure 3.5). In addition, few transcripts with Wu dORFs have RP reads aligned with around 200 transcripts included for each RP dataset, fewer for tumour sample 3 datasets (Figure 3.5). As seen in Figure 3.4, there is a spread of the datapoints. There is no biologically meaningful correlation between 3' UTR ribosomal presence and the translational regulation of transcripts containing Wu dORFs.

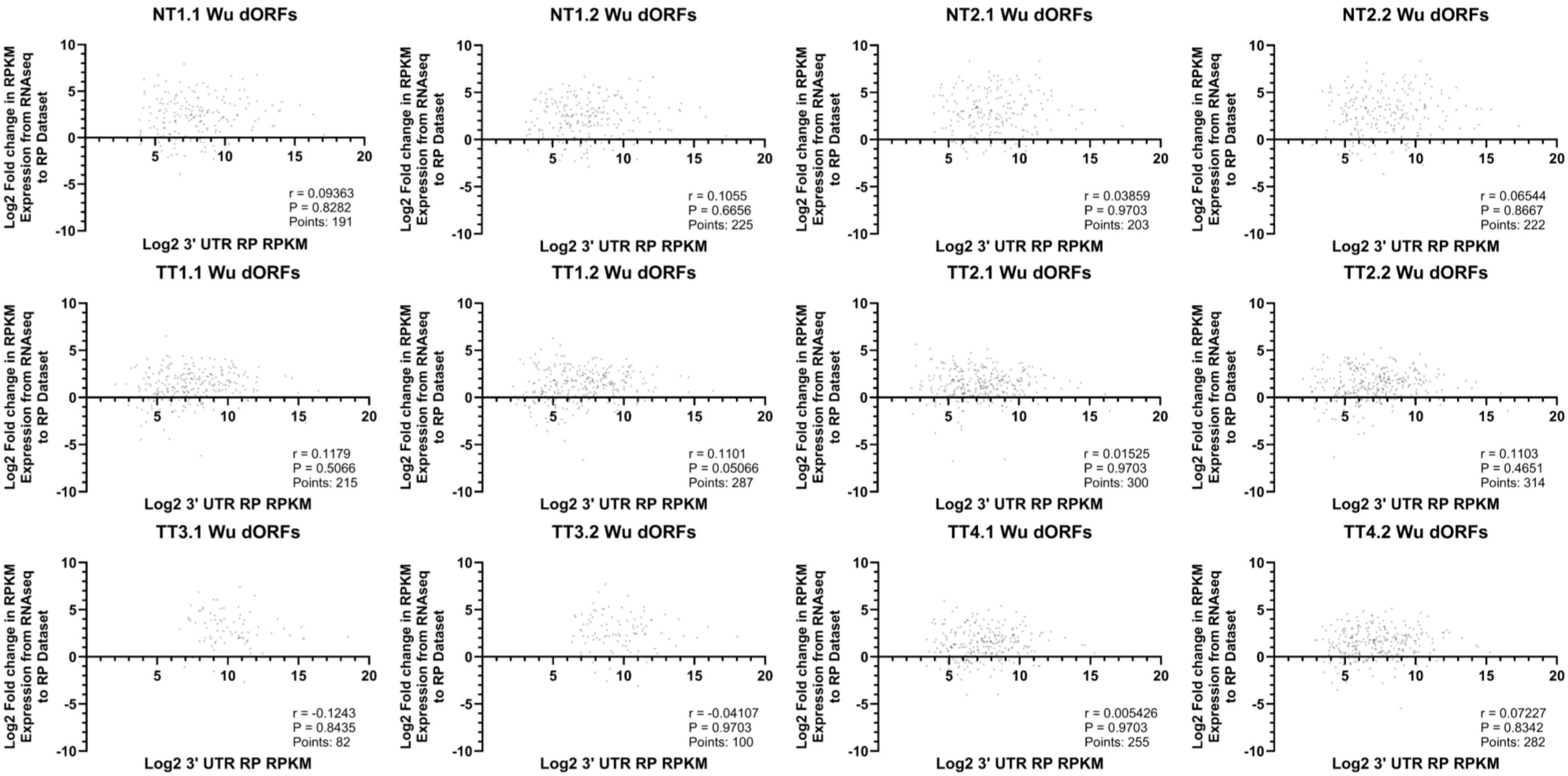

Figure 3.5: Correlation between translational regulation and 3' untranslated region (UTR) expression in ribosome profiling (RP) datasets for RefSeq mRNA ID transcripts containing Wu dORFs in PRJNA256316 datasets. *Relative transcript translation calculated as Log2 fold change in expression from RNAseq to RP dataset for each transcript. 3' UTR expression was calculated as Log2 values. The correlation was quantified using the Pearson correlation coefficient. Each chart title describes which RP dataset was used in the analysis e.g. NT1.1. The Pearson r coefficient, P value and number of data points are included on the charts. P values adjusted for multiple comparisons using Holm-Šídák method.*

The relationship between 3’ UTR and CDS expression in RP datasets, a potential model of their ribosomal association and translation, was investigated. As discussed previously, dORFs translation is suggested to increase CDS translation in a transcript (Wu *et al.*, 2020b), meaning increasing ribosomal presence in the 3’ UTR would be expected alongside increased CDS ribosomal presence. Investigating the correlation between 3’ UTR and CDS RP expression can help understand what happens to CDS ribosome occupancy as the 3’ UTR ribosome occupancy changes. The datasets from human healthy kidney (NT) and kidney tumour (TT) samples show a consistent positive correlation between CDS and 3’ UTR ribosome occupancy (Figure 3.6). The overall correlation is shown by the statistically significant (P<0.001) Pearson’s r values ranging from 0.4776 to 0.3971 across all datasets, indicating that with increasing CDS ribosome occupancy, 3’ UTR ribosome occupancy also increases (Figure 3.6). The distribution of data points shows variation with the spread (Figure 3.6), but to a reduced extent than in Figures 3.4 and 3.5. This variation highlights the range of different relationships between 3’ UTR and CDS ribosome occupancy for transcripts: some have reduced CDS ribosome occupancy with increased 3’ UTR ribosome occupancy. TT3.1 and TT3.2 charts in Figure 3.6 differ due to fewer data points. Between the healthy and tumour datasets, however, the number of transcripts with RP expression in the CDS and 3’ UTR is increased in most of the tumour tissue datasets, except those from tumour tissue sample 3 (Figure 3.6). Although the number of transcripts differed, the size and direction of the correlation was very similar in both healthy and tumour tissue samples. The correlation analysis indicates further potential problems with tumour tissue sample 3, and a relationship between 3’ UTR and CDS ribosome occupancy in transcripts.

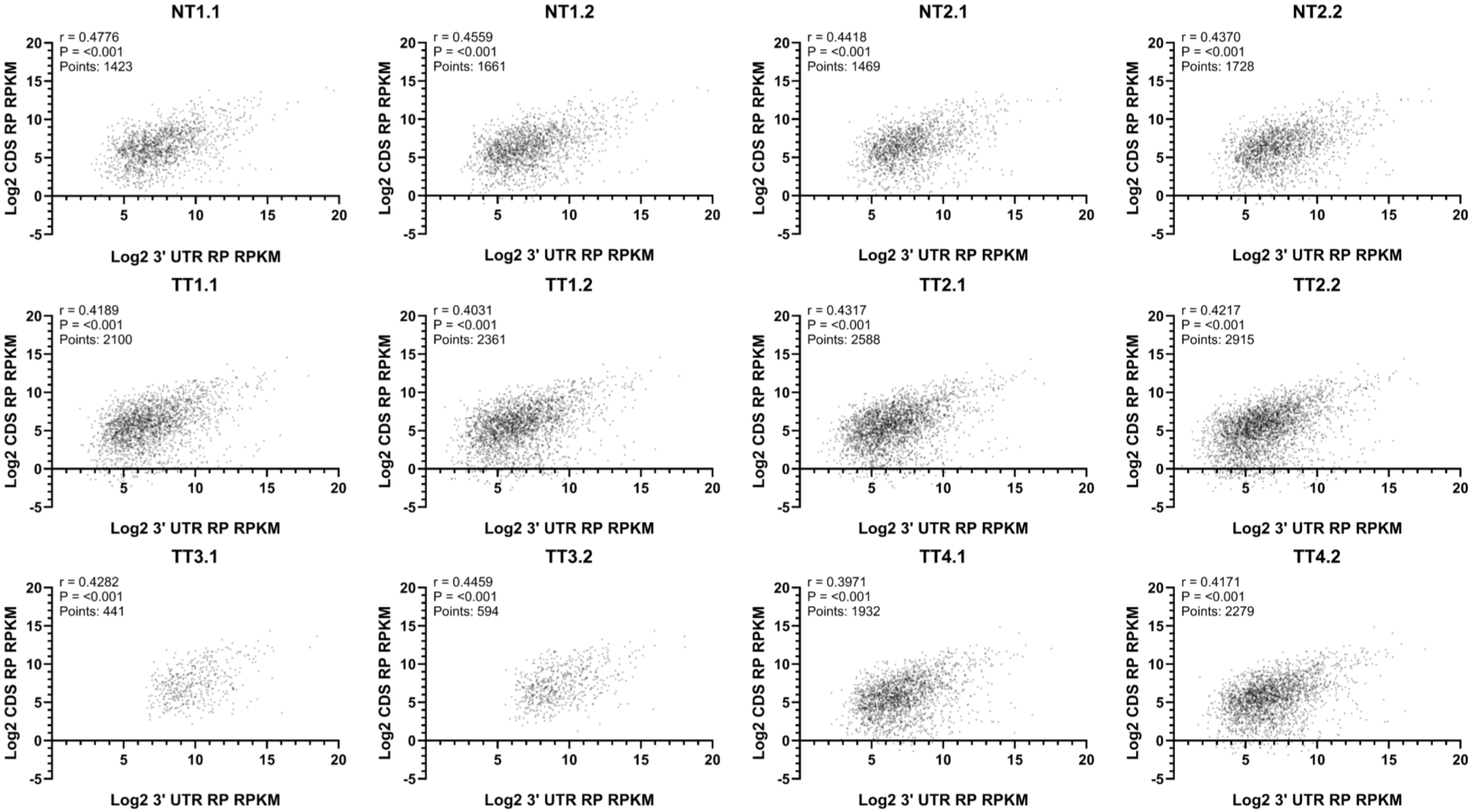

Figure 3.6: Correlation between coding sequence (CDS) and 3' untranslated region (UTR) expression in ribosome profiling (RP) datasets for RefSeq mRNA ID transcripts in PRJNA256316 datasets. *CDS and 3' UTR expression were calculated as Log2 values. The correlation was quantified using the Pearson correlation coefficient. Each chart title describes which RP dataset was used in the analysis e.g. NT1.1. The Pearson r coefficient, P value and number of data points are included on the charts. P values adjusted for multiple comparisons using Holm-Šídák method.*

To support the previous analysis, the same correlation analysis between CDS and 3' UTR ribosome occupancy was carried out on transcripts containing Wu dORFs. If these dORFs are translated, and they increase translation of the CDS, then it would be expected that the increased 3' UTR ribosomal association would correlate with increased CDS ribosomal association (Wu *et al.*, 2020b). Very similar results are seen to Figure 3.6, across all healthy and tumour datasets there is a consistent positive correlation between the CDS and 3' UTR ribosome occupancy (Figure 3.7). The distribution, and spread, of data points varies less compared to Figure 3.6, (Figure 3.7). The correlations for transcripts containing Wu dORFs is greater for all transcripts, shown by the statistically significant (P<0.001) Pearson's r values ranging from 0.6215 to 0.4908 across all datasets (Figure 3.7). This indicates that Wu dORF presence increases the correlation. These correlations indicate that transcripts with increased CDS ribosome occupancy also have increased 3' UTR ribosome occupancy (Figure 3.7). TT3.1 and TT3.2 charts in Figure 3.7 again have considerably fewer data points. This analysis, similar to Figure 3.5, again highlights that few of the Wu dORF-containing transcripts had data reported.

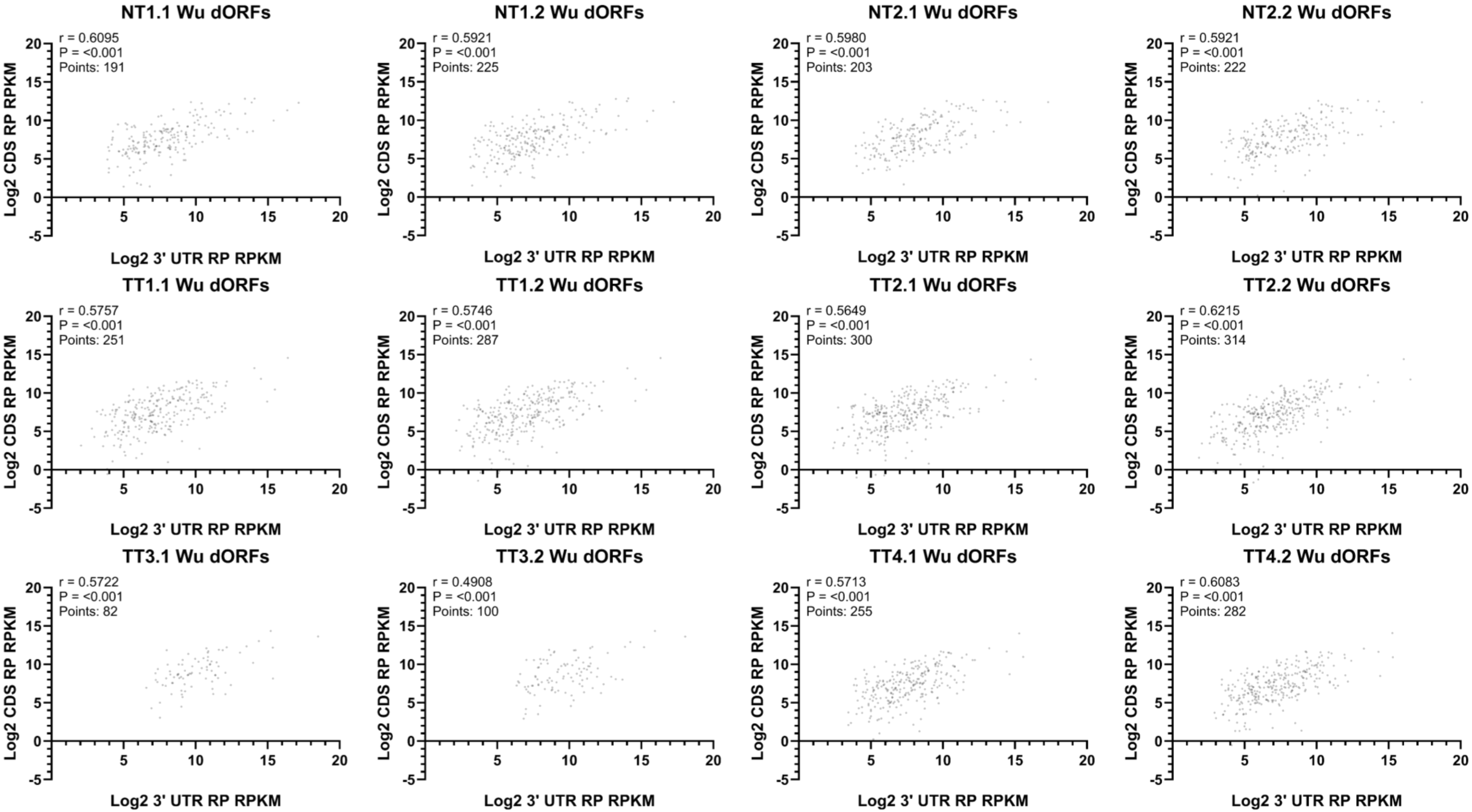

Figure 3.7: Correlation between coding sequence (CDS) and 3' untranslated region (UTR) expression in ribosome profiling (RP) datasets for RefSeq mRNA ID transcripts containing Wu dORFs in PRJNA256316 datasets. *CDS and 3' UTR expression were calculated as Log2 values. The correlation was quantified using the Pearson correlation coefficient. Each chart title describes which RP dataset was used in the analysis e.g. NT1.1. The Pearson r coefficient, P value and number of data points are included on the charts. P values adjusted for multiple comparisons using Holm-Šídák method.*

3.4 Human kidney tumour sample 3

Results from human kidney tumour sample 3 (RNAseq: TT3, RP: TT3.1, TT3.2) differ from other tumour samples. In particular, the number of transcripts included in the differential expression and correlation results (Figures 3.3-3.7). In the differential expression analysis, the results from human kidney tumour sample 3 appear more similar to the healthy sample datasets (Figures 3.2 and 3.3). The MultiQC tool (Ewels *et al.*, 2016) pulled together the statistical and quality reports generated by FASTQC (Andrews, 2010) for comparison in Table 3.1. In addition, the number of RP (CDS) or RNAseq (Transcript) reads aligned were reported for the processed datasets, part of the 2.4.1 methods. This analysis aimed to explore these tumour tissue sample 3 differences. The human kidney tumour sample 3 datasets are generally smaller compared to the other datasets, a difference that increases when duplicates are removed (Table 3.1). The RP datasets from tumour tissue sample 3 have between 10-15% more duplicated reads than the other datasets, whereas the RNAseq dataset has a lower proportion of duplicate reads (Table 3.1). The proportion of the total reads made up of over-represented sequences is much larger for the RP and RNAseq datasets generates from tumour tissue sample 3 (Table 3.1). When aligned against the human non redundant database using the NCBI BLASTn tool (Altschul *et al.*, 1997; Camacho *et al.*, 2009; Cock *et al.*, 2015) the top over-represented sequences in TT3.1 and TT3.2 RP datasets were found to align to 28S ribosomal subunit ribosomal RNA (rRNA). This could indicate an issue with the sample preparation. The datasets also have considerably fewer reads aligned to the CDS or transcript, not explained by differences in the dataset sizes (Table 3.1). In the RNAseq datasets there are fewer than half as many RP reads aligned to the transcript (Table 3.1), and in the RP datasets this difference ranges from six to twenty times smaller (Table 3.1). This difference would have an impact on the number of results and differential expression data generated and could be an explanation for the differences seen. These dataset differences and possible confounding issues led to the exclusion of this dataset in further analyses.

Table 3.1: Comparing datasets generated from human kidney tumour sample 3 with other datasets generated from the other kidney tumour samples from PRJNA256316. *This table includes the ribosome profiling (RP) and RNAseq datasets with those generated from kidney tumour sample 3 highlighted in bold. The data in all but the final column are from the quality control checks on the unprocessed datasets. The final column includes the number of reads aligned to the coding sequence (CDS), for RP datasets, and the whole RefSeq mRNA ID transcripts for RNAseq datasets.*

	TT Datasets	**Total Reads**	**Unique Reads**	**Duplicate Reads (%)**	**Top over-represented read (%)**	**Sum of remaining over-represented sequences (%)**	**Processed reads aligned to CDS (RP) or transcript (RNAseq)**
Ribosome Profiling	SRR1528690	22206265	4948005	77.71798	3.787661	29.66966	886446
	SRR1528691	31147450	6374228	79.53531	3.740817	29.39058	1261188
	SRR1528692	25657299	5426717	78.84923	3.662689	33.14905	947381
	SRR1528693	36181563	7059984	80.48734	3.632444	33.02474	1349268
	SRR1528694	**14984509**	**616793**	**95.8838**	**17.34048**	**59.42281**	**48071**
	SRR1528695	**20492315**	**763279**	**96.27529**	**17.6626**	**59.24783**	**68755**
	SRR1528696	20302267	3037963	85.03634	6.769392	42.58618	431777
	SRR1528697	28365554	3912558	86.20666	6.770599	42.42306	610820
RNAseq	SRR2064426	31392714	20425623	34.93515	0.675803	1.085988	3450583
	SRR2064427	27246329	17780835	34.74044	0.613198	0.874591	3103017
	SRR2064428	**18512001**	**13180868**	**28.79826**	**0.3963**	**2.032617**	**735572**
	SRR2064429	25087358	16647061	33.64363	0.49515	0.903607	2381260

3.5 Exploring the impact of Wu dORFs altered in the tumour transcriptome generated from PRJNA256316

The previous findings indicate that the activity of Wu dORFs is reduced in tumour tissue compared to healthy tissue (Figures 3.2 and 3.3). In cancers, increased mutation rates mean that a wide range of 3' UTRs can be altered and often truncated (Carninci *et al.*, 2006; Mayr and Bartel, 2009; Singh *et al.*, 2009). These 3' UTR changes could disrupt, or remove, dORF sequences, preventing function, explaining the differences observed for tumour datasets. To explore this, the differential expression results generated previously were compared in transcripts with Wu dORFs that were fully conserved, or not, in the tumour transcriptome generated from the tumour sample RNAseq datasets. The tumour tissue transcriptome was generated using the Trinity tool (Grabherr *et al.*, 2011) with default settings, on the processed TT1, TT2 and TT4 RNAseq datasets from the RNAseq Galaxy workflow described in section 2.4.1. Wu dORFs were aligned against these Trinity assembled the NCBI BLAST+ blastn tool (Altschul *et al.*, 1997; Camacho *et al.*, 2009; Cock *et al.*, 2015). If altered 3' UTR processing caused the observed differences, the relative translation of transcripts would be reduced when the Wu dORFs were not conserved, compared to those that were, in the tumour transcriptome. Across the datasets there are between 143 and 158 transcripts with conserved Wu dORFs and around twice as many, 301-350, where the Wu dORF is not conserved. There were no statistically significant (P>0.05) differences between the mean differential expression values of transcripts with or without conserved Wu dORFs (Figure 3.8). There is very little differences in the means in healthy datasets, ranging from 0.04171 to 0.06781 (Figure 3.8). Although statistically not significant, the mean for transcripts with conserved Wu dORFs is slightly reduced compared to the transcripts without conserved Wu dORFs in all the tumour datasets, with the size of this difference ranging from 0.2550 to 0.4105 (Figure 3.8).

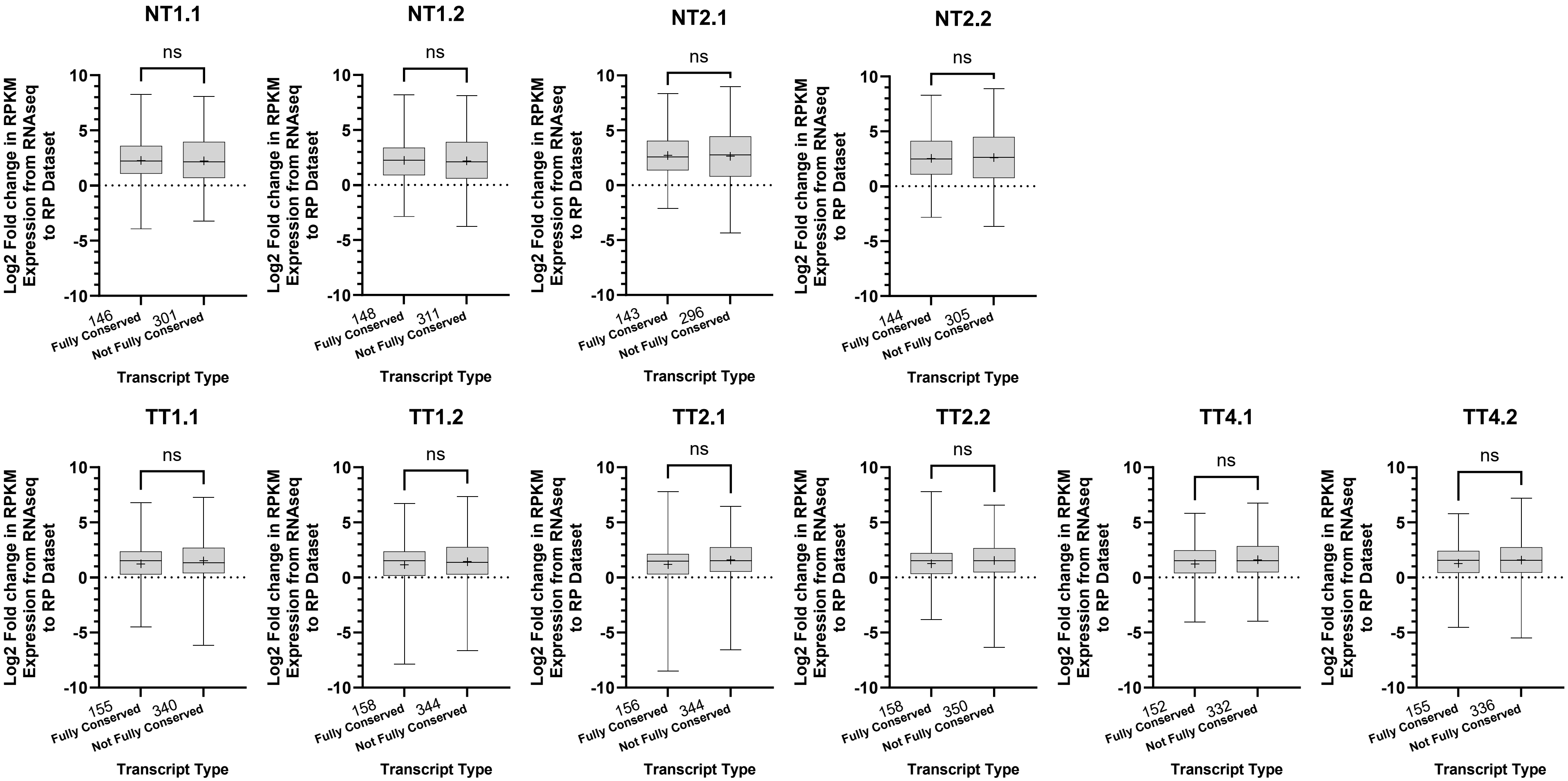

Figure 3.8: Translational regulation of RefSeq mRNA ID transcripts containing Wu downstream open reading frames (dORFs) are unaffected by whether dORFs are fully conserved in the tumour transcriptome. *Tumour transcriptome generated with kidney tumour RNAseq datasets from PRJNA256316. Relative transcript translation calculated as Log2 fold change in expression from RNAseq to RP dataset for each transcript. Transcript Type refers to whether the transcript contains a Wu dORF that was fully conserved or not in the tumour tissue transcriptome. Box and whisker plots for each group of transcripts, the horizontal line within the box marks the median, the whiskers denote the maximum and minimum values, and the box marks the 25th to the 75th percentile of the values distribution, with the mean marked with '+'. One-Way ANOVA multiple comparisons test with Šídák adjusted P values was used to compare the means. Each chart title describes which RP dataset was used in the analysis. Transcript group sizes included in the x axis labels. ns – P>0.05.*

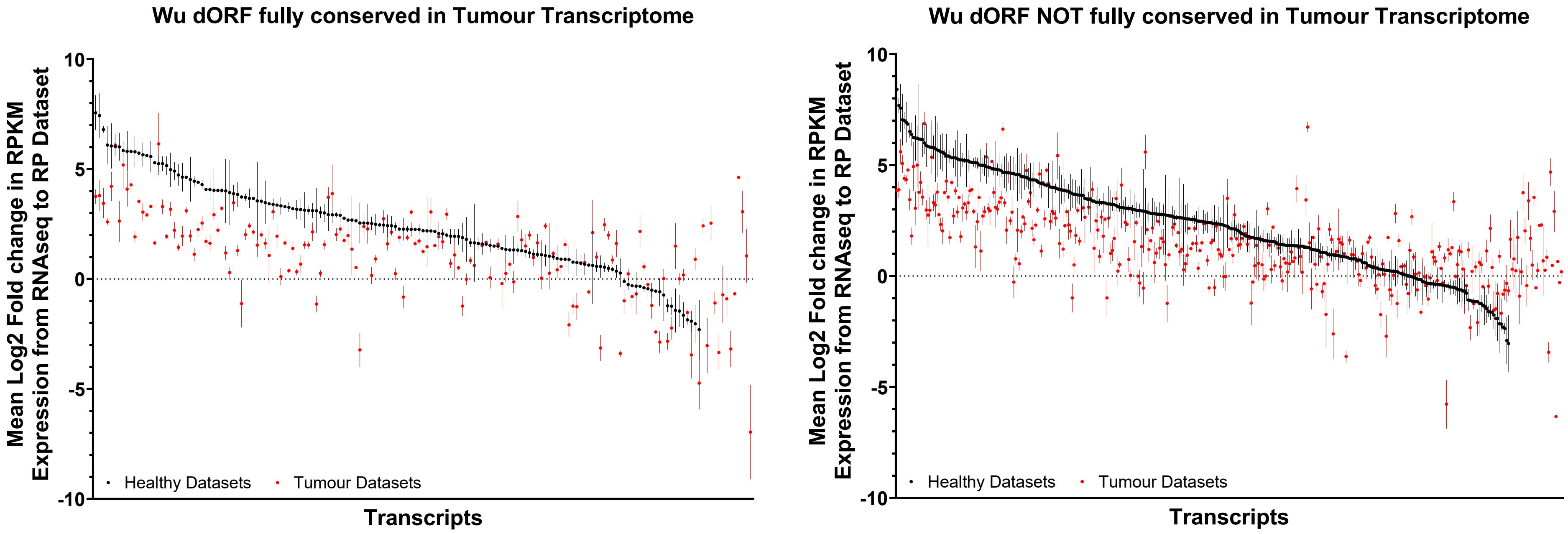

Figure 3.9: Changes in possible translational regulation between healthy and kidney tumour datasets from PRJNA256316 of RefSeq mRNA ID transcripts containing Wu downstream open reading frames (dORFs) are similar whether dORFs are fully conserved in the tumour transcriptome, or not. *Relative transcript translation calculated as Log2 fold change in expression from RNAseq to RP dataset for each transcript. Tumour transcriptome generated from Kidney tumour sample RNAseq datasets from PRJNA 256316. The mean Log2 fold change in expression from RNAseq to ribosome profiling (RP) dataset for each transcript was plotted with bars representing the standard deviation, with the mean for the healthy kidney datasets in black and the kidney tumour datasets in red. The data was plotted in order of the healthy dataset mean from high to low. The chart on the left contains transcripts with Wu dORFs which are fully conserved in the tumour transcriptome, and on the right are those which are not. Number of transcripts in charts: Left - 167, Right - 365.*

When comparing the mean differential expression values in the healthy and tumour tissue datasets for each transcript, whether the Wu dORF is conserved or not in the tumour transcriptome, most transcripts have a reduced differential expression mean in the tumour dataset (Figure 3.9). Although there are more transcripts without conserved Wu dORFs, the patterns in Figure 3.9 are similar. Figure 3.9 also shows the variation for these transcripts, even within datasets from the healthy or tumour samples, indicated by the large standard deviation (SD) bars, especially when considering the Log2 scale. Although it could explain the change in Wu dORF activity for a few transcripts with Wu dORFs, 3' UTR changes in the tumour transcriptome does not explain the potentially reduced Wu dORF activity in tumour datasets.

3.6 Investigating Ribosomal Association with CDSs, 3' UTRs and Wu dORFs using RP datasets generated from human healthy and tumour kidney samples

This section focuses on the ribosome profiling (RP) datasets used so far to look at the ribosomal association with different transcript regions and Wu dORFs, presented by Wu *et al.* (2020b), using the methods described in section 2.5. To compare with a translated region, the same analysis looking at RP read alignment to 3' UTRs, was done with CDSs with RefSeq mRNA IDs (Harrison *et al.*, 2024). The steps were the same as for the 3' UTR, however instead of discarding RP reads also aligning to 5' UTRs and CDSs, CDS aligned RP reads aligning to the 5' UTRs or 3' UTRs were discarded. This resulted in RP reads fully aligned to the CDS with confidence that those RP reads were not aligned to other regions. The correlation data in Figures 3.4-3.7 show that there are RP read alignments to the 3' UTR. This supports the literature which suggest the 3' UTR has ribosomal association, although at lower ribosome density than the CDS or 5' UTR (Ingolia *et al.*, 2009; Guydosh and Green, 2014; Ji *et al.*, 2015; Miettinen and Björklund, 2015; Young *et al.*, 2015; Hsu *et al.*, 2016). Looking into the ribosomal association with these regions could indicate translation, proposed as an integral part of Wu dORF function (Wu *et al.*, 2020b). The alignment process used in this analysis differed from the previous HISAT2 (Kim *et al.*, 2019), allowing investigation of RP read alignment to Wu dORFs and identification of precise relative alignment locations. The NCBI BLAST+ blastn tool (Altschul *et al.*, 1997; Camacho *et al.*, 2009; Cock *et al.*, 2015) was used to align the RP reads and transcripts with RefSeq mRNA IDs (O'Leary *et al.*, 2016). The analysis continued to explore differences in ribosomal association in healthy and tumour tissue, but also investigated possible translation of Wu dORFs and 3' UTRs.

Table 3.2 considers all RP reads in the processed datasets, however, there are some issues with RP reads aligned to the same location repeatedly, which will be discussed below. This led to the inclusion of RP reads that align to a unique location, the 'uniquely aligned reads'. There are fewer reads aligned to CDSs, 3' UTRs and Wu dORFs when using the uniquely aligned reads (Table 3.2). Whether considering unique or all RP reads, across all the RP datasets there are considerably more reads aligned to the CDS compared to the 3' UTR of transcripts with RefSeq mRNA IDs (Table 3.2). The CDS has the greatest number of RP reads aligned, with thirty to fifty times as many as the 3' UTR (Table 3.2), highlighting just how few reads align to the 3' UTR compared to the CDS. Compared to the 3' UTR there are far fewer reads aligned to Wu dORFs (Table 3.2). The CDS, 3' UTR and Wu dORFs vary drastically in size indicating the need for the subsequent RP read density to account for the differences (section 3.6.1). The RP datasets generated from tumour tissue show alignment to the CDS, 3' UTR and Wu dORFs in transcripts with RefSeq mRNA IDs compared to RP datasets generated from healthy kidney samples (Table 3.2). The increased alignment in the tumour tissue compared to the healthy tissue is not purely driven by RP dataset size differences. Table 3.2 highlights the lack of RP read alignment to Wu dORFs and shows a large proportion of RP reads aligned to 3' UTRs do not align to Wu dORFs. Increased RP reads alignment in the tumour tissue datasets continues to indicate possible differences in ribosomal association in tumour tissue.

Table 3.2: Number of ribosome profiling (RP) reads aligned to different RefSeq mRNA ID transcript regions in each RP dataset from PRJNA256316. *The NCBI BLAST+ blastn tool was used to align RP datasets from healthy kidney samples (NT) and kidney tumour samples (TT). The aligned regions were: coding sequences (CDSs), 3' untranslated regions (3' UTRs) and Wu downstream open reading frames (dORFs). The table includes all aligned RP reads and uniquely aligned reads, referring to RP reads which align once to a region in a unique location, where no other read aligns to the same location. Unpaired t tests used to compare means and P values adjusted for multiple comparisons using Holm-Šídák method (P<0.05 - *).*

RP Dataset	Reads Aligned to CDSs		Reads Aligned to 3' UTRs		Reads Aligned to Wu dORFs	
	All	Unique	All	Unique	All	Unique
NT1.1	525485	205030	18423	5201	251	129
NT1.2	722863	251351	24940	6719	350	169
NT2.1	659479	273700	14196	5179	286	144
NT2.2	904098	334037	19226	6727	404	180
TT1.1	2103901	571962	53059	11516	1242	334
TT1.2	2873273	679064	71974	15178	1770	430
TT2.1	2335977	609356	156361	18374	1197	530
TT2.2	3215772	724836	211876	24365	1694	683
TT4.1	936877	335221	54469	10969	582	289
TT4.2	1278343	409362	74187	14588	820	400
NT Mean	**702981**	**266030**	**19196**	**5957**	**322.8**	**155.5**
TT Mean	**2124024**	**554967**	**103654**	**15832**	**1218**	**444.3**
Significance	*	*	*	*	*	*

Table 3.3: Human healthy kidney sample example (NT1.1) ribosome profiling (RP) dataset with RP reads repeatedly aligning to the top five RefSeq mRNA IDs coding sequence (CDS) locations. *The table also shows the proportion of the total reads aligned to the CDSs that these repeated alignments make up. The transcript version ID, gene, and start location of the RP read alignment relative to the CDS start are included.*

NT1.1 CDS	
Top 5 CDS Alignments	Number of RP reads Aligned
ENST00000388825.9_GPX3_202	3662
ENST00000252486.9_APOE_838	440
ENST00000647789.2_ALDOB_259	283
ENST00000646664.1_ACTB_535	271
ENST00000646664.1_ACTB_934	254
Total RP reads Aligned to CDSs	525485
RP reads aligned to top 5 alignments	4910
Proportion of total reads (%)	0.93

Table 3.4: Human kidney tumour sample example (TT1.1) ribosome profiling (RP) Dataset with RP reads repeatedly aligning to the top five RefSeq mRNA IDs coding sequence (CDS) locations. *The table also shows the proportion of the total reads aligned to the CDSs that these repeated alignments make up. The transcript version ID, gene, and start location of the RP read alignment relative to the CDS start are included.*

TT1.1 CDS	
Top 5 CDS Alignments	Number of RP reads Aligned
ENST00000225964.10_COL1A1_391	2988
ENST00000646664.1_ACTB_910	1433
ENST00000451311.7_TMSB4X_1	1008
ENST00000291568.7_CSTB_19	1003
ENST00000646664.1_ACTB_193	1001
Total RP reads Aligned to CDSs	2103901
RP reads aligned to top 5 alignments	7433
Proportion of total reads (%)	0.35

The difference between the results in table 3.2 for all and the unique aligned RP reads highlighted the impact of duplicate RP read alignment. The large number of repeated RP read alignments to transcripts is shown in Tables 3.3 and 3.4, displaying the top five alignment locations for a human healthy kidney RP dataset and a tumour RP dataset within the CDSs. Large numbers of RP reads align to the same locations in both healthy and tumour RP datasets (Tables 3.3 and 3.4). Tables 3.3 and 3.4 show that although a larger number of reads do align to the top five locations, that these reads only account for a very small proportion of the total RP reads aligned to the CDS. This reduces the impact of these repeated alignments in CDS RP alignments to the CDS regions, but it will still be considered in further analysis.

Table 3.5: Human healthy kidney sample example (NT1.1) ribosome profiling (RP) Dataset with RP reads repeatedly aligning to the top five RefSeq mRNA IDs 3' untranslated region (3' UTR) locations. *The table also shows the proportion of the total reads aligned to the 3' UTRs that these repeated alignments make up. The transcript version ID, gene, and start location of the RP read alignment relative to the 3' UTR start are included.*

NT1.1 3' UTR	
Top 5 3'UTR Alignment	Number of RP reads Aligned
ENST00000651323.1_CTC1_1541	3106
ENST00000651323.1_CTC1_1540	2031
ENST00000322434.8_ZNF354B_447	473
ENST00000620804.1_RIMBP3B_612	149
ENST00000219821.9_TMC5_754	138
Total RP reads Aligned to 3' UTRs	18423
RP reads aligned to top 5 alignments	5897
Proportion of total reads (%)	32.01

Table 3.6: Human kidney tumour sample example (TT1.1) ribosome profiling (RP) Dataset with RP reads repeatedly aligning to the top five RefSeq mRNA IDs 3' untranslated region (3' UTR) locations. *The table also shows the proportion of the total reads aligned to the 3' UTRs that these repeated alignments make up. The transcript version ID, gene, and start location of the RP read alignment relative to the 3' UTR start are included.*

TT1.1 3' UTR	
Top 5 3' UTR Alignments	Number of RP reads Aligned
ENST00000219821.9_TMC5_754	25262
ENST00000219821.9_TMC5_753	6335
ENST00000651323.1_CTC1_1540	3076
ENST00000651323.1_CTC1_1541	862
ENST00000219821.9_TMC5_755	475
Total RP reads Aligned to 3' UTRs	53059
RP reads aligned to top 5 alignments	36010
Proportion of total reads (%)	67.87

The same analysis investigated the top five 3' UTR RP read alignment locations. In these examples from human healthy kidney and tumour RP datasets, a large number of RP reads aligned to these repeated locations (Tables 3.5 and 3.6), an issue present in all RP datasets. In contrast to the CDS repeated alignments, a large proportion of the total 3’ UTR aligned RP reads aligned to the top five alignment locations (Tables 3.5 and 3.6), leading to a large impact on the alignments seen. The next few sections take account of these repeated alignments by using all the RP reads in the datasets but also RP read subsets. The ribosomes generating RP reads with large numbers aligning to the same place may be stalled and not translating, or another artefact. To avoid issues with large numbers of repeated RP alignments to the same location, subsets of RP reads were generated for further analysis. These subsets were RP reads with: only one RP read (unique), five or fewer RP reads, ten or fewer RP reads, aligned to the same location. These subsets were generated by assessing the number of RP reads aligned to each location. The last two subsets allow some amount of repeated alignment while excluding large numbers of repeated RP read alignments. Using the subsets of RP reads allows an insight into the influence of the repeated RP read alignment locations on subsequent analyses.

3.6.1 RP read density of alignments with CDSs, 3’ UTRs and Wu dORFs

The investigation of ribosomal association, and possible translation of 3’ UTRs and Wu dORFs is described below, taking account of differing RP dataset sizes and lengths of the 3’ UTR, CDS and Wu dORFs. The RP read density analysis used the full RP datasets and the subsets described previously and considers only transcripts with RefSeq mRNA IDs. Longer sequences have an increased likelihood of RP read alignments, making RP read density a more suitable comparison when also adjusted for the differing sizes of the RP datasets. Comparing the RP read density to CDSs, 3’ UTRs and Wu dORFs in RefSeq transcripts across the RP datasets provides insight into ribosomal association with these regions in healthy and tumour tissue. Across all RP datasets, regardless of RP read subset, the CDS had the greatest RP read density, followed by the Wu dORFs, and then the remaining 3’ UTR region (Figure 3.10). In all RP subsets, the RP read density across all regions was increased in the tumour RP datasets compared to the healthy kidney RP datasets (Figure 3.10). Restricting the RP reads into the subset groups decreased the read density seen in all datasets and regions, shown by the altered axes (Figure 3.10). This decrease was consistent and the fewer RP reads allowed to align to the same location in the subset, the greater the reduction in the read density. The read density analysis supports the previous analysis looking into the number of RP reads aligned to the different mRNA regions (Table 3.2), showing that the CDS had the greatest ribosomal association and tumour tissue datasets had more ribosomal association with the mRNA regions (Figure 3.10). However, the read density reduces the difference between the mRNA regions, taking into account the much larger length of the CDS and 3' UTR compared to Wu dORFs. Although the number of reads in the Wu dORFs was the fewest (Table 3.2), the RP read density in these Wu dORFs was greater than the 3' UTR (Figure 3.10). The ribosomal association with Wu dORFs is greater than that seen in the 3' UTR as a whole, however the CDS has greater ribosomal association.

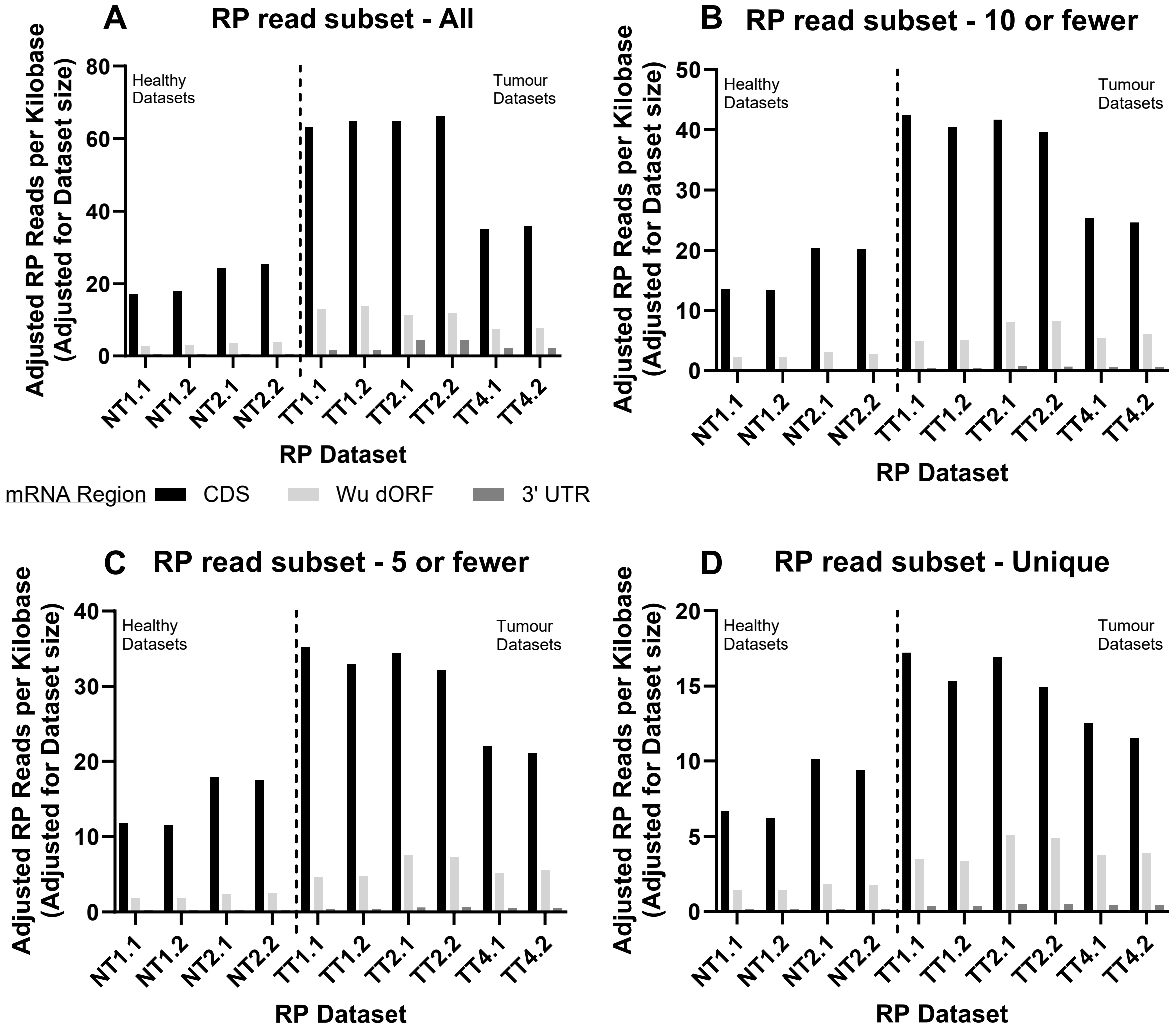

Figure 3.10: Ribosome profiling (RP) read density was greatest across coding sequences (CDSs), followed by Wu dORFs, and then 3' untranslated regions (3' UTRs). *Transcripts with RefSeq mRNA IDs included and PRJNA256316 RP datasets used. Read density is represented as the number of RP reads per kilobase of the region of concern in transcripts with RP read alignment. This read density was adjusted to acknowledge differing number of RP reads in the datasets, normalising all densities to datasets with 20,000,000 RP reads. A – Includes all RP reads aligned to each mRNA region. B – Only includes RP reads aligned to each mRNA region which have 10 or fewer RP reads to each location. C – Only includes RP reads aligned to each mRNA region which have 5 or fewer RP reads to each location. D – Only includes RP reads aligned to each mRNA region which have one RP read to each location. The human healthy (NT) and tumour tissue (TT) datasets are displayed divided by a dashed vertical line.*

3.6.2 Reading Frame Analysis

Ribosomal presence, represented by RP alignments to regions of a transcript, could give an indication of the relative translation. Analysing the reading frame of the RP alignments investigated translation in these regions further, with a focus on dORF translation, suggested to be required for their proposed regulatory function (Wu *et al.*, 2020b). A preference for one reading frame over another, suggesting possible periodicity, could indicate translation in RP read aligned mRNA regions. This analysis used RP reads aligned to the CDS, 3' UTR and Wu dORFs of transcripts with RefSeq mRNA IDs. Again, the RP read subsets for each dataset were also used. The CDS was used for comparison to the 3' UTR and Wu dORFs as ribosomes present in the CDS are likely to be translating. The analysis doesn't determine the biological frame that the ribosome was in and instead represents the reading frames with letters relative to the CDS, 3' UTR, or dORF start. The reading frame analysis used the first base of the RP read alignment to the transcript to represent the reading frame of the ribosome, assuming that the first base could be a consistent distance from the translated codon. Instead of determining the genomic reading frame, this analysis determined which RP reads were in the same reading frames, due to the distance between the codon and 5' end of the RP read not being determined. This reading frame analysis was done on the RP read alignment data. First the RP aligned sequence, e.g. transcript 3' UTR, CDS or Wu dORF, was split into three reading frames from the start of the sequence (Figure 3.11). Then the start of the alignment at the 5' end of the aligned RP read was assigned a reading frame (Figure 3.11). This is why it was important to avoid trimming the 5' end of RP reads, to prevent the distance between the codon and the 5' end from changing. Potentially this distance may not be consistent, meaning this method may not represent the reading frame for all RP reads. However, if this affected a small proportion of reads, this would still provide useful data and possible support of translation.

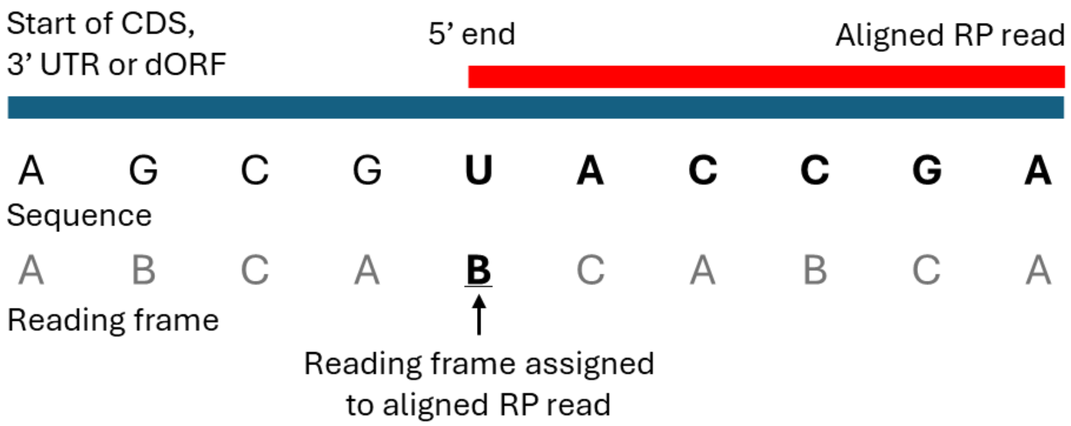

Figure 3.11: Diagram demonstrating how reading frame analysis determines the reading frame assigned to each ribosome profiling (RP) read aligned to coding sequences (CDSs), 3' untranslated regions (3' UTRs) or Wu downstream open reading frames (dORFs). *Reading frames A, B, and C are assigned to each base in the sequences aligned by the RP read, staring from the beginning of the region of interest. The 5' end of the aligned RP read is used to determine the reading frame of the RP read.*

In CDS aligned RP reads, reading frame A was preferred (Figure 3.12). The proportions varied in different datasets, but across the RP subsets, reading frame A had the greatest proportion of reads (Figure 3.12). Both healthy kidney and kidney tumour datasets had similar results with at least 50% of the reads fall in reading frame A. This proportion increased as the RP read subset groups became less stringent (Figure 3.12). When considering all 3' UTR aligned RP reads, reading frame A had the greatest proportion of RP reads to varied extents across the datasets (Figure 3.12). Healthy datasets had similar preference for reading frame A and B, reduced in reading frame C, whereas tumour datasets had a greater proportion of reading frame A RP reads (Figure 3.12). When using the subsets of RP reads, and regardless of the RP dataset being from health or tumour samples, there was an even split of RP across the reading frames, with no reading frame preference in the 3' UTR (Figure 3.12). The 3' UTR results indicated the influence of RP reads repeatedly aligned to the same location, removed these showed consistent results (Figure 3.12).

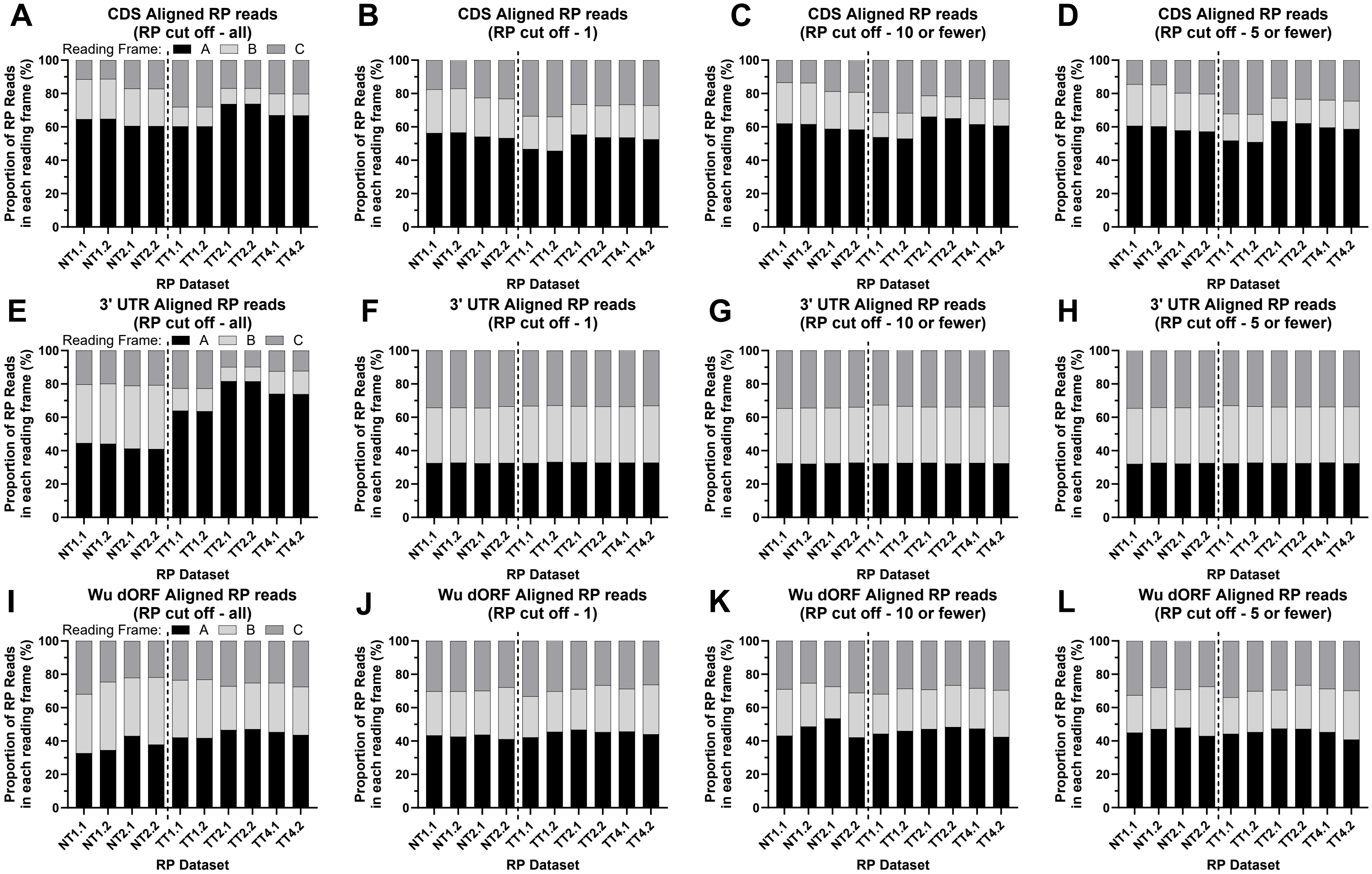

Figure 3.12: Proportion of ribosome profiling (RP) reads aligned in each reading frame (A, B, or C) in coding sequences (CDSs), 3' untranslated regions (3' UTRs), and Wu downstream open reading frames (dORFs). *Transcripts with RefSeq mRNA IDs and RP datasets from PRJNA256316were used. The proportion of RP reads aligned in different reading frames are displayed as a percentage of the total reads aligned to CDS (A-D), 3' UTR (E-H), and Wu dORFs (I-L). The charts include: all RP reads aligned (A, E, I), RP reads aligned with: one RP read per location (B, F, J), 10 or fewer RP reads to each location (C, G, K), 5 or fewer RP reads to each location (D, H, L). The human healthy (NT) and tumour tissue (TT) datasets are divided by the dashed line. Reading Frames: A – Black, B – Pale Grey, C – Mid Grey.*

The reading frame analysis of all RP reads aligned to Wu dORFs indicated healthy RP datasets had a greater proportion of reads in reading frame A and B, split evenly between the two (Figure 3.12). In tumour RP datasets, when considering all RP reads, a greater proportion of reads aligned in reading frame A, around 40% (Figure 3.12). RP read subsets were consistent, with between 40-50% of reads in reading frame A in Wu dORFs (Figure 3.12). As the RP read subsets became more stringent, moving from 10 or fewer to uniquely aligned, the reading frame results became more consistent across the RP datasets (Figure 3.12). These results were consistent in both the healthy and tumour RP datasets (Figure 3.12). This proportion of reads in reading frame A in Wu dORFs was reduced compared to RP reads aligned across CDSs, but greater than the 3' UTR, suggesting some preference for reading frame A, not seen with the 3' UTR. This could indicate some translation occurred in the Wu dORFs, not to the extent of the CDS, but more than the little evidence of translation seen in the 3' UTR.

3.6.3 Using results from RP read alignment to Wu dORFs to investigate evidence of Wu dORF translation

Wu dORFs were investigated further to identify those individual dORFs with consistent RP read alignment across the RP datasets with a similar reading frame A preference seen in the CDS, suggesting dORF translation. RP datasets from healthy kidney samples were used as these datasets had suggested dORF activity (Figure 3.2). Very few dORFs were found with a similar reading frame preference to the CDS, with at least 50% of the reads aligned in reading frame A (Table 3.7). The few dORFs with similar reading frame preference to the CDS in each RP dataset were not found consistently across the human healthy kidney datasets (Table 3.7). Very few Wu dORFs had more than two RP reads aligned in reading frame A, meaning that even if the preference was similar to the CDS there was little ribosomal association with these Wu dORFs (Table 3.7). Wu dORFs generally lacked ribosomal association and there was little evidence of translation of Wu dORFs in the healthy RP datasets.

Table 3.7: Few Wu downstream open reading frames (dORFs) have evidence of translation when considering all ribosome profiling (RP) reads aligned to these Wu dORFs in human healthy kidney samples datasets (NT) from PRJNA256316. *This table reports the number of Wu dORFs with similar reading frame preference to the overall coding sequence (CDS) alignments (50+% of RP reads in reading Frame A). Then how many Wu dORFs appeared with consistency across all healthy human kidney RP datasets in PRJNA256316. Before looking at Wu dORFs with more than 2 RP reads in reading frame A.*

<u>All</u> BLASTn Alignments	**Number of Wu dORFs with reading frame preference similar to CDS**	**Wu dORFs appearing in multiple datasets**	**Wu dORFs with >2 RP reads in reading frame A**
All Healthy RP Datasets	36 (NT1.1 - 11, NT2.2 - 13, NT2.1 - 14, NT2.2 - 14)	9 (3 appear in all healthy datasets)	9

3.7 Investigating Wu dORF function in further paired RNAseq and RP datasets generated from cancer cells

Sections 3.2 and 3.3 showed potential function of Wu dORFs as translational regulators in both healthy and tumour tissue, with their activity reduced in tumour tissue. The same methods described in section 2.4, were used with datasets from two further SRA projects with paired RNAseq and RP datasets from A549 pulmonary adenocarcinomic human alveolar basal epithelial cells with control (Ctrl) and *FKBP10* knock down with shRNA (KD) (PRJNA532400), and RKO human colorectal carcinoma cells under fed (400 uM arginine) or starved (12.5 uM arginine) conditions for 24 hours (PRJNA880902). These datasets were used to investigate whether the previous tumour tissue results were consistent in other cancer cells. The second project also provides an opportunity to investigate whether cell stress, which can influence uORF activity (Renz, Valdivia Francia and Sendoel, 2020) and 3’ UTR ribosomal presence (Ingolia *et al.*, 2009), influences Wu dORF activity, using arginine starvation. Analysis included the differential expression of a transcript from the RNAseq to the RP dataset, used again to model translational regulation, and the correlations between 3’ UTR RP dataset expression and the differential expression values, or the CDS RP dataset expression.

The datasets from A549 pulmonary adenocarcinomic human alveolar basal epithelial cells had consistent results whether with control cells (Ctrl) or *FKBP10* knock down with shRNA (KD) (Figure 3.13). *FKBP10* knock down has been shown to reduce lung tumour growth in mice and decreases translation elongation at the start of ORFs (Ramadori *et al.*, 2020). Across these cancer cell datasets there was an increased Log2 fold change in CDS expression from RNAseq to RP datasets for RefSeq mRNA ID transcripts with a Wu dORF compared to those without (Figure 3.13). The mean difference between transcripts with and without Wu dORFs were 0.7290 (95% confidence interval for difference 0.9804 to 0.4776), 0.8379 (1.091 to 0.5847), 0.6274 (0.8854 to 0.3693), 0.7716 (1.024 to 0.5191) across datasets, supported by shifted distribution of the box and whisker plots (Figure 3.13). All of these differences in means were statistically significantly (P<0.0001). These mean differences and the data distributions were similar to the tumour tissue results in Figure 3.3, suggesting consistent Wu dORF activity in these cells and the tumour tissue analysed previously. Similar considerations remain from section 3.2, with large variation in results indicated by the whiskers and the difference in transcript group sizes. A similar number of transcripts with Wu dORFs were included in this analysis and in the tumour tissue previously. These results indicate that Wu dORF activity remains in these cancer cells with similarity to the tumour tissue results, however activity remains reduced compared to healthy tissue results.

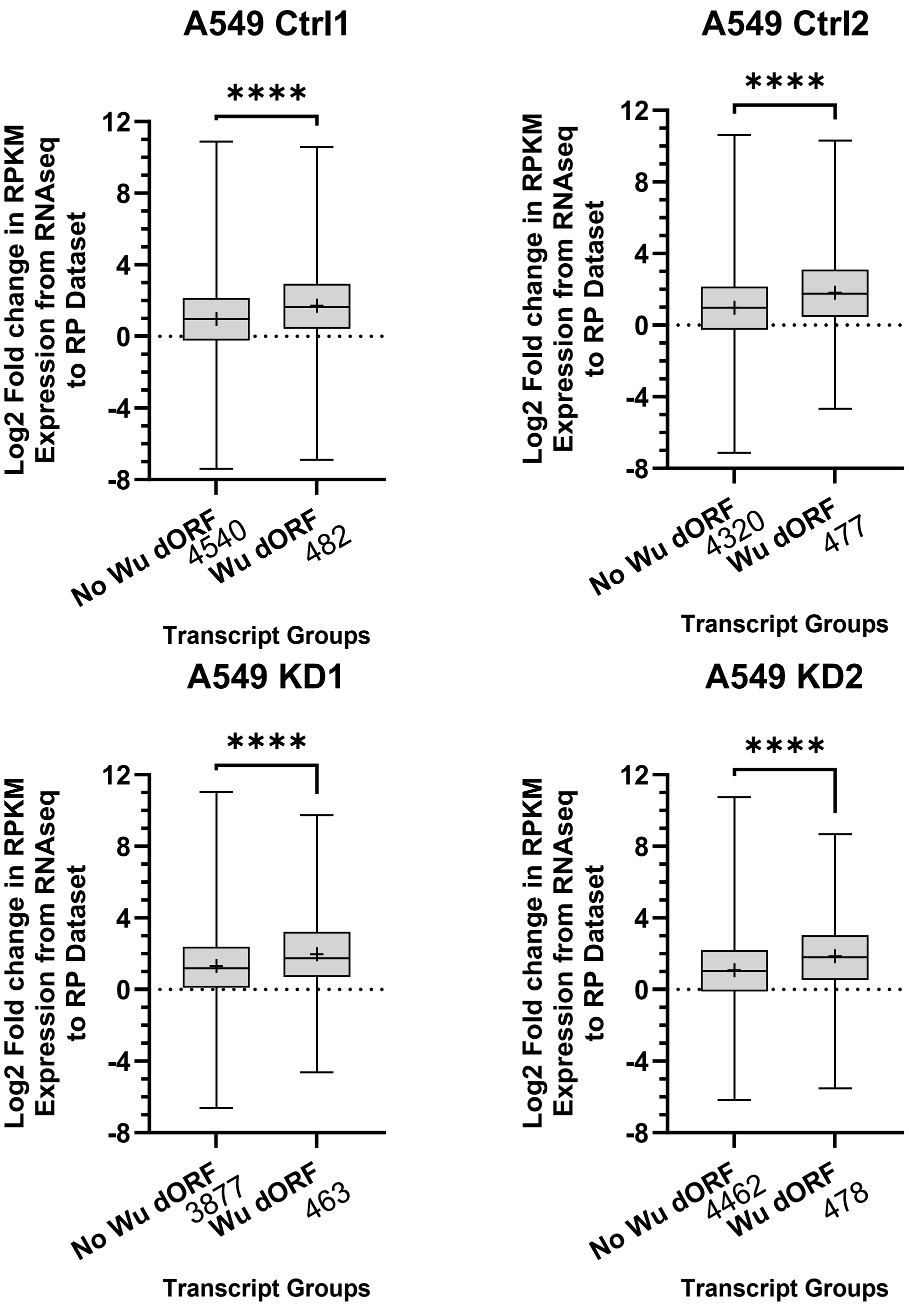

Figure 3.13: RefSeq mRNA ID transcripts with Wu downstream open reading frames (dORFs) appear to be translationally upregulated in A549 pulmonary adenocarcinomic human alveolar basal epithelial cells with control (Ctrl) and FKBP10 knockdown with shRNA (KD) from PRJNA532400. *Relative transcript translation calculated as Log2 fold change in expression from RNAseq to RP dataset for each transcript. Box and whisker plots for each group of transcripts, the horizontal line within the box marks the median, the whiskers denote the maximum and minimum values, and the box marks the 25th to the 75th percentile of the values distribution, with the mean marked with '+'. One-Way ANOVA multiple comparisons test with Šidák adjusted P values was used to compare the means. Each chart title describes which RP dataset was used in the analysis. Transcript group sizes included in the x axis labels. **** - P<0.0001.*

The same datasets from cancer cells were then used to explore whether the correlation results gathered previously from the tumour tissue datasets were also consistent in these datasets. The correlation between 3’ UTR RP expression and the Log2 fold change in expression from RNAseq to RP dataset, modelling translational regulation was carried out (Figure 3.14), before correlating against the CDS RP expression (Figure 3.15). In both cases, initially all RefSeq mRNA transcripts with RP read alignment to the 3’ UTR and CDS with differential expression values were used and then only the transcripts containing a Wu dORF. The correlations assessed whether there was a relationship between the extent of 3’ UTR ribosomal association and the potential transcript translational regulation, or CDS ribosomal association. The number of transcripts, and those with Wu dORFs, included in these correlations (Figures 3.14 and 3.15) were similar to the results for the tumour tissue datasets in Figures 3.4 – 3.7. Whether transcripts contained a Wu dORF or not, and whether they were the control or knock down cells, there was a consistent small positive correlation between the Log2 fold change in expression from RNAseq to RP dataset and the 3’ UTR RP dataset expression. The Pearson r values for these correlations ranged from 0.1650 and 0.2032 and all were statistically significant ($P<0.05$) (Figure 3.14). Once again there was variation and a spread of data points, but the correlation was greater than that seen previously in healthy or tumour tissue datasets. The correlation between the CDS and 3’ UTR ribosome occupancy was similar to that seen previously in healthy and tumour datasets. There was a statistically significant ($P<0.001$) positive correlation across all the datasets from these cancer cells whether considering all transcripts or those with Wu dORFs (Figure 3.15). As seen with the healthy and tumour tissue before (Figure 3.6 and 3.7), when considering the transcripts with Wu dORFs the correlation increased, with Pearson r values ranging from 0.6095 to 0.6749, whereas with all transcripts Pearson r values ranged from 0.4661 to 0.5130 (Figure 3.15). The datapoints are more clustered, especially in the transcripts with Wu dORFs (Figure 3.15). These results showed a small correlation between 3’ UTR ribosome occupancy and potential translational regulation, not seen in tumour tissue datasets previously, and a positive correlation between 3’ UTR and CDS ribosome occupancy that was slightly increased in transcripts with Wu dORFs.

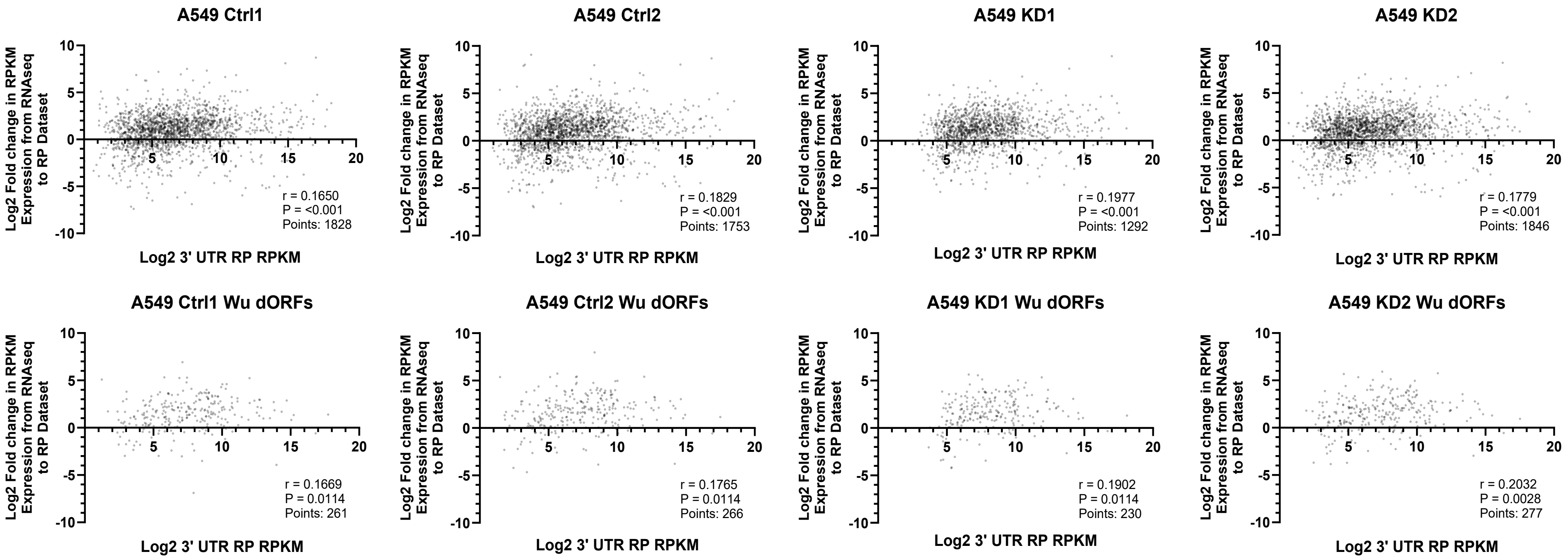

Figure 3.14: Correlation between translational regulation and 3' untranslated region (UTR) expression in ribosome profiling (RP) datasets for RefSeq mRNA ID transcripts, and those with Wu dORFs in PRJNA532400 datasets. *Relative transcript translation calculated as Log2 fold change in expression from RNAseq to RP dataset for each transcript. The correlation was quantified using the Pearson correlation coefficient. Each chart title describes which RP dataset was used in the analysis e.g. A549 Ctrl1. The Pearson r coefficient, P value and number of data points are included on the charts. Top row of charts includes all RefSeq mRNA ID transcripts, and the bottom row includes transcripts with Wu dORFs. P values adjusted for multiple comparisons using Holm-Šídák method.*

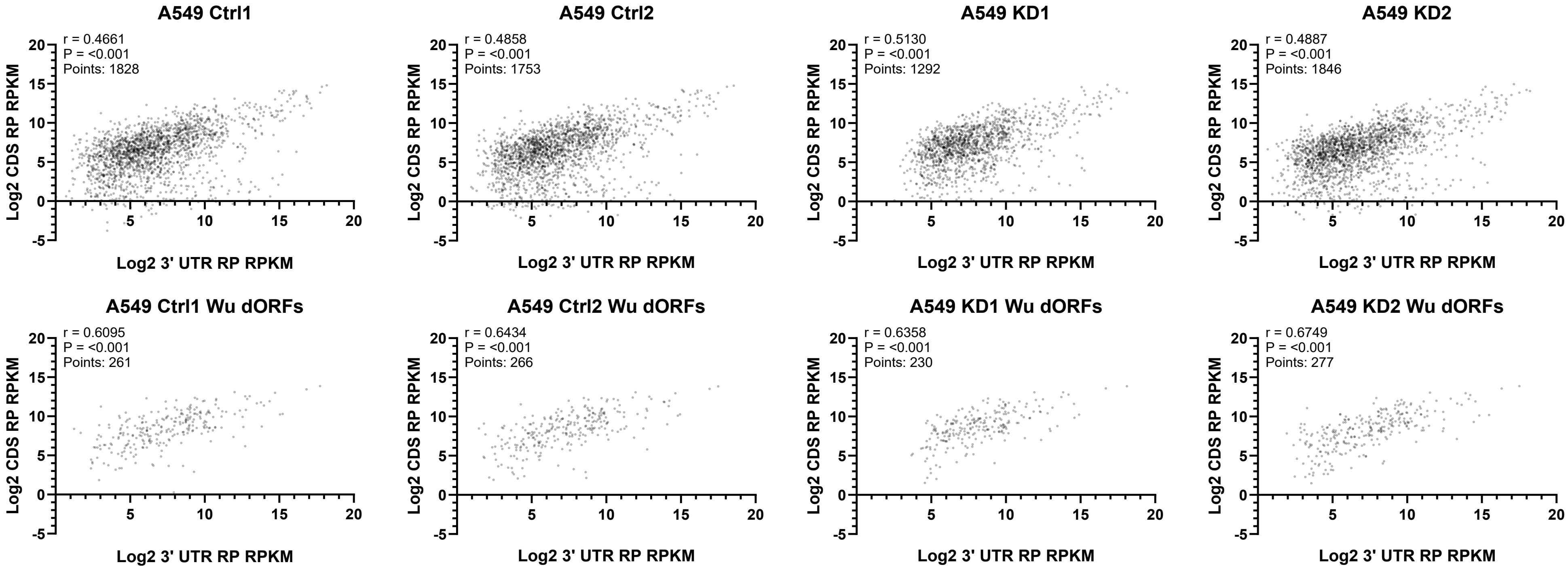

Figure 3.15: Correlation between coding sequence (CDS) and 3' untranslated region (UTR) expression in ribosome profiling (RP) datasets for RefSeq mRNA ID transcripts in PRJNA532400 datasets. *CDS and 3' UTR expression were calculated as Log2 values. The correlation was quantified using the Pearson correlation coefficient. Each chart title describes which RP dataset was used in the analysis e.g. A549 Ctrl1. The Pearson r coefficient, P value and number of data points are included on the charts. Top row of charts includes all RefSeq mRNA ID transcripts, and the bottom row includes transcripts with Wu dORFs. P values adjusted for multiple comparisons using Holm-Šídák method.*

The investigation of other cancer cell datasets continued with RKO human colorectal carcinoma cells under fed (400 µM arginine) or starved (12.5 µM arginine) conditions for 24 hours. This also allowed the effect of cell stress caused by arginine starvation to be assessed on Wu dORF activity alongside gathering data from another cancer cell type. Initially the analysis exploring potential translational regulation of transcripts with and without Wu dORFs was carried out. The results from datasets from RKO cells under fed or starved conditions were very similar (Figure 3.16). There was a small increase in Log2 fold change in CDS expression from RNAseq to RP datasets for RefSeq mRNA ID transcripts with a Wu dORF compared to those without (Figure 3.16). The mean difference between transcripts with and without Wu dORFs in fed cell datasets were 0.1964 (95% confidence interval for difference 0.3821 to 0.01070, P = 0.0319), 0.1488 (0.3368 to -0.03915, P = 0.2037), 0.1969 (0.3831 to 0.01062, P = 0.0321). For starved cell datasets the mean differences were 0.1718 (0.3592 to -0.01549, P= 0.0911), 0.1540 (0.3426 to -0.03460, P = 0.1757), 0.2244 (0.4118 to 0.03693, P = 0.0098). Across all datasets there was little shift in the box and whisker plot distributions (Figure 3.16). Four of the mean differences were statistically not significant (P>0.05). There was a more compact distribution of results around the mean and median indicated by the shorter boxes (Figure 3.16). The whiskers showed variation in the results and transcript group sizes were similar to the previous cancer analysis (Figure 3.16). These results were not consistent with the other cancer results in Figure 3.3 and Figure 3.13, with reduced dORF activity compared to the other cancer datasets. In RKO cells there was little evidence of Wu dORF activity. This suggests dORF activity varies across cancer types. Nevertheless, there is reduced Wu dORF activity compared to the healthy tissue dataset. In addition, cell stress, induced by arginine starvation, had no effect on Wu dORF activity in these cells.

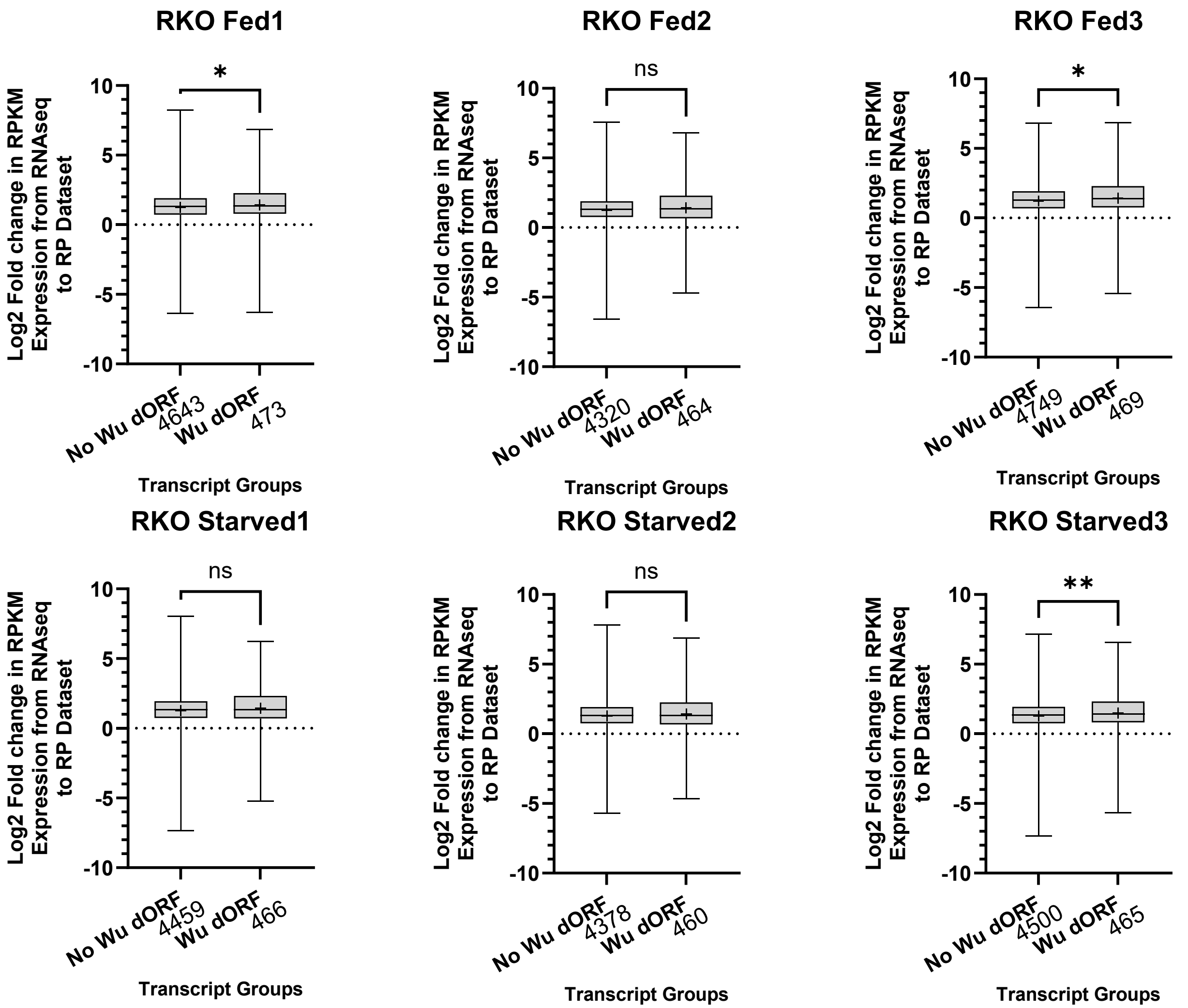

Figure 3.16: Translational regulation of RefSeq mRNA ID transcripts with, and without, Wu downstream open reading frames (dORFs) appears to be similar in RKO human colorectal carcinoma cells under fed (400 uM arginine) or starved (12.5 uM arginine) conditions for 24 hours from PRJNA880902. *Relative transcript translation calculated as Log2 fold change in expression from RNAseq to RP dataset for each transcript. Box and whisker plots for each group of transcripts, the horizontal line within the box marks the median, the whiskers denote the maximum and minimum values, and the box marks the 25th to the 75th percentile of the values distribution, with the mean marked with '+'. One-Way ANOVA multiple comparisons test with Šídák adjusted P values was used to compare the means. Each chart title describes which RP dataset was used in the analysis. Transcript group sizes included in the x axis labels. ns – P>0.05, * - P<0.05, ** - P<0.005.*

The same datasets from cancer cells under fed and starved conditions were then used to explore whether the previous correlation results (Figures 3.14-15) were consistent, and to investigate the effect of cell stress. Correlations were between 3' UTR RP expression and the Log2 fold change in expression from RNAseq to RP dataset, modelling translational regulation (Figure 3.17), or the CDS RP expression (Figure 3.18). All RefSeq mRNA transcripts with RP read alignment to the 3' UTR and CDS and differential expression values were included, and then only those with a Wu dORF. The numbers of transcripts included were similar to previous cancer cell or tumour tissue results (Figures 3.17 and 3.18). Both fed and starved cell datasets showed a consistent small negative correlation between the 3' UTR RP dataset expression and the Log2 fold change in expression from RNAseq to RP dataset (Figure 3.17). The correlation was slightly more negative in Wu dORF-containing transcripts (Figure 3.17). The statistically significant (P<0.001) Pearson r values for all transcripts ranged from -0.2648 to -0.2804, and for transcripts with Wu dORFs the range was from -0.3670 to -0.4087 (Figure 3.17). Once again there was variation and a spread of data points, and the negative correlation differed from the previous positive correlation in A549 cancer cell datasets (Figure 3.14) and the lack of biologically meaningful correlation in tumour tissue datasets (Figures 3.4 and 3.5). The correlation CDS and 3' UTR ribosome occupancy correlation showed a statistically significant (P<0.001) consistent positive correlation across all the fed and starved cell datasets regardless of whether the transcripts contained Wu dORFs (Figure 3.18). The Pearson r values ranged from 0.5602 to 0.6592 (Figure 3.18). This is similar to all datasets analysed previously. Stressful cellular conditions do not appear to affect the correlations. These results indicate a negative correlation between potential translational regulation and 3' UTR ribosome occupancy, not seen previously, and another positive correlation between 3' UTR and CDS ribosome occupancy.

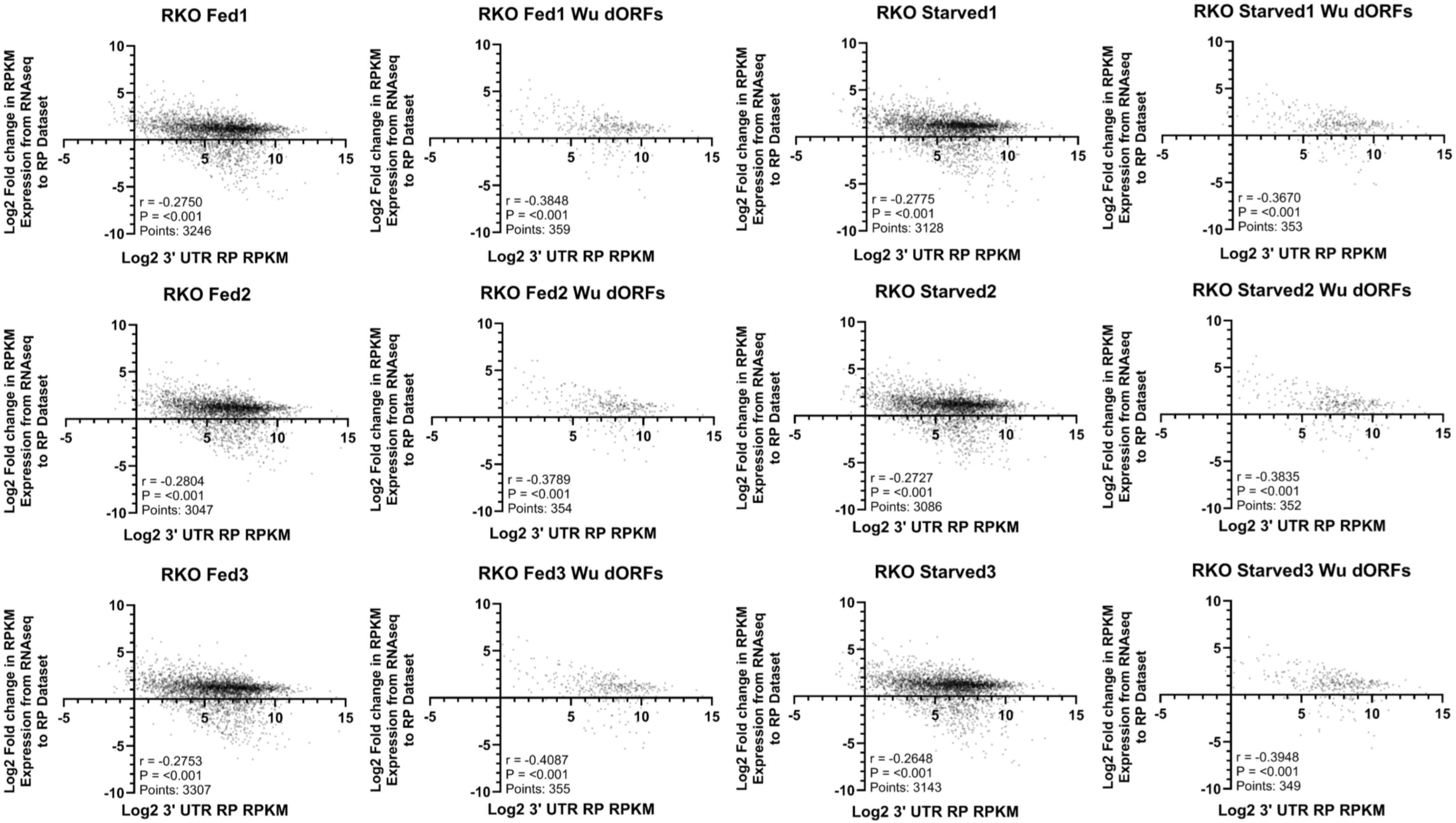

Figure 3.17: Correlation between translational regulation and 3' untranslated region (UTR) expression in ribosome profiling (RP) datasets for RefSeq mRNA ID transcripts, and those with Wu dORFs in PRJNA880902 datasets. *Relative transcript translation calculated as Log2 fold change in expression from RNAseq to RP dataset for each transcript. 3' UTR expression was calculated as Log2 values. The correlation was quantified using the Pearson correlation coefficient. Each chart title describes which RP dataset was used in the analysis e.g. RKO Fed1. The Pearson r coefficient, P value and number of data points are included on the charts. Columns of charts alternate with first and third column including all RefSeq mRNA ID transcripts and the second and fourth column including transcripts with Wu dORFs. P values adjusted for multiple comparisons using Holm-Šídák method.*

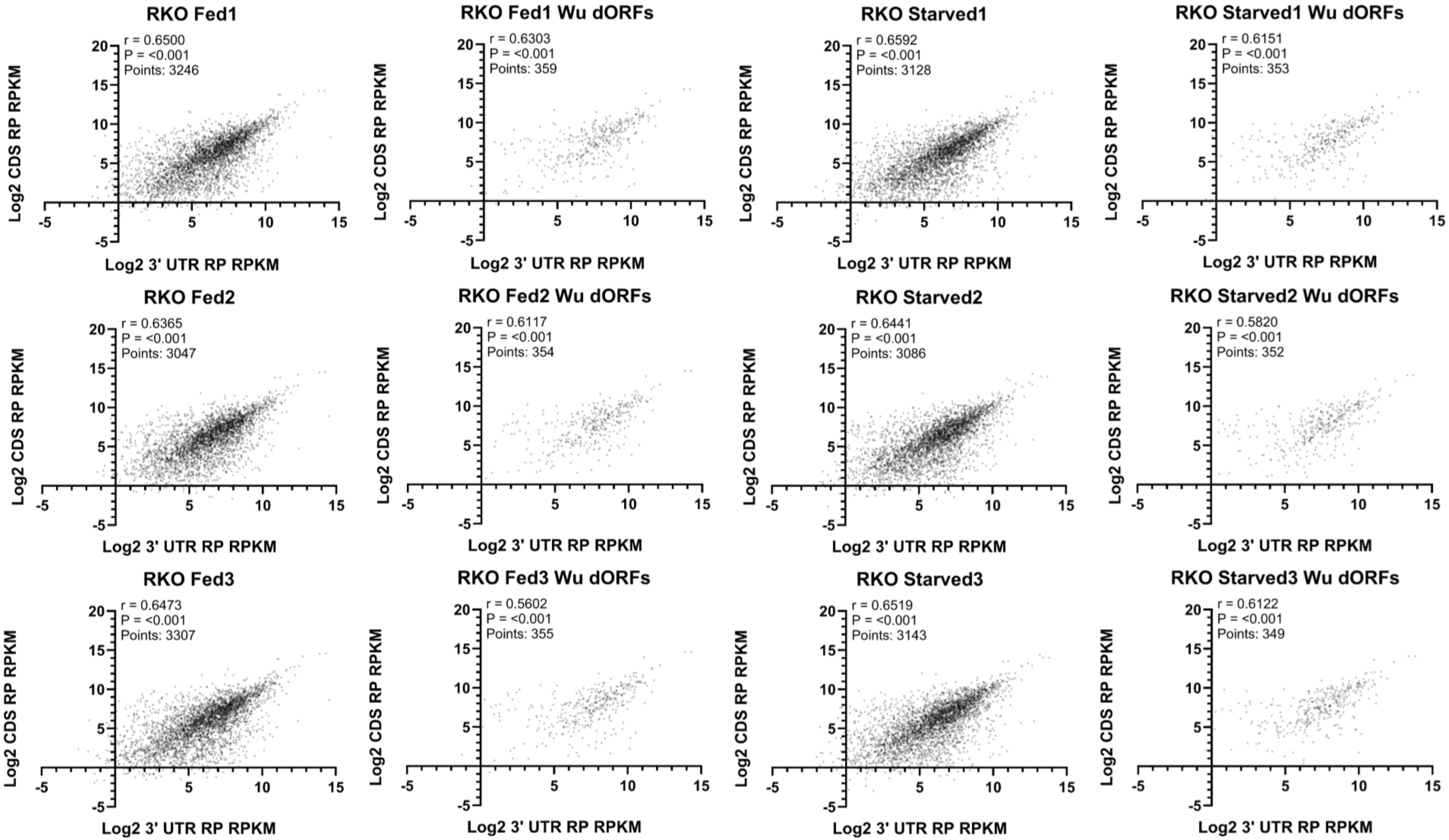

Figure 3.18: Correlation between coding sequence (CDS) and 3' untranslated region (UTR) expression in ribosome profiling (RP) datasets for RefSeq mRNA ID transcripts, and those with Wu dORFs, in PRJNA880902 datasets. *CDS and 3' UTR expression were calculated as Log2 values. The correlation was quantified using the Pearson correlation coefficient. Each chart title describes which RP dataset was used in the analysis e.g. RKO Fed1. The Pearson r coefficient, P value and number of data points are included on the charts. Columns of charts alternate with first and third column including all RefSeq mRNA ID transcripts and the second and fourth column including transcripts with Wu dORFs. P values adjusted for multiple comparisons using Holm-Šídák method.*

3.8 Shortlist of Wu dORFs

The result from section 3.6 aimed to investigate ribosomal investigation and possible translation of the Wu dORFs, dORFs presented by Wu *et al.* (2020b). However, doubt remains over whether there is evidence of Wu dORF translation. The methods described in section 2.6 were used to identify a Wu dORF shortlist with greater confidence in their translation, using databases of MS validated short proteins: OpenProt (Leblanc *et al.*, 2024), SmProt (Y. Li *et al.*, 2021) and MetamORF (Choteau *et al.*, 2021). The shortlist, known as 'MSVW dORFs' (MS Validated Wu dORFs), contains 26 Wu dORFs that encode MS validated proteins, evidencing their translation. Tables 3.8 and 3.9 contain details, and previous results, gathered about these dORFs. The length of MSVW dORFs ranged from 90 to 282 nucleotides, all use an AUG start codon, and they were found from the start of the 3' UTR to several thousand bases downstream of the CDS stop codon (Table 3.8). Although translated, in the human healthy kidney and kidney tumour tissue RP datasets only eleven of these MSVW dORFs were RP read aligned in any of the datasets (Table 3.8). The genes containing these MSVW dORFs varied in their functions, from cell growth and proliferation to neuronal development, the immune system and inflammation. Most commonly these genes are associated with transcription, RNA binding, and splicing, with ten genes associated with these processes (Table 3.8). Table 3.8 also reported links between the genes and cancers, of the 26 genes, fourteen were associated with cancers. Three genes were reportedly tumour suppressors, ten were reported to promoting cancers, and one had differing suppressive or oncogenic properties depending on the cancer type. The relevance of some MSVW dORF genes to cancer highlights the impact these dORFs could have as an avenue for future cancer treatments.

Table 3.8 Brief details about shortlist of Wu downstream open reading frames (dORFs) with mass spectroscopy (MS) validation of the encoded protein, MSVW dORFs. *The dORF length and location in the 3' untranslated region (UTR) of 26 dORFs were included. There is a brief description of the gene function for those containing the dORFs from the NCBI Gene database (Sayers et al., 2022). Ribosome profiling (RP) read aligned refers to whether any RP reads aligned to the dORF in the PRJNA256316 RP datasets. The final column includes details about whether the gene with the dORF is reported to have a role in cancers.*

Wu dORF	**Length**	**3' UTR Location**	**NCBI Gene Description**	**RP read aligned**	**Association of gene with cancers**
3 ENST00000006015 HOXA11	117	50:166	Homeobox A11: encodes a DNA-binding transcription factor which may regulate gene expression, morphogenesis, and differentiation.	FALSE	*HOXA11* referred to as a tumour suppressor in renal cell carcinoma, often silenced by promoter methylation (Wang *et al.*, 2017).
22 ENST00000175756 PTPN18	165	492:656	Protein tyrosine phosphatase non-receptor type 18: encodes a member of the protein tyrosine phosphatase (PTP) family, signalling molecules regulating cellular processes including cell growth, differentiation, the mitotic cycle, and oncogenic transformation. This PTP contains a PEST motif, often a protein-protein interaction domain and can differentially dephosphorylate autophosphorylated tyrosine kinases overexpressed in tumour tissues. Suggested to regulate HER2 (epidermal growth factor receptor family of receptor tyrosine kinases).	FALSE	*PTPN18* frequently highly expressed in colorectal cancer (CRC) suggested to promote CRC development by stabilizing the MYC protein level, activating the MYC-CDK4 axis (C. Li *et al.*, 2021).
103 ENST00000225504 SUPT4H1	291	230:520	SPT4 homolog, DSIF elongation factor subunit: encodes the small subunit of DRB (5,6-dichloro-1-beta-d-ribofuranosylbenzimidazole) sensitivity-inducing factor (DSIF) complex, localized to the nucleus, which regulates mRNA processing and transcription elongation by RNA polymerase II.	FALSE	N/A

Wu dORF	Length	3’ UTR Location	NCBI Gene Description	RP read aligned	Association of gene with cancers
226 ENST00000256015 BTG1	123	15:137	BTG anti-proliferation factor 1: encodes a member of an anti-proliferative gene family that regulates cell growth and differentiation. Expression is highest in the G0/G1 cell cycle phases and is downregulated following G1. Interacts with several nuclear receptors, as a coactivator of cell differentiation.	TRUE	Reduced *BTG1* expression is associated with increased disease severity and progression in kidney cancer (Sun *et al.*, 2015). *BTG1* is often deleted or mutated in B-cell leukaemia and lymphoma and downregulated in solid tumours, correlating with malignant cell behaviour and poorer outcomes (Yuniati *et al.*, 2019).
548 ENST00000293525 KRT86	105	396:500	Keratin 86: encodes a type II keratin protein, which heterodimerizes with type I keratins to form hair and nails.	FALSE	N/A
846 ENST00000325630 SPINK6	114	11:124	Serine peptidase inhibitor Kazal type 6: encodes a Kazal-type serine protease inhibitor that acts on kallikrein-related peptidases in the skin.	FALSE	SPINK6 can lead to tumour progression by interacting with the EGFR protein (Liao *et al.*, 2022).
921 ENST00000333007 TNFAIP2	282	362:643	TNF alpha induced protein 2: Tumour necrosis factor alpha (TNF) and retinoic acid can induce gene expression. In acute promyelocytic leukaemia this gene could be a retinoic acid target gene.	TRUE	TNFAIP2 suggested to be involved in carcinogenesis, upregulated in head and neck squamous cell carcinoma, stomach adenocarcinoma, diffuse large B-cell lymphoma, glioblastoma multiforme, kidney renal clear cell carcinoma, kidney renal papillary cell carcinoma (Lin *et al.*, 2022).
1275 ENST00000370783 MOSPD1	123	453:575	Motile sperm domain containing 1: Predicted to be involved in regulation of RNA polymerase II transcription.	FALSE	*MOSPD1* upregulated in colorectal cancer and may be involved in cancer progression through the Wnt/β-catenin signaling pathway (Horie *et al.*, 2022).
1400 ENST00000375446 NINJ1	273	92:364	Ninjurin 1: encodes a protein with homophilic adhesion molecule properties and it promotes neurite outgrowth in dorsal root ganglion neurons and is upregulated after nerve injury in Schwann cells and dorsal root ganglion neurons.	TRUE	*Ninj1* is overexpressed in human cancer and its role in tumorigenesis may relate to the p53 tumour suppressor (Cho *et al.*, 2013).

Wu dORF	Length	3' UTR Location	NCBI Gene Description	RP read aligned	Association of gene with cancers
1408 ENST00000375856 IRS2	273	287:559	Insulin receptor substrate 2: encodes a cytoplasmic signalling molecule that mediates effects of insulin, insulin-like growth factor 1, and other cytokines by acting as a molecular adaptor between diverse receptor tyrosine kinases, such as insulin and interleukin 4, and downstream effectors.	TRUE	N/A
1548 ENST00000393939 HAP1	99	3134:3232	Huntingtin associated protein 1: encodes a protein suggested to be involved with organelle transport or vesicular trafficking that interacts with huntingtin, with two cytoskeletal proteins (dynactin and pericentriolar autoantigen protein 1), and hepatocyte growth factor-regulated tyrosine kinase substrate.	TRUE	N/A
1575 ENST00000395123 SLC43A3	165	215:379	Solute carrier family 43 member 3: Predicted to be an integral membrane component that has a role in transmembrane transport.	FALSE	N/A
1849 ENST00000483063 POLE4	90	212:301	DNA polymerase epsilon 4, accessory subunit: encodes a histone-fold protein that interacts with other histone-fold proteins and larger enzymatic complexes to bind DNA for transcription, replication, and packaging.	TRUE	In mice POLE4 deficiency can predispose to tumour formation (Bellelli *et al.*, 2018).
2024 ENST00000589123 NFIC	108	836:943	Nuclear factor I C: encodes a dimeric DNA-binding protein, part of the CTF/NF-I family, functioning as transcription factors and as replication factors for adenovirus DNA replication.	FALSE	The NFI family may have roles in various cancers, as both tumour suppressors and oncogenes. *NFIC* is upregulated in in kidney chromophobe cell carcinoma, but downregulated in kidney papillary cell carcinoma (Li *et al.*, 2020).

Wu dORF	Length	3' UTR Location	NCBI Gene Description	RP read aligned	Association of gene with cancers
2043 ENST00000593274 C19orf53	201	47:247	Chromosome 19 open reading frame 53: no summary given	TRUE	N/A
2082 ENST00000612661 MARCKS	108	1046:1153	Myristoylated alanine rich protein kinase C substrate: encodes a substrate for protein kinase C that localizes to the plasma membrane and crosslinks actin filaments. Phosphorylation by protein kinase C or binding to calcium-calmodulin inhibits acting and membrane association. Suggested role in cell motility, phagocytosis, membrane trafficking and mitogenesis.	TRUE	MARCKS is suggested to be involved in kidney cancer growth. *MARCKS* upregulation is suggested to promote angiogenesis and growth in renal cell carcinoma, potentially involved upstream of the AKT/mTOR pathway, possibly involving vascular endothelial growth factor-A (Chen *et al.*, 2017).
142 ENST00000238618 ACYP1	96	124:219	Acylphosphatase 1: encodes a small cytosolic enzyme that catalyzes the hydrolysis of the carboxyl-phosphate bond of acylphosphates.	FALSE	*ACYP1* is overexpressed across cancers, suggested to be involved in tumorigenesis and progression, with high expression correlating with a poor prognosis in most tumor types (Zhou *et al.*, 2022a; Wang *et al.*, 2023).
209 ENST00000252804 PXDN	147	58:204	Peroxidasin: encodes a heme-containing peroxidase that is secreted into the extracellular matrix, involved in extracellular matrix formation, and may function in the physiological and pathological fibrogenic response in fibrotic kidney.	FALSE	PXDN suggested to promote cancer cell invasion, angiogenesis, and metastasis (Wyllie, Panagopoulos and Cox, 2023).
584 ENST00000296581 LSM6	99	139:237	LSM6 homolog, U6 small nuclear RNA and mRNA degradation associated: contains the Sm sequence motif, consisting of 2 regions with a linker that folds as a loop. Suggested to form a stable heteromer found in tri-snRNP particles, important for pre-mRNA splicing.	FALSE	N/A

Wu dORF	Length	3’ UTR Location	NCBI Gene Description	RP read aligned	Association of gene with cancers
849 ENST00000326005 OAZ2	201	13:213	Ornithine decarboxylase antizyme 2: belongs to the ornithine decarboxylase antizyme family, with a role in cell growth and proliferation by regulating intracellular polyamines. Antizyme expression requires +1 ribosomal frameshifting, enhanced by high levels of polyamines. Antizymes bind and inhibit ornithine decarboxylase (ODC), the key enzyme in polyamine biosynthesis. This gene encodes antizyme 2 with broad tissue distribution, inhibits ODC activity and polyamine uptake, and stimulates ODC degradation.	TRUE	N/A
1663 ENST00000401089 SZRD1	174	1:174	SUZ RNA binding domain containing 1: no summary given.	FALSE	SZRD1 may be a tumour suppressor in cervical cancer and is downregulated in many tumours. SZRD1 suppresses growth by downregulating ERK1/2, AKT and STAT3 phosphorylation, inducing G2/M cell cycle arrest and apoptosis by inducing P21 expression (Zhao *et al.*, 2017).
1711 ENST00000414487 SNRPE	117	115:231	Small nuclear ribonucleoprotein polypeptide E: encodes a core component of U small nuclear ribonucleoproteins which functions in processing the 3’ end of histone transcripts, forming part of the pre-mRNA processing spliceosome.	FALSE	N/A
1770 ENST00000434618 TAPBP	90	1948:2037	TAP binding protein: encodes a transmembrane glycoprotein which mediates interaction between newly assembled major histocompatibility complex (MHC) class I molecules and the transporter associated with antigen processing (TAP), this protein and up to four complexes of MHC class I can bind to one TAP molecule.	FALSE	N/A

Wu dORF	**Length**	**3' UTR Location**	**NCBI Gene Description**	**RP read aligned**	**Association of gene with cancers**
1833 ENST00000471115 C1orf52	195	7:201	Chromosome 1 open reading frame 52: Enables RNA binding activity. Located in nucleoplasm.	TRUE	N/A
1975 ENST00000555028 LIN52	222	19:240	lin-52 DREAM MuvB core complex: Predicted to be involved in transcription, DNA-templated, and part of the DRM complex. Predicted to be located in nucleoplasm.	TRUE	N/A
2021 ENST00000588572 COX6B2	249	112:360	Cytochrome c oxidase subunit 6B2: Predicted to be involved in oxidative phosphorylation. Located in mitochondrial crista.	FALSE	COX6B2, through enhancing oxidative phosphorylation function, may drive metastasis in pancreatic ductal adenocarcinoma (Nie *et al.*, 2020).

Alongside the details included in Table 3.8 about MSVW dORFs, Table 3.9 includes the data from investigation of MSVW dORF-containing transcript translational regulation across healthy and tumour datasets and cancer datasets under stressful conditions. Eleven MSVW dORFs had data from the transcript translational regulation analysis and the mean Log2 fold change in CDS expression from RNAseq to RP datasets, a model of translational regulation. Table 3.9 shows the varied potential translational regulation of MSVW dORF transcripts across, and within, different cell and tissue types with relative translation suggested to be increased and decreased. Most MSVW dORF transcripts, except two, suggest there was increased relative translation in the healthy kidney tissue datasets (Table 3.9). However, only four of these transcripts suggested reduced dORF activity in the tumour sample datasets, which was the overall trend seen in previous data (Table 3.9). When comparing the mean healthy tissue relative translation to the tumour tissue and other cancer cell datasets, three transcripts suggested the relative translation reduced, three transcripts suggested it increased, and five transcripts showed varied potential translational regulation across the different datasets (Table 3.9). No clear patterns emerged when comparing the potential translational regulation of these transcripts in the fed and starved conditions and the control and knockdown cancer cell datasets (Table 3.9). These results show the potential variation in MSVW dORF activity across different cell and tissue types, and that individually some MSVW dORFs followed the previous results and others did not. Most of these shortlisted MSVW dORFs did show potential translational upregulation of their transcript across the various datasets, matching the described Wu dORF function (Wu *et al.*, 2020b).

Table 3.9: Summary of the mean translational regulation data for the Wu downstream open reading frame (dORF) shortlist with mass spectroscopy (MS) validation of encoded protein (MSVW dORFs). *Data from the PRJNA256316, PRJNA880902 and PRJNA532400 datasets were used. Relative translation of transcript calculated as Log2 fold change in expression from RNAseq to ribosome profiling (RP) dataset for each transcript. For each MSVW dORF, the mean (reported to two decimal places) for the datasets from different treatments/conditions in each Sequence Read Archive (SRA) project are included. The colours range from the lowest value in the darkest red, becoming more yellow as values approach zero, and then from zero positive values become greener as values approach the highest value in the darkest green.*

SRA Project	**PRJNA256316**		**PRJNA880902**		**PRJNA532400**	
Wu dORF	**Relative Translation Healthy Kidney Tissue Mean**	**Relative Translation Kidney Tumour Tissue Mean**	**Relative Translation RKO cells Fed Mean**	**Relative Translation RKO cells Starved Mean**	**Relative Translation A549 Control Mean**	**Relative Translation A549 FKBP10 KD Mean**
3_ENST00000006015 HOXA11	2.32	N/A	1.09	1.75	N/A	1.28
22_ENST00000175756 PTPN18	0.00	1.16	0.22	0.75	-0.05	0.76
209_ENST00000252804 PXDN	-0.05	0.86	N/A	N/A	0.11	0.61
226_ENST00000256015 BTG1	3.41	2.98	2.43	2.65	3.93	3.55
849_ENST00000326005 OAZ2	2.29	3.10	N/A	3.03	N/A	N/A
1400_ENST00000375446 NINJ1	1.81	1.61	0.58	0.76	1.57	1.75
1408_ENST00000375856 IRS2	0.46	1.00	0.39	0.38	2.66	2.38
1711_ENST00000414487 SNRPE	3.91	2.94	3.39	3.33	5.39	5.69
1833_ENST00000471115 C1orf52	0.58	0.95	1.31	1.04	-0.68	-0.37
1849_ENST00000483063 POLE4	1.88	2.25	1.51	1.55	2.62	2.82
2082_ENST00000612661 MARCKS	4.26	3.84	-0.74	-0.08	2.75	2.52

3.9 Discussion

Downstream open reading frames (dORFs) are 3' UTR regulators suggested to be translated, leading to increased CDS translation (Wu *et al.*, 2020b). Although the dORF mechanism of action is unknown. Similar 5' UTR, or long non-coding RNA (lncRNA), short ORFs (sORFs) have function in a range of diseases, such as cancer (S. Wu *et al.*, 2020) and immunology (Niu *et al.*, 2020), associated with their translation, not the encoded protein (Barbosa, Peixeiro and Romão, 2013; Couso and Patraquim, 2017). Dysregulated translation and altered 3' UTR processing are associated with cancer (Mayr and Bartel, 2009; Singh *et al.*, 2009), with the regulation of many cancer-associated processes linked to sORF encoded microproteins (Merino-Valverde, Greco and Abad, 2020). Some sORFs have importance in cancer cell survival (Prensner *et al.*, 2021). The importance of other translational regulators in cancer and the altered activity of some upstream ORFs (uORFs) in cancer (Young and Wek, 2016; Sendoel *et al.*, 2017), highlights the potential importance of dORFs in cancer. Bioinformatic analysis was used to investigate dORF function and translation, contrasting this in healthy and cancerous tissue.

The bioinformatic translational regulation analysis had some assumptions. The model for translational regulation of a transcript, or CDS, used the differential expression of a transcript, or CDS, from paired RNAseq to ribosome profiling (RP) datasets, removing the influence of transcriptional regulation. A similar method was used by Wu *et al.* (2020b) when presenting 'Wu dORFs'. Initiation is assumed to be the rate-limiting translation step and transcript, or CDS, RP dataset expression is representative of their translation. Although other factors, such as elongation regulation, could impact ribosomal association with transcripts. This analysis could not determine where translational regulation comes from; other regulators such as micro-RNAs (miRNAs) or uORFs could also act on transcripts.

The translational regulation analysis was used to look at transcripts with and without Wu dORFs as whole groups, rather than at individual dORFs, meaning results for individual dORFs could be missed or masked and the conclusions are generalised across dORFs. Choosing to only include transcripts with RefSeq mRNA IDs increased transcript annotation accuracy, although 72 Wu dORFs were excluded. The small changes in the workflows for the RNAseq (Figure 2.2) and RP (Figure 2.1) datasets, particularly aligning to CDSs for RP datasets and whole transcripts for RNAseq datasets, ensured the transcriptional and translational expression were represented for transcripts. CDSs are the relevant region for translation of transcripts, preventing 5’ UTRs and 3’ UTRs from influencing the transcript expression values, even though they could align RP reads and undergo translation (Mueller and Hinnebusch, 1986; Krishna M. Vattem and Wek, 2004; Ingolia *et al.*, 2009, 2014; Guydosh and Green, 2014; Ji *et al.*, 2015; Miettinen and Björklund, 2015; Young *et al.*, 2015; Hsu *et al.*, 2016). Some Wu dORF-containing transcripts were excluded due to unavailable transcript data in the Ensembl Biomart database (Harrison *et al.*, 2024), or where transcripts were unaligned in RNAseq and/or RP datasets. This could be influenced by the dataset depth or differing patterns of transcription and translation between cell and tissue types. Alignments were carried out using HISAT2, a widely used, sensitive, fast, and splice-aware aligner suitable for these datasets (Kim *et al.*, 2019).

The results in Figure 3.2 suggest that in healthy kidney tissue, transcripts containing Wu dORFs had greater relative translation compared to transcripts without. Although supportive of dORF function (Wu *et al.*, 2020b), there is variation in the results, suggesting Wu dORF activity varies. Other regulators could influence the translational regulation of Wu dORF-containing transcripts, such as miRNAs inducing mRNA degradation (Ha and Kim, 2014). Presence of a Wu dORF in a transcript does not guarantee translational upregulation, even if it is more likely. The results for the various cancer datasets showed a potential reduction in Wu dORF activity. Kidney tumour tissue (Figure 3.3) and A549 pulmonary adenocarcinomic human alveolar basal epithelial cells (Figure 3.13) showed potentially reduced Wu dORF activity in transcripts, whereas there was little evidence of Wu dORF activity in RKO human colorectal carcinoma cells (Figure 3.16). This could mean different cancers affect dORF activity in different ways. Local conditions and cancer can influence uORF translation, allowing adaptation to these conditions through translational regulation of the CDS (Young and Wek, 2016; Sendoel *et al.*, 2017; Renz, Valdivia Francia and Sendoel, 2020). The reduced Wu dORF activity in cancer datasets could relate to local conditions within the tumour, or cancer more generally, affecting dORF translation. The A549 cell datasets also allowed comparison between control and FKBP10 knockdown, a gene involved in extracellular matrix formation and potentially tumour formation (Ishikawa *et al.*, 2008; Solassol, Mange and Maudelonde, 2011; Yao *et al.*, 2011). However, FKBP10 knockdown did not appear to influence Wu dORF activity. The inclusion of stressful conditions of arginine starvation on RKO cells also seemed to have no effect on dORF activity.

Another possible explanation for reduced and varying dORF activity in cancer involves altered 3’ UTR processing in cancers, such as truncation or rearrangement (Kuersten and Goodwin, 2003; Sandberg *et al.*, 2008; Mayr and Bartel, 2009; Singh *et al.*, 2009). The analysis into the tumour tissue transcriptome compared whether full Wu dORFs sequences appeared or not (Wu *et al.*, 2020b). Although small mutations in Wu dORFs may not disrupt the ORF, meaning they could still be translated and function. The overall dORF activity change between healthy and tumour tissue results was not explained by changes in the tumour transcriptome (Figures 3.8 and 3.9), but on an individual dORF basis this could remove a dORF, and its regulatory effect, from a transcript. dORF sequences are impacted by changes in 3’ UTR processing in cancer, which should be considered in future investigations of dORFs in cancer. The altered transcript processing in cancers can also affect the 5’ UTR and CDS (Kuersten and Goodwin, 2003; Sandberg *et al.*, 2008; Mayr and Bartel, 2009; Singh *et al.*, 2009). Rearrangements and mutations can affect 5’ UTR regulatory elements, such as introducing or removing uORFs, which could also influence the translation of the transcripts. The CDS can also be affected, leading to alternative transcripts which can also affect transcript translation when compared to the unmodified transcripts in the healthy tissue. To investigate whether changes in the 5’ UTR or CDS were influential, further exploration of the tumour transcriptome in the future would be useful.

The correlation data compared transcript 3’ UTR expression in RP datasets against CDS RP expression and the potential translational regulation of transcripts. One consideration is that 3’ UTR RP expression was only calculated for transcripts with 3’ UTR aligned RP reads and details about 3’ UTR length included in the Ensembl Biomart database (Harrison *et al.*, 2024). Importantly, correlations do not indicate causation, only suggesting possible relationships. The correlation between 3’ UTR ribosome occupancy and the potential translational regulation of transcripts was very small and varied across healthy and cancer RP datasets, whether transcripts contained Wu dORFs or not (Figures 3.4, 3.5, 3.14, 3.17). Overall, there was little evidence that with increased 3' UTR ribosomal association, and possible translation, transcripts were undergoing greater relative translation. This does not support the suggestion that dORF translation in the 3' UTR leads to increased translation of the CDS of the transcript (Wu *et al.*, 2020b). By contrast, there was a consistent positive correlation between 3' UTR and CDS ribosome occupancy (Figures 3.6, 3.7, 3.15, 3.18). This correlation is potentially supportive of dORF function where increased 3' UTR ribosomal presence increases CDS ribosomal presence (Wu *et al.*, 2020b), especially considering these correlations were often slightly greater in transcripts with Wu dORFs. Instead, and perhaps more likely, it could be that increasing CDS ribosomal presence leads to increased 3' UTR ribosomal presence. This could involve stop codon readthrough or issues with ribosomal recycling (Doronina and Brown, 2006; Namy and Rousset, 2010).

The translational regulation and correlation analysis results from human kidney tumour sample 3 datasets were more similar to healthy tissue results. This could indicate an issue with the original sampling of tumour tissue; however, this does not explain the reduced mRNA alignments. There could have been sample processing issues with possible ribosomal, or subunit, contamination with the most overrepresented sequences aligning to 28S rRNA. This could affect the number of reads included and mRNA alignment. One future option would be to explore the 28S ribosomal subunit region where the alignment occurs to investigate whether this is on the subunit surface and could be exposed.

To investigate RP read alignment to Wu dORFs, CDSs, and 3' UTRs, the NCBI BLAST+ blastn tool (Altschul *et al.*, 1997; Camacho *et al.*, 2009; Cock *et al.*, 2015) was used, providing an easily accessible, adjustable tool which quantified precise alignment locations. The different regions provided comparisons between translated regions, CDSs, generally untranslated regions, 3' UTRs, and Wu dORFs. RP read alignments are assumed to show ribosomal association and possible translation (Ingolia *et al.*, 2009). However, the data generated raised questions about whether all read alignments indicate translation. Different RP reads from the same dataset can align to precisely the same location (Tables 3.3-3.6). Reads could be identical if ribosomes were translating the same region on different copies of the same transcript. However, the repeated read alignments can be more than a few reads. The issue with these reads is where there are large numbers of RP reads aligned to the same location; it is unlikely be from a translating ribosome. Instead, these may be stalled ribosomes or another artefact. If these sites do indicate stalled ribosomes this would mean ribosomes would not translate beyond these sites, potentially preventing translation of some dORFs. Further investigation into these potential stall sites could be done to establish whether ribosomes are found downstream of these sites.

Table 3.2 showed that considerably more ribosomes associated with CDSs of transcripts with RefSeq mRNA IDs than 3' UTRs or Wu dORFs. This implied ribosomal association with, and possible translation, of some 3' UTRs and Wu dORFs. The results support the expectation that most translation occurs in CDSs and a possible smaller amount in 3' UTRs (Ingolia *et al.*, 2009; Guydosh and Green, 2014; Ji *et al.*, 2015; Miettinen and Björklund, 2015; Young *et al.*, 2015; Hsu *et al.*, 2016). Comparatively, very few RP reads align to Wu dORFs, explaining why dORFs may have been disregarded or missed previously. More reads aligned to all regions in tumour tissue datasets, implying greater ribosomal association and possible translation (Ingolia *et al.*, 2009).

The read density analysis considered the varying size of different mRNA regions and RP datasets, providing an improved indication of ribosomal association. Although there is debate, 3' UTR RP reads have been suggested to show ribosomal presence but not translation (Guydosh and Green, 2014; Miettinen and Björklund, 2015). The greatest read density was in CDSs, followed by Wu dORFs, then 3' UTRs (Figure 3.10). This could indicate that some translation occurred in Wu dORFs, especially compared to 3' UTRs. However, only a small proportion of RP reads aligned to Wu dORFs and many dORFs were unaligned. Read density was increased in all regions in tumour datasets compared to healthy RP datasets (Figure 3.10). As mentioned previously dORF function is linked to their translation (Wu *et al.*, 2020b). However, the results suggest that greater dORF translation in the tumour tissue is seen alongside reduced dORF activity. It could be that in tumour tissue dORF function is reduced regardless of dORF translation, but this would need to be explored further.

RP reading frame analysis searched for further evidence of dORF translation. The CDS preference for reading frame A modelled reading frame preference for translating ribosomes (Figure 3.12). The repeated RP alignments affected the 3’ UTR results, skewing the results towards reading frame A due to large numbers of repeated reads aligned in this reading frame, which may be due to stalled ribosomes or another artefact. Using the subsets of RP reads avoided these potential issues. In these RP subsets, 3’ UTRs appeared to be untranslated as there was no reading frame preference with ribosomes distributed evenly across the reading frames (Figure 3.12). This also means that readthrough does not generally explain 3’ UTR RP reads, as ribosomes would continue in the CDS reading frame into the 3’ UTR, maintaining the CDS preference (Doronina and Brown, 2006; Namy and Rousset, 2010). Some of the ribosomes associated with Wu dORFs may be translating, due to the preference for reading frame A compared to 3’ UTRs, however not to the extent of CDSs (Figure 3.12). The small number of reads aligned to Wu dORFs means that a few reads in a different frame could have a large impact on overall preferences.

The potential dORF translation and function found previously led to investigation into individual dORFs with strongest evidence of translation. Across healthy kidney RP datasets not many Wu dORFs had evidence of translation, requiring RP read alignment and similar reading frame preference to CDSs (Table 3.7). There is little evidence that Wu dORFs are translated consistently in these RP datasets. The small size of Wu dORFs makes it less likely for multiple ribosomes to associate due to the size of ribosomal footprints. Greater RP dataset read depth could increase the number of RP reads aligned to Wu dORFs or the consistency of alignment across datasets. The small size also means that dORF translation would be relatively rapid compared to CDSs, reducing the duration of ribosome association. These challenges indicate the difficulties of using RP datasets to investigate dORF translation.

The shortlisted MSVW dORFs (MS validated Wu dORFs) with MS validated encoded proteins show dORF translation. MSVW dORFs still had varied RP read alignment in the datasets and translational regulation results, further evidence of variable dORF activity. Alternatively, some MSVW dORFs could encode functional proteins. The genes containing MSVW dORFs have varied functions, and some have importance in cancers as tumour suppressors, oncogenes, or both. If dORFs are translational regulators appearing in genes associated with important processes, for example the *HOXA11* MSVW dORF, a gene suggested to be involved in transcription, by increasing the protein production from these genes the dORFs may be important in processes such as transcription. This could be used when considering future cancer therapies, such as the MSVW dORF in *NINJ1*, which is suggested to be involved in tumour development (Cho *et al.*, 2013). This dORF is suggested to increase transcript translation, observed in the translational regulation analysis, meaning disruption of this dORF in cancer could reduce *NINJ1* translation and potentially tumour development.

3.10 Conclusions

The overall findings support dORFs as translational regulators, increasing translation of transcripts they appear in, especially in healthy tissue. dORF activity is shown to be reduced in cancer when compared to healthy datasets. Altered 3' UTR sequences in the tumour transcriptome that disrupt dORFs do not appear to explain this altered activity in cancer. There is no biologically meaningful correlation between the number of ribosomes associated with the 3' UTR and the potential translational regulation of the transcript. There is a positive correlation between the number of ribosomes associated with the 3' UTR and CDS of a transcript across healthy and cancer datasets. Ribosomes associate with dORFs, but this analysis has not definitively shown dORF translation. The analysis also showed ribosomal association with 3' UTRs and suggested that these ribosomes were not translating. Ribosomal association with CDSs, 3' UTRs and dORFs was suggested to increase in the tumour tissue. MS validated proteins have shown that some dORFs are translated and some of these translated dORFs associate with ribosomes and show potential regulatory activity. These results raise interesting questions about dORF function in cancer, whether dORFs are translated, and a possible change in ribosomal association with 3' UTRs in cancer.

Chapter 4: dORF Composition is More Comparable to 3' UTRs than CDSs

4.1 Introduction

The previous chapter provided support for downstream open reading frames (dORFs) acting as translational regulators (Wu *et al.*, 2020b), increasing relative translation of transcripts containing Wu dORFs, dORFs presented by Wu *et al.* (2020b). Ribosomes associated with 3' untranslated regions (UTRs) and although some aligned to Wu dORFs, many 3' UTR aligned RP reads aligned outside of Wu dORFs. It is likely that there are more dORFs outside of Wu dORFs, suggested by these 3' UTR aligned RP reads and sORFs being widespread with underestimated abundance (Slavoff *et al.*, 2013; Bazzini *et al.*, 2014; Couso and Patraquim, 2017; Lu *et al.*, 2019; van Heesch *et al.*, 2019; Chen *et al.*, 2020; Ouspenskaia *et al.*, 2020; Ruiz Cuevas *et al.*, 2021). One difficulty in identifying more dORFs is that sORF transcription does not guarantee function and sORFs could be randomly generated in a genome (Guttman *et al.*, 2013; Couso and Patraquim, 2017). The Wu dORFs have no reported surrounding 3' UTR motifs and little discussion of Wu dORF composition other than a sORF sequence with ribosomal association (Wu *et al.*, 2020b). Understanding Wu dORF composition better could help identify other similar dORFs or differentiate these Wu dORF from other, nonfunctional, or randomly generated, dORF sequences. The 3' UTR composition is understood better and can influence regulatory elements. 3' UTR composition, especially when influencing the structure, is associated with translation termination (Zahdeh and Carmel, 2016). The 3' UTR has reduced GC composition (Pesole et al., 2001; Larizza et al., 2002; Mignone et al., 2002), possibly due to increased tolerance of mutations which reduce GC content caused by methylation of CpG residues or CpG islands (Ehrlich and Wang, 1981; Karlin and Mrázek, 1997; Takata *et al.*, 2017), grouped CpG dinucleotides with a high proportion of GC nucleotides (Esteller, 2008; Portela and Esteller, 2010; Jang *et al.*, 2017). Another explanation for 3' UTR aligned RP reads could be stop codon readthrough, where translation continues into 3' UTRs. This can be affected by the stop codon, UGA is the most common for readthrough and UAA is the least (Howard *et al.*, 2000; Manuvakhova, Keeling and

Bedwell, 2000; Bidou *et al.*, 2004; Floquet *et al.*, 2012; Wangen and Green, 2020). Wu dORFs-containing transcripts could be more likely to undergo readthrough based on the CDS stop codon, leading to 3' UTR ribosomal association. The hypotheses for this chapter are that potential dORFs will be abundant in 3' UTRs and some of these will associate with ribosomes and share the Wu dORF regulatory function, with reduced activity in cancer datasets. Additionally, the nucleotide composition of Wu dORFs will differ from 3' UTRs and appear more like a mini CDS, due to the suggestion that they are translated. The composition of Wu dORF-containing 3' UTRs will differ from more general 3' UTR composition.

This chapter contains the results from bioinformatic analyses used to investigate the abundance of potential dORF sequences (referred to as 'Potential dORFs' throughout this thesis), ribosomal association with 'Potential dORFs', potential translational regulation of 'Potential dORFs' with ribosomal association (referred to as 'RP Potential dORFs' throughout this thesis), and the composition of Wu dORFs compared to 3' UTRs and CDSs. The analysis of the abundance of 'Potential dORFs' in 3' UTRs compared to shuffled 3' UTRs used methods described in section 2.7. There is also consideration of 3' UTRs with ribosomal association in the PRJNA256316 RP datasets from healthy human kidney and kidney tumour samples. These RP datasets were then used to investigate ribosomal association with 'Potential dORFs', and potential translational regulation of 'RP Potential dORFs'. Then correlations between the relative translation of transcripts containing 'RP Potential dORFs' and the extent of ribosomal association with the 'RP Potential dORF' were analysed. This analysis sought to expand the number of dORFs identified and gather further evidence of dORF translation and function. The methods described in section 2.8 were then used to analyse the nucleotide, dinucleotide, trinucleotide, and codon composition of Wu dORFs and 3' UTRs containing these dORFs. This composition was compared to shuffled 3' UTRs, 3' UTRs and CDSs. The genome wide and Wu dORF transcript CDS stop codon preference were also compared. The analysis aimed to improve understanding of Wu dORF composition and assist in identifying similar dORFs in the future.

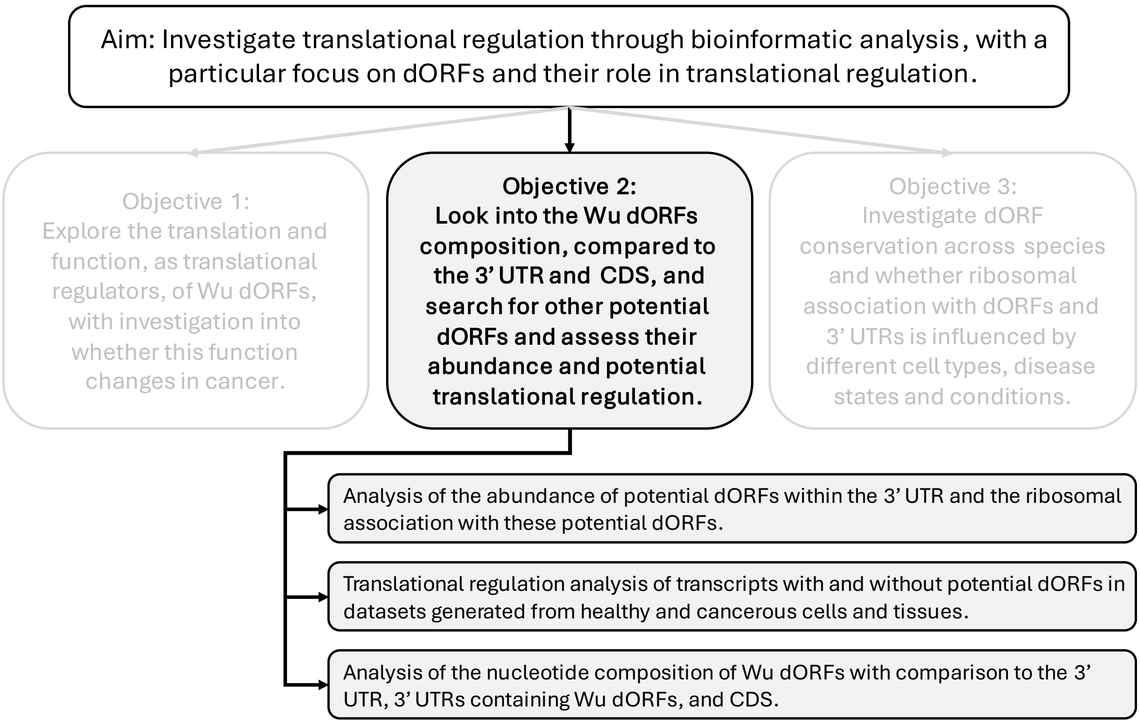
Aim: Investigate translational regulation through bioinformatic analysis, with a particular focus on dORFs and their role in translational regulation.
Objective 1: Explore the translation and function, as translational regulators, of Wu dORFs, with investigation into whether this function changes in cancer.
Objective 2: Look into the Wu dORFs composition, compared to the 3' UTR and CDS, and search for other potential dORFs and assess their abundance and potential translational regulation.
Objective 3: Investigate dORF conservation across species and whether ribosomal association with dORFs and 3' UTRs is influenced by different cell types, disease states and conditions.
Analysis of the abundance of potential dORFs within the 3' UTR and the ribosomal association with these potential dORFs.
Translational regulation analysis of transcripts with and without potential dORFs in datasets generated from healthy and cancerous cells and tissues.
Analysis of the nucleotide composition of Wu dORFs with comparison to the 3' UTR, 3' UTRs containing Wu dORFs, and CDS.

Figure 4.1: Summary of approaches used in section 4 to meet the chapter objectives and overall study aim. *dORF – downstream open reading frame, 3' UTR – 3' untranslated region, CDS – coding sequence.*

4.2 Coverage of 3' UTR by 'Potential dORFs' and proportion of RP reads aligned to these dORFs

The abundance of 'Potential dORFs', in terms of the number in 3' UTRs and the proportion of 3' UTRs covered by these dORFs, was investigated to find more potentially functional dORFs. dORFs could be generated by random mutation in 3' UTRs without necessarily conferring function (Guttman *et al.*, 2013; Couso and Patraquim, 2017). Including comparison between genomic 3' UTRs and randomly shuffled versions of these 3' UTRs with sequences of equivalent nucleotide content and lengths, allows consideration of the 3' UTR dORF abundance which could occur by chance. This abundance analysis was also used with RP read aligned 3' UTRs in the PRJNA256316 RP datasets to consider whether 'Potential dORF' abundance changes when 3' UTRs are associated with ribosomes. When considering 3' UTRs with RefSeq mRNA IDs, 'Potential dORF' sequences are very abundant and cover 84.2% of these 3' UTRs (Table 4.1). In contrast when looking across the three shuffled 3’ UTR versions, maintaining nucleotide composition and length, fewer ‘Potential dORF’ sequences were found. In all three shuffled versions there were more than 35000 fewer dORF sequences (Table 4.1). There was 6.77%, 6.84% and 6.62% fewer dORFs in the shuffled 3’ UTRs compared to genomic 3’ UTRs. Unsurprisingly, the coverage of ‘Potential dORFs’ in the shuffled 3’ UTR sequences was reduced to 78.4%, 78.4% and 78.5% coverage (Table 4.1). When considering RP read aligned 3’ UTRs the number of 3’ UTRs reduced from 27134 to 10225 sequences. In these RP read aligned 3’ UTRs ‘Potential dORFs’ were more abundant in genomic 3’ UTRs compared to shuffled versions, and in all RP read aligned 3’ UTRs more dORFs were found per 3’ UTR (Table 4.1). The coverage by ‘Potential dORFs’ remained similar whether 3’ UTRs were RP read aligned or not, 84.7% in genomic RP read aligned 3’ UTRs, reduced to 79.5%, 79.5% and 79.4% in the shuffled RP read aligned 3’ UTRs (Table 4.1). Whether an RP read aligned or not, the abundance of ‘Potential dORFs’ in number and coverage reduced in randomly shuffled 3’ UTR versions, suggesting dORF presence does not solely rely on chance. ‘Potential dORFs’ are very abundant in 3’ UTRs, but they can appear by chance.

Table 4.1: Potential downstream open readings frame (dORFs) are more common and cover a slightly greater proportion of 3' untranslated regions (UTRs) in genomic sequences compared to shuffled sequences. *Comparing the number of 'Potential dORFs' and the proportion of 3' UTRs covered by these dORFs in genomic and shuffled 3' UTR sequences. 3' UTRs with RefSeq mRNA IDs and those 3' UTRs with ribosome profiling (RP) reads aligned in one or more of the PRJNA256316 RP datasets were used. Tables include the total 3' UTR and 'Potential dORF' sequences and their length with the percentage total coverage of 3' UTRs by 'Potential dORFs' in genomic and shuffled 3' UTR sequences, separated into all RefSeq 3' UTRs and those with RP read alignment.*

All RefSeq 3' UTR

Genomic or Shuffled sequences	Total 3' UTR Length	3' UTR Sequences	Total 'Potential dORF' length	'Potential dORF' Sequences	% 'Potential dORF' Coverage
Genomic	33440369	27134	28172736	535907	84.2
Shuffle 1	33440369	27134	26226758	499619	78.4
Shuffle 2	33440369	27134	26228799	499274	78.4
Shuffle 3	33440369	27134	26237265	500430	78.5

RP read aligned RefSeq 3' UTR

Genomic or Shuffled sequences	Total 3' UTR Length	3' UTR Sequences	Total 'Potential dORF' length	'Potential dORF' Sequences	% 'Potential dORF' Coverage
Genomic	23095652	10225	19551733	370025	84.7
Shuffle 1	23095652	10225	18358592	348693	79.5
Shuffle 2	23095652	10225	18351766	348198	79.5
Shuffle 3	23095652	10225	18331743	348416	79.4

Investigation into which ‘Potential dORFs’ had ribosomal association, suggesting possible translation, known as ‘RP Potential dORFs’, was done to determine which of the highly abundant ‘Potential dORFs’ could be functional, considering that dORFs could appear by chance (Couso and Patraquim, 2017). The number of ‘RP Potential dORFs’ and the proportion of 3’ UTR aligned RP reads that aligned to ‘RP Potential dORFs’ was investigated. This also explored whether other dORFs could explain the 3’ UTR aligned RP reads not found to align to Wu dORFs. The PRJNA256316 healthy kidney and kidney tumour sample RP datasets and the subsets of RP reads, outlined in section 3.6, were used. RP read subsets reduced the influence of repeated RP reads alignments, by only including reads in the subset if there are ten, five, or one, read(s) aligned to the same location. ‘RP Potential dORFs’ were found across the RP datasets and the RP read subsets had little impact on the number of ‘RP Potential dORFs’ found (Table 4.2). Very few ‘Potential dORFs’ in RP read aligned 3’ UTRs (Table 4.1) had RP read alignment (Table 4.2). Across the RP datasets there were at most 16263 ‘RP Potential dORFs’, compared to 370025 ‘Potential dORFs’ found in RP read aligned 3’ UTRs. This suggests a huge number of ‘Potential dORFs’ were unaligned. Although RP dataset size varied, across the RP read subsets there were more ‘RP Potential dORFs’ in tumour tissue sample datasets compared to healthy kidney sample datasets (Table 4.2). The occurrences of large numbers of repeated RP alignments, which were likely artefacts, meant inconsistent proportions of 3’ UTR aligned RP reads aligned to ‘Potential dORFs’, ranging from around 58-92%, showing a large proportion of RP reads can align to dORFs, especially in tumour RP datasets (Table 4.2). Reducing the influence of these potential artefacts in the RP read subsets (10 or fewer, 5 or fewer, and unique) produced consistent results across healthy and tumour RP datasets; 66-70% of 3’ UTR aligned RP reads were found within ‘Potential dORFs’ (Table 4.2). ‘Potential dORFs’ covered 84.7% of the RP read aligned 3’ UTRs (Table 4.1), meaning the proportion of RP reads aligned to ‘Potential dORFs’ was lower than expected. A small proportion of ‘Potential dORFs’ had RP read alignment, although this did increase in tumour sample RP datasets. ‘Potential dORFs’ could explain around two thirds of 3’ UTR aligned RP reads, however ‘Potential dORFs’ cover around 85% of 3’ UTRs, meaning less RP reads aligned to the dORFs than expected. This could be because some of these ‘Potential dORFs’ are not translated, reducing the number of ribosomes associated with dORFs.

Table 4.2: Across PRNJNA256316 ribosome profiling (RP) datasets, around two thirds of 3' untranslated region (UTR) aligned RP reads aligned within potential downstream open reading frame (dORF) sequences and the tumour tissues RP datasets align to a greater number of 3' UTRs and 'Potential dORFs'. *Comparing the number of 'Potential dORFs' and RefSeq mRNA ID 3' UTRs with RP read alignment and the proportion of 3' UTR aligned RP reads that align within 'Potential dORFs' for each RP dataset and for each RP read subset. RP read subsets denote the number of RP reads allowed to align to the same location, 10 or fewer, 5 or fewer, only 1 (unique). Included in the square brackets are the number of sequences normalised to 20,000,000 RP reads in each dataset to account for the varying dataset sizes. PRJNA256316 RP datasets from healthy human kidney samples are annotated NT and RP datasets from human kidney tumour samples are annotated TT.*

		RP read subset							
		All RP reads		**10 or fewer RP reads**		**5 or fewer RP reads**		**Unique RP reads**	
RP Dataset	**3' UTR Sequences with RP read alignment**	**% 3' UTR RP reads aligned to 'Potential dORFs'**	**dORF sequences with RP read alignment**	**% 3' UTR RP reads aligned to 'Potential dORFs'**	**dORF sequences with RP read alignment**	**% 3' UTR RP reads aligned to 'Potential dORFs'**	**dORF sequences with RP read alignment**	**% 3' UTR RP reads aligned to 'Potential dORFs'**	**dORF sequences with RP read alignment**
NT1.1	3079 [3198]	57.9	4351 [4519]	68.4	4348 [4516]	67.7	4348 [4516]	66.5	4228 [4392]
NT1.2	3603 [2725]	58.4	5405 [4089]	67.5	5401 [4086]	67.1	5395 [4081]	66.5	5227 [3954]
NT2.1	3022 [3526]	74.8	4285 [5000]	67.1	4283 [4998]	66.7	4281 [4995]	65.8	4171 [4867]
NT2.2	3691 [3137]	74.9	5485 [4662]	66.7	5482 [4659]	66.2	5477 [4655]	66.0	5306 [4510]
TT1.1	4677 [4467]	90.1	8674 [8284]	69.0	8671 [8281]	68.0	8668 [8278]	67.5	8507 [8124]
TT1.2	5420 [3709]	90.0	11175 [7648]	69.8	11174 [7647]	68.7	11174 [7647]	68.3	10914 [7469]
TT2.1	5949 [4816]	94.4	12686 [10270]	67.7	12683 [10267]	66.8	12682 [10266]	66.9	12425 [10058]
TT2.2	6812 [3931]	94.4	16263 [9385]	68.5	16262 [9384]	67.4	16262 [9384]	67.1	15956 [9207]
TT4.1	4344 [4570]	91.8	8066 [8486]	69.3	8063 [8483]	68.9	8663 [9114]	68.9	7899 [8311]
TT4.2	5082 [3843]	91.6	10315 [7800]	69.1	10314 [7800]	68.7	10313 [7799]	68.8	10113 [7648]

4.3 Translational regulation of ‘RP Potential dORFs’ in PRJNA256316 RP datasets

‘RP Potential dORFs’ had RP read alignment, suggestive of ribosomal association and possible translation. These potentially translated dORFs were an expansion of the Wu dORFs. dORF translation is suggested to increased CDS translation (Wu *et al.*, 2020b), leading investigation into whether the translational regulation indicated with Wu dORFs was also present for ‘RP Potential dORFs’. Using the PRJNA256316 datasets allowed comparison of potential translational regulation in transcripts with, and without, ‘RP Potential dORFs’ in healthy and tumour datasets using the same method used in the previous chapter. Compared to the number of transcripts with ‘RP Potential dORFs’ used, a reduced number of transcripts had differential expression values reported, 1019 to 1230 transcripts in healthy tissue datasets and increased in tumour sample datasets with 1475 to 2543 transcripts (Figure 4.2). The difference between the mean Log2 fold change in transcript occurrence from RNAseq to RP datasets for transcripts with and without ‘RP Potential dORFs’ is very small across all datasets (Figure 4.2). In healthy tissue, the transcripts with ‘RP Potential dORFs’ had increased mean Log2 fold change values, by 0.1431 to 0.2774, compared to transcripts without, although only NT1.2 (P = 0.0409) and NT2.2 (P = 0.0001) were statistically significant differences (P<0.05) (Figure 4.2). This small difference and little change in the box and whisker plots suggested ‘RP Potential dORFs’ had little influence on relative translation in healthy tissue datasets. In the tumour tissue datasets there were even smaller differences, ranging from 0.08899 to -0.1364, between the mean Log2 fold changes in transcripts with and without ‘RP Potential dORFs’ with no statistically significant differences (P >0.05)(Figure 4.2). In tumour tissue, transcripts with ‘RP Potential dORFs’ showed little function as potential translational regulators. Overall, there was a slight loss of dORF activity in tumour tissue compared to healthy tissue, however, in both tissues ‘RP Potential dORFs’ showed little evidence of translational regulation.

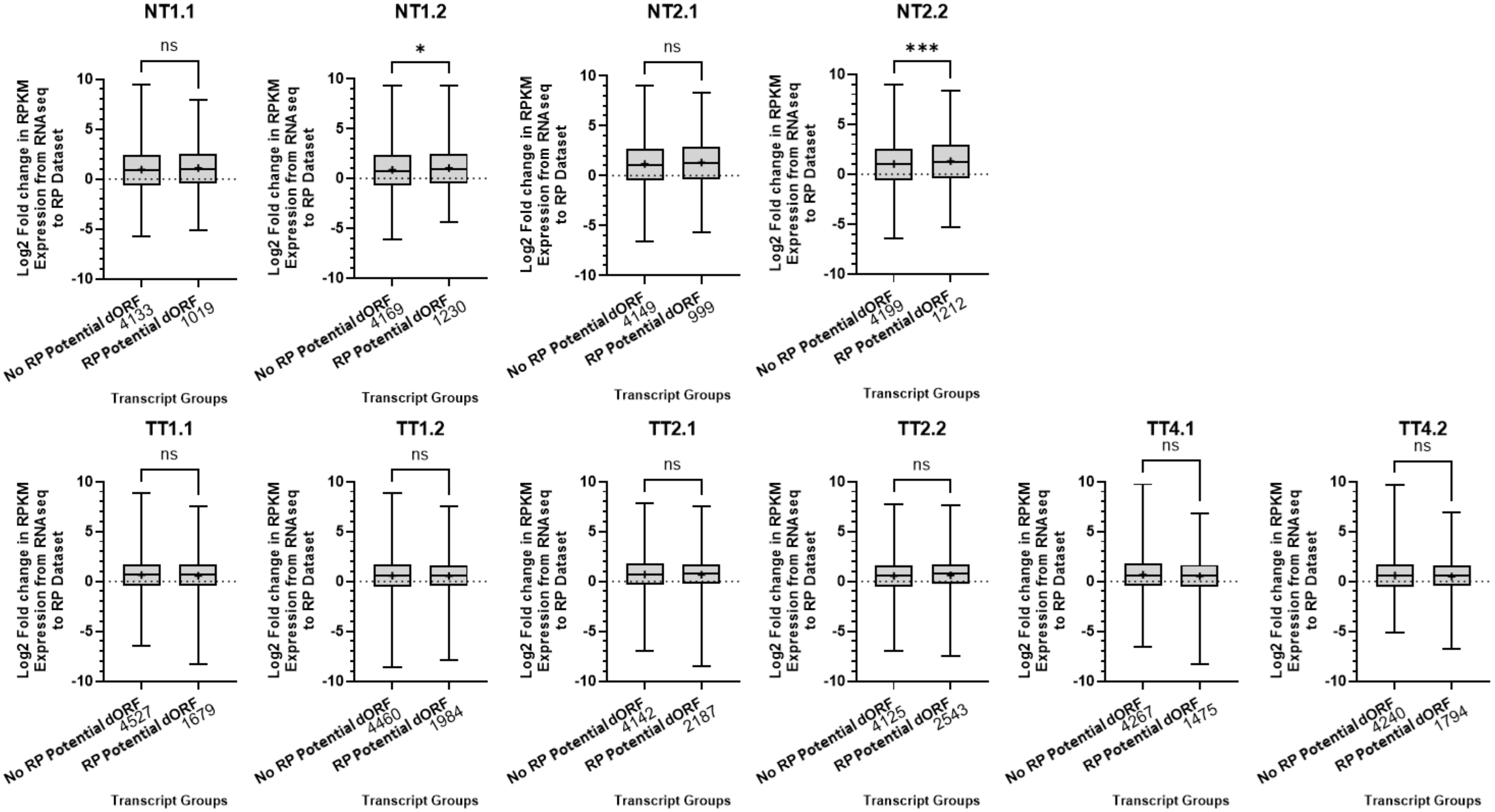

Figure 4.2: RefSeq mRNA ID transcripts with ‘RP Potential downstream open reading frames (dORFs)’ appear to show no change in translational regulation in both healthy kidney (NT) and kidney tumour samples (TT) from PRJNA256316 compared to transcripts without ‘RP Potential dORFs’. *‘RP Potential dORFs’ are potential dORF sequences with ribosome profiling (RP) read alignment. Relative transcript translation calculated as Log2 fold change in expression from RNAseq to RP dataset for each transcript. Box and whisker plots for each group of transcripts, the horizontal line within the box marks the median, the whiskers denote the maximum and minimum values, and the box marks the 25th to the 75th percentile of the values distribution, with the mean marked with ‘+’. One-Way ANOVA multiple comparisons test with Šídák adjusted P values was used to compare the means. Each chart title describes which RP dataset was used in the analysis. Transcript group sizes included in the x axis labels. ns – P>0.05, * - P<0.05, *** - P<0.0005.*

Although there was little overall evidence of ‘RP Potential dORF’ function, the box and whisker plots showed that the relative translation was increased for some individual transcripts (Figure 4.2). Some individual ‘RP Potential dORFs’ could have function, potentially driven by greater translation of these dORFs compared to ‘RP Potential dORF’ overall, especially considering dORF and sORF function more generally, is often linked to their translation (Barbosa, Peixeiro and Romão, 2013; Couso and Patraquim, 2017; Wu *et al.*, 2020b). Investigating the RP read density, or read location density, of ‘RP Potential dORFs’ and correlating this with the relative translation of the transcript could show whether ‘RP Potential dORFs’ with evidence of more ribosomal association, and possible translation, had greater dORF activity. Due to the repeating 3’ UTR RP alignment artefacts, RP read location density, based on the number of different RP read alignment locations in the dORF, was used as a better indication of possible translation. In healthy kidney and kidney tumour RP datasets there was likely no biologically meaningful correlation between the ‘RP Potential dORF’ RP read location density and the relative translation of the transcript (Figure 4.3). In healthy tissue RP datasets, Pearson r values ranged from 0.01748 to 0.06131, with between 1674 and 2278 transcripts and one statistically significant ($P<0.05$) r value (Figure 4.3). In the tumour tissue RP datasets, Pearson r values ranged from -0.04201 to 0.06966 with three statistically significant ($P<0.05$) r values, and between 3571 to 7605 transcripts included (Figure 4.3). Most ‘RP Potential dORFs’ had very low RP read location density, between one and two RP read locations per 100 bases, in healthy datasets. This increased to between one and five in tumour datasets, implying that ‘RP Potential dORF’ ribosomal association increased. More transcripts were considered in tumour tissue RP datasets correlation analysis. Some Figure 4.3 correlation charts showed vertical lines of points with the same x axis value. These occurrences either indicated different transcripts with differing RP read location density which share the same relative translation, or, perhaps more likely, these lines of points were where transcripts contain multiple ‘RP Potential dORFs’. There was no relationship found between ribosomal association with, or translation of, ‘RP Potential dORFs’ and the relative translation of transcripts containing these dORFs. ‘RP Potential dORFs’ generally had low levels of ribosomal association, although this did increase slightly in tumour tissue compared to healthy tissue.

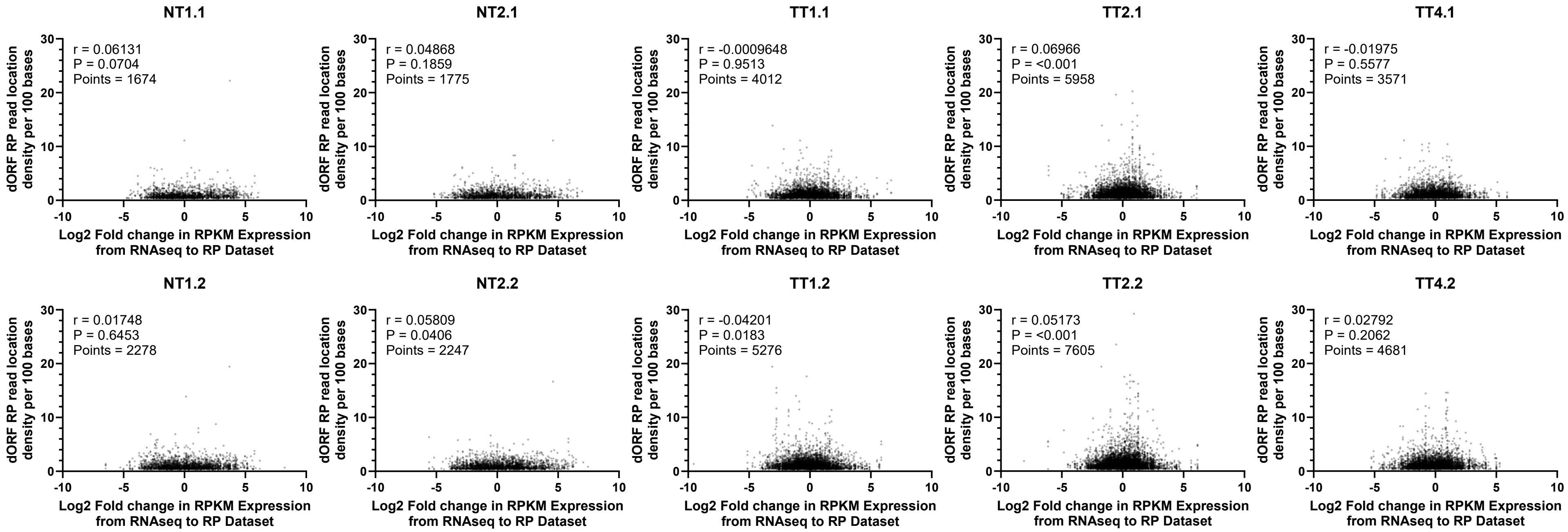

Figure 4.3: 'RP Potential downstream open reading frames' (dORFs) show no biologically meaningful correlation between the ribosome profiling (RP) read alignment location density in the dORF and the possible translational regulation of the transcript in PRJNA256316 datasets. *'RP Potential dORFs' are potential dORF sequences with RP read alignment Relative transcript translation reported as Log2 fold change in expression from RNAseq to RP dataset. RP read location density refers to the density of different RP read alignment locations within a 'RP Potential dORF'. The correlation was quantified using the Pearson correlation coefficient. Each chart title describes which PRJNA256316 RP dataset was used in the analysis. The Pearson r coefficient, P value and number of data points are included on the charts. P values adjusted for multiple comparisons using Holm-Šídák method.*

The same correlation analysis was also done with the RP read density of ‘RP Potential dORFs’ in addition to RP read location density. This was done to confirm the results were consistent whether using the RP read location or RP read density to explore the effect of ribosomal association, and suggested translation, of ‘RP Potential dORFs’. The Pearson r and P values, with the number of comparisons, quantifying the correlation between transcript Log2 fold change in expression from RNAseq to RP dataset and ‘RP Potential dORF’ RP read density were included in Table 4.3. Across the healthy and tumour datasets there were no statistically significant P values (P<0.05) and Pearson r values ranged from -0.03661 to -0.007705 (Table 4.3). These results supported those in Figure 4.3, implying there was no biologically meaningful correlation between the relative translation of transcripts and the RP read density in ‘RP Potential dORFs’. The ‘RP Potential dORFs’ as a group do not show the function seen previously with, and described for, Wu dORFs (Wu *et al.*, 2020b), even with increased potential translation of ‘RP Potential dORFs’.

Table 4.3: ‘RP Potential downstream open reading frames’ (dORFs) show no correlation between the ribosome profiling (RP) read density in the dORF and the possible translational regulation of the transcript in PRJNA256316 datasets. *‘RP Potential dORFs’ are potential dORF sequences with RP read alignment. The correlation between ‘RP Potential dORF’ RP read density and relative transcript translation, reported as Log2 fold change in expression from RNAseq to RP dataset of a transcript, was quantified using the Pearson correlation coefficient. The Pearson r coefficient, P value and number of data points are included in the tables. P values adjusted for multiple comparisons using Holm-Šídák method.*

PRJNA256316 RP Dataset	**Pearson r**	**P value**	**Number of comparisons**
NT1.1	-0.02176	0.7541	1674
NT1.2	-0.03444	0.5707	2278
NT2.1	-0.03117	0.6498	1775
NT2.2	-0.007705	0.8055	2247
TT1.1	-0.02993	0.4159	4012
TT1.2	-0.03661	0.0753	5276
TT2.1	-0.01959	0.5707	5958
TT2.2	-0.01251	0.7245	7605
TT4.1	-0.02695	0.5707	3571
TT4.2	-0.008542	0.8055	4681

4.4 Comparison of nucleotide composition of Wu dORFs with 3' UTR and CDS

The previous section suggested that 'RP Potential dORFs' do not share the potential regulatory function of Wu dORFs (Wu *et al.*, 2020b). Although dORFs may be very abundant, they may not all be functional as translational regulators (Wu *et al.*, 2020b). What is the difference between the 'RP Potential dORFs' and Wu dORFs? There is little known about Wu dORFs composition (Wu *et al.*, 2020b), meaning analysing of this could support identification of other functional dORFs. The nucleotide, dinucleotide, trinucleotide, and codon composition of Wu dORFs, and 3' UTRs containing these dORFs were analysed, with comparison to 3' UTRs and CDSs. 3' UTRs provided an untranslated comparison for Wu dORFs, with CDSs providing a translated region comparison. As whole groups the nucleotide composition of Wu dORFs was compared to that of 3' UTRs and CDSs with RefSeq mRNA IDs. Uracil (U) (30.0%) was the most common nucleotide followed closely by adenine (A) (27.5%), then with reduced guanine (G) (21.4%) and cytosine (C) (21.2%) nucleotide composition in 3' UTRs (Table 4.4). This matches the expected 3' UTR GC composition (Pesole et al., 2001; Larizza et al., 2002; Mignone et al., 2002). The CDS nucleotide composition differed from the 3' UTR with G and A nucleotides being the most common followed very closely by C, U was the least common nucleotide (Table 4.4). In Wu dORFs, U (31.1%) was the most common nucleotide, followed by A (27.9%), then G (21.0%), and C (20.1%) was the least common nucleotide (Table 4.4). The Wu dORF nucleotide composition was more similar to 3' UTRs than CDSs, with reduced G and C, and increased U nucleotide composition.

Table 4.4: Nucleotide composition of Wu downstream open reading frames (dORFs) appears to be more similar to 3' untranslated regions (UTRs) than coding sequences (CDSs). *The table provides the nucleotide composition of the RefSeq mRNA ID 3' UTR and CDS sequences to contrast against the nucleotide composition of Wu dORF sequences.*

Sequences	**Nucleotide Composition of sequences (%)**			
	A	**C**	**G**	**U**
RefSeq 3' UTR	27.5	21.2	21.4	30.0
RefSeq CDS	26.2	25.8	26.2	21.8
Wu dORF	27.9	20.1	21.0	31.1

Dinucleotide composition can have importance in regulation of sequences. An important example is the CG dinucleotide, which can affect gene expression, through methylation of CpG residues or CpG islands (Esteller, 2008; Portela and Esteller, 2010; Jang *et al.*, 2017). The results in Table 4.4 can influence dinucleotide composition due to the varied nucleotide compositions, however, preferences for particular dinucleotides could also influence the nucleotide composition. The dinucleotide composition analysis included the mean dinucleotide composition of three separate shuffles of RefSeq mRNA ID 3' UTR sequences, where the length and nucleotide content were maintained, providing a model of 3' UTR composition if left to chance. In addition to Wu dORFs, RefSeq mRNA ID 3' UTRs and CDSs, the 3' UTRs containing Wu dORFs were included to investigate whether these differed from 3' UTRs more generally. When comparing 3' UTRs to the mean shuffled 3' UTRs the proportion of most dinucleotides changed, evidenced by the differing colours in the heat map, some of the largest changes were in CG, AU, UA and UG dinucleotides, suggested by the extent of the colour change (Figure 4.4). CG results matched the expected reduced GC composition seen and described previously in 3' UTRs (Pesole et al., 2001; Larizza et al., 2002; Mignone et al., 2002). The differences across most dinucleotides implied that 3' UTR dinucleotide composition is not generated by chance. 3' UTR sequences and the Wu dORF containing 3' UTRs shared very similar dinucleotide compositions; although the UU dinucleotide was more frequent in the Wu dORF containing 3' UTRs (Figure 4.4). Wu dORF dinucleotide composition was fairly similar to the 3' UTRs with a few dinucleotides, such as AA and UG, showing some small changes (Figure 4.4). The CDSs and Wu dORF dinucleotide compositions were not very similar with many dinucleotides showing different frequencies (Figure 4.4). Wu dORF dinucleotide composition was more similar to 3’ UTRs than CDSs, supported by the statistically significant (P<0.0001) chi-square values, with a smaller chi-square value when comparing the Wu dORFs and 3’ UTRs, suggesting a smaller difference when comparing the sequences (Figure 4.4). Wu dORF dinucleotide composition was more similar to the often untranslated 3’ UTR than the translated CDS and the presence of Wu dORFs did not affect 3’ UTR dinucleotide composition much.

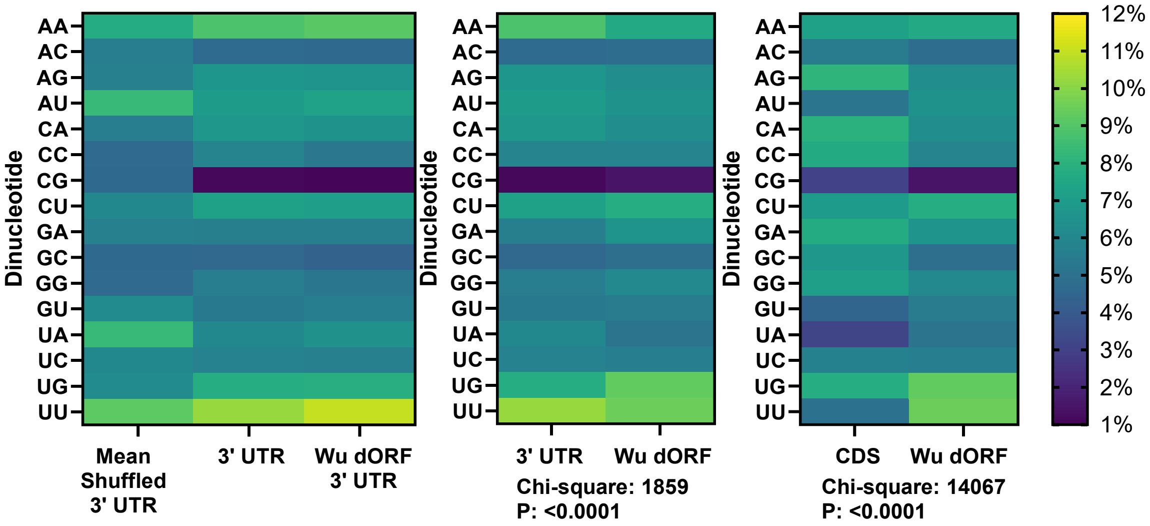

Figure 4.4: Dinucleotide composition of Wu downstream open reading frames (dORFs) appears to be more similar to 3’ untranslated regions (UTRs) than coding sequences (CDSs). *Heat maps represent the proportion (%) of sequences made up of each dinucleotide and are grouped for comparisons. The heat maps compare RefSeq mRNA ID 3’ UTR sequences against: the mean of three repeats of shuffled RefSeq 3’ UTR sequences, the RefSeq 3’ UTRs that contain Wu dORFs, and the Wu dORF sequences. The heat map on the right compares RefSeq mRNA ID CDS sequences against Wu dORF sequences. These comparisons were quantified with Chi-square goodness of fit tests comparing the difference and significance between expected and observed distributions. In each heat map for the Chi-square tests the left column of observed values represents the expected values and the right column the observed.*

The importance of the CpG dinucleotide as a site of mutation has been discussed previously in section 1.2, and CG was shown to be the least common dinucleotide in Wu dORFs and 3’ UTRs (Figure 4.4). The CG dinucleotide frequency was largely reduced, around four times, from the shuffled (4.69%) to genomic 3’ UTR sequences

(1.19%) (Figure 4.5). The CG dinucleotide frequency was very similar when comparing the overall 3' UTRs and Wu dORF-containing 3' UTRs (1.12%) (Figure 4.5). Compared to 3' UTRs the Wu dORFs (1.54%) had a very slight increase in CG dinucleotide frequency, which increased further in CDSs (3.13%) (Figure 4.5). CG dinucleotides were more than twice as frequent in CDSs than in Wu dORFs and 3' UTRs. Compared to what is expected by chance, the CG dinucleotide appeared much less frequently in 3' UTRs, and compared to the CDSs and 3' UTRs, Wu dORF CG dinucleotide frequency was more similar to 3' UTRs.

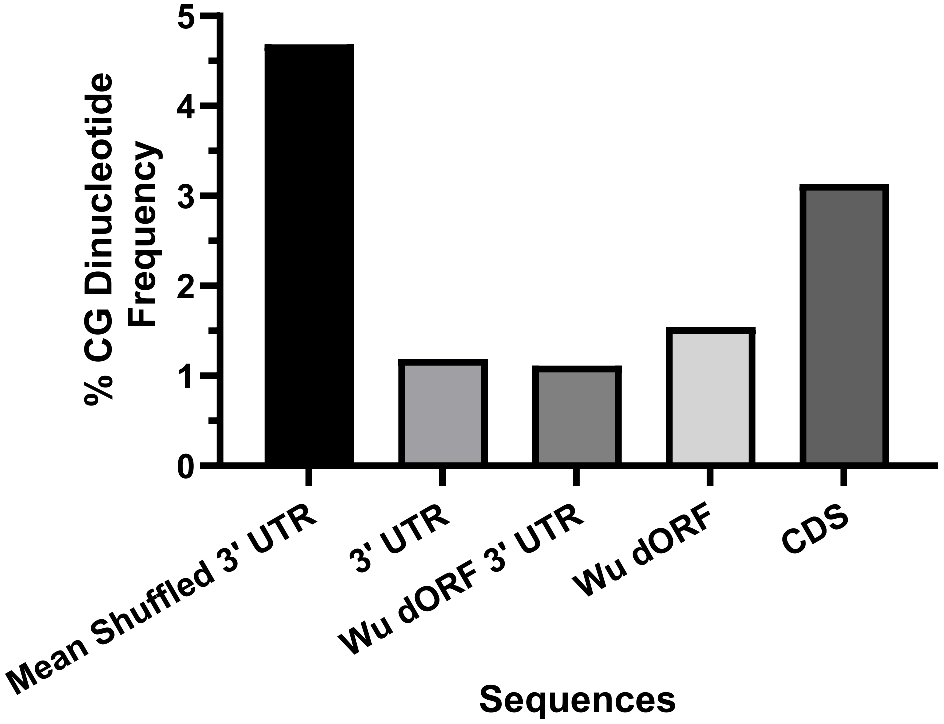

Figure 4.5: Cytosine Guanine (CG) dinucleotide frequency is greatest in shuffled 3' untranslated regions (UTRs), followed by coding sequences (CDSs), Wu downstream open reading frames (dORFs), then 3' UTRs. *The CG dinucleotide frequency is represented as a percentage of all the dinucleotides in different sequences, including: shuffled RefSeq mRNA ID 3' UTRs, RefSeq 3' UTRs, RefSeq 3' UTRs containing Wu dORFs, Wu dORFs, and RefSeq CDSs.*

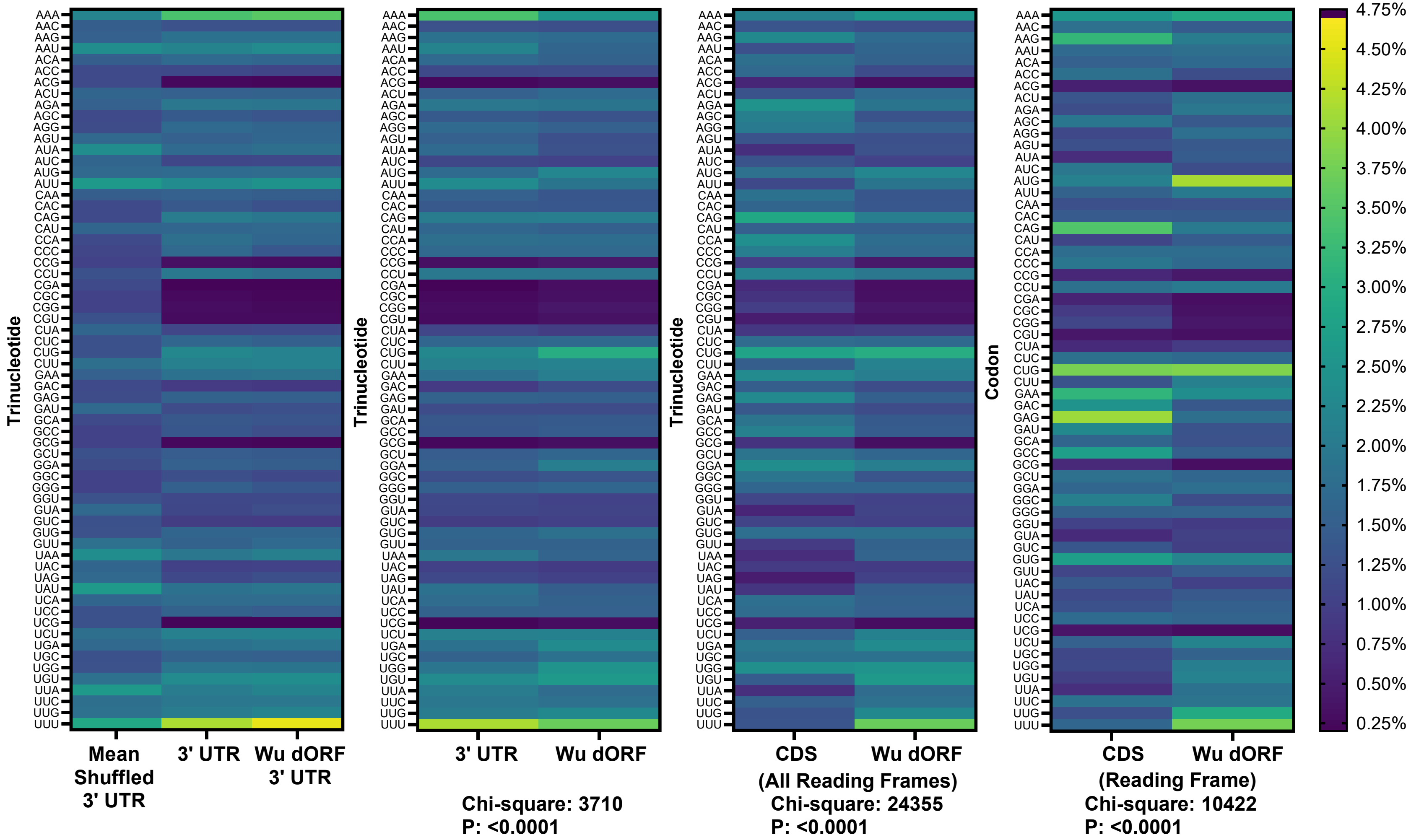

Figure 4.6: Trinucleotide or codon composition of Wu downstream open reading frames (dORFs) appears to be more similar to 3' untranslated regions (UTRs) than coding sequences (CDSs). *Heat maps represent the proportion (%) of sequences made up of each trinucleotide/codon and are grouped for comparisons. The heat maps compare RefSeq mRNA ID 3' UTR sequences against: the mean of three repeats of shuffled 3' UTR sequences, the 3' UTRs that contain Wu dORFs, and Wu dORF sequences. The two heat maps on the right compare RefSeq mRNA ID CDS sequences against Wu dORF sequences, across either all reading frames (trinucleotides in any frame) or the codons in the reading frame with stop codons (UAA, UAG, UGA) excluded. These comparisons were quantified with Chi-square goodness of fit tests comparing the difference and significance between expected and observed outcomes. In each heat map for the Chi-square tests the left column of observed values represents the expected values and the right column the observed.*

The nucleotide and dinucleotide compositions have the potential to affect the trinucleotide, or codon, compositions, and conversely trinucleotide, or codon, compositions could drive nucleotide or dinucleotide compositions. The suggested importance of Wu dORF translation to their regulatory function (Wu *et al.*, 2020b), means their trinucleotide composition or codon usage could be important. Although ribosomes associate with 3' UTRs, potential 3' UTR translation is relatively rare and 3' UTRs are generally considered to be untranslated (Ingolia *et al.*, 2009; Guydosh and Green, 2014; Ji *et al.*, 2015; Miettinen and Björklund, 2015; Young *et al.*, 2015; Hsu *et al.*, 2016). The sequences included in the analysis were the same from Figure 4.4, with the addition of codon composition analysis in the reading frame of CDSs and Wu dORFs. The trinucleotide frequency of RefSeq mRNA ID 3' UTRs differed from the frequency expected by chance, when compared with shuffled 3' UTRs (Figure 4.6). The trinucleotide compositions of 3' UTRs overall and Wu dORF-containing 3' UTRs were very similar (Figure 4.6). Similar to the UU dinucleotide, the UUU trinucleotide was also slightly more frequent in the 3' UTRs that contain Wu dORFs compared to those that did not (Figure 4.6). The Wu dORFs trinucleotide frequency differed slightly from the 3' UTRs, which could be driven by the potential translation of these dORFs (Figure 4.6). Wu dORF and CDS trinucleotide compositions across all reading frames were more different than when comparing Wu dORFs and 3' UTRs (Figure 4.6). The difference between Wu dORFs and CDSs remained when comparing codon composition (Figure 4.6). The statistically significant (P<0.0001) chi-square values supported the differences shown by the heat maps between the trinucleotide, or codon, compositions of the 3' UTR and Wu dORF-containing 3' UTRs, the 3' UTR and Wu dORFs, and the CDS and Wu dORFs. Similar to the dinucleotide composition analysis, the chi square values were increased when comparing Wu dORFs and CDSs suggesting that the difference was greater than when comparing Wu dORFs and 3' UTRs. These results are similar to the nucleotide and dinucleotide composition analysis. 3' UTR trinucleotide composition does not occur by chance and the Wu dORF-containing 3' UTRs did not differ from other 3' UTRs. Although potentially translated, Wu dORFs composition was more similar to 3' UTRs than CDSs, and Wu dORF codon composition was not similar to CDSs. The influence of CG dinucleotide frequency was seen in this trinucleotide and codon composition analysis where the least frequent trinucleotides, or codons, in 3' UTRs and Wu dORFs were those containing the CG dinucleotide (Figure 4.6).

The pattern of CG dinucleotide frequency across different sequence groups (Figure 4.4), was present to some extent for CG dinucleotide-containing trinucleotides ('CG trinucleotides' or 'CG codons'), and codons (Figure 4.7). The different CG trinucleotide frequencies varied from one another across the different sequences. The mean frequency of these trinucleotides or codons was used to investigate the general effect, with standard deviation bars indicating the variation between trinucleotides, or codons (Figure 4.7). The CG trinucleotides were around four times less abundant in 3' UTRs compared to shuffled 3' UTR sequences, with means of 1.17% and 0.30% respectively, and a statistically significant difference (P<0.0001). The CG trinucleotide frequency was similar in RefSeq 3' UTRs and those containing Wu dORFs, with slightly increased abundance in Wu dORFs (Figure 4.7). The difference between mean frequencies in Wu dORFs (0.39%) and 3' UTRs (0.30%), and Wu dORFs and Wu dORF-containing 3' UTRs (0.28%) was statistically significant (P<0.0001). This could mean that in Wu dORF containing 3' UTRs, the 3' UTR sequence outside of the dORF could have reduced frequency of trinucleotides that contain CG dinucleotides when compared to other 3' UTRs without Wu dORFs. This could be explored further to establish if it could be a feature. A larger difference was found between Wu dORFs and RefSeq CDSs, with increased CG trinucleotide abundance in CDSs (Figure 4.7). The mean CG trinucleotide CDS frequency was 0.78%, at least twice as frequent in CDSs, although variation between CG trinucleotides was greater (Figure 4.7). When considering CG codons in Wu dORFs and CDSs, the mean frequency in CDSs (0.70%) was statistically significantly (P = 0.0242) greater than that of Wu dORFs (0.39%) (Figure 4.7). The trinucleotide, or codon, frequency may be influenced by other dinucleotide frequencies, such as CGG, which could be affected by GG frequency (Figure 4.4), which was greater in CDSs compared to Wu dORFs, potentially increasing CGG frequency in CDSs (Figure 4.7). The Wu dORFs continue to be more similar to 3' UTRs than CDSs, and the CG dinucleotide frequency also had an impact on CG dinucleotide-containing trinucleotide frequency.

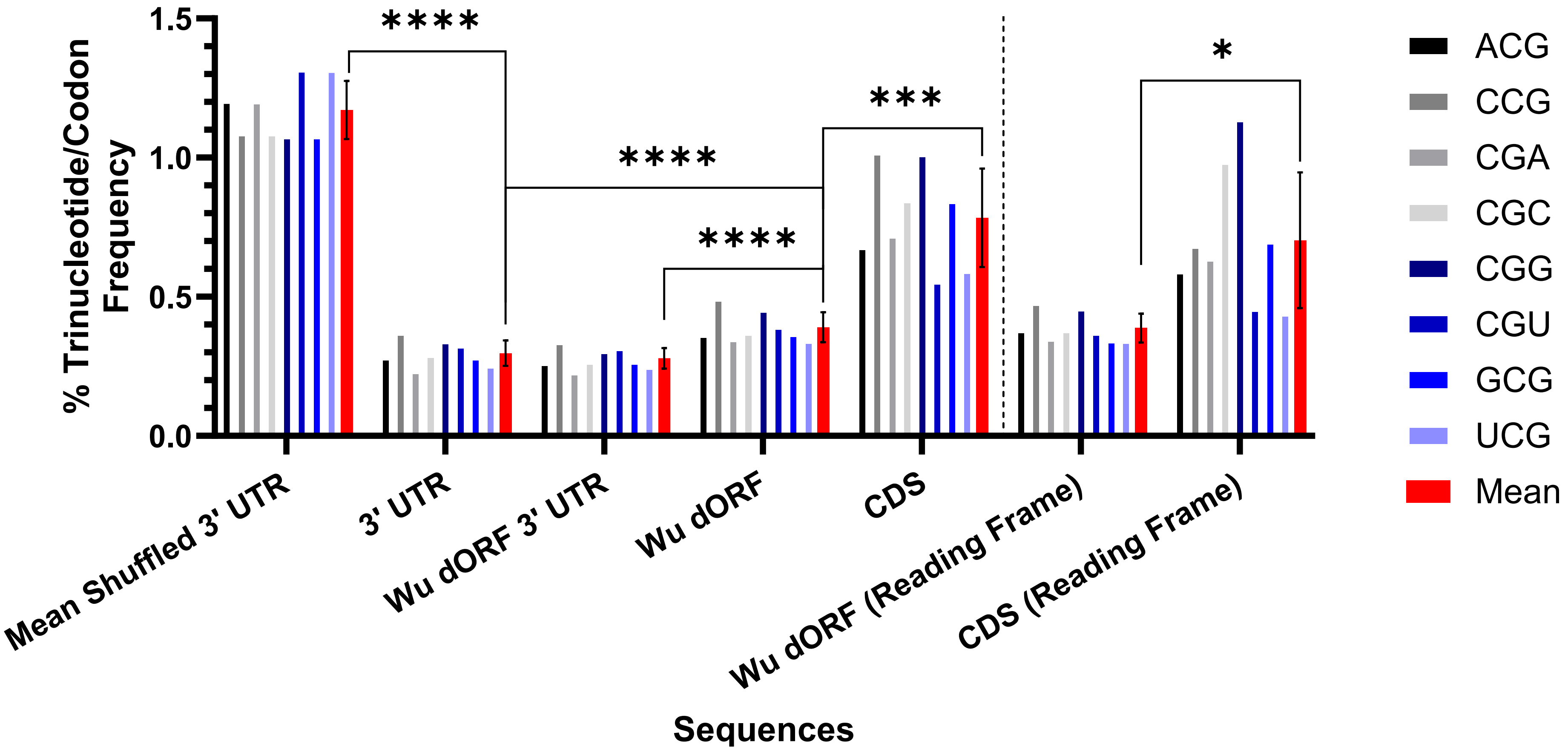

Figure 4.7: The frequency of trinucleotides, or codons, containing Cytosine Guanine (CG) dinucleotides varies across Wu downstream open reading frames (dORFs), coding sequences (CDSs) and 3' untranslated regions (3' UTRs). *The CG containing trinucleotide, or codon, frequency (represented as a percentage of all the trinucleotides, or codons) in different sequences was compared, including: shuffled RefSeq mRNA ID 3' UTRs, 3' UTRs, 3' UTRs containing Wu dORFs, Wu dORFs (in all frames and in the reading frame), and RefSeq mRNA ID CDS (in all reading frames and the reading frame). For each group of sequences, the mean across the CG containing trinucleotides/codons with standard deviation bars are included in red. The legend includes the which coloured bar represents which trinucleotide/codon. One-Way ANOVA multiple comparisons test with Šídák adjusted P values was used to compare the means. * - P<0.05, *** - P<0.0005, **** - P<0.0001.*

The trinucleotide composition analysis also explored the frequency of AUG and non-AUG start codons, or trinucleotides, (CUG, GUG, UUG) in shuffled 3’ UTR sequences, RefSeq 3’ UTRs, and Wu dORF-containing 3’ UTRs. This investigated whether start trinucleotides appeared more frequently in 3’ UTRs than was expected by chance, and whether Wu dORF-containing 3’ UTRs had more potential start codons. Start codons could provide dORF start sites. UUG was most common in shuffled 3’ UTR sequences, followed by AUG (Figure 4.8). Start trinucleotide frequency was very similar in RefSeq 3’ UTRs and Wu dORF-containing 3’ UTRs, CUG was the most common, followed by UUG, then AUG and GUG (Figure 4.8). AUG frequency was similar across shuffled 3’ UTRs, RefSeq 3’ UTRs, and Wu dORF-containing 3’ UTRs (Figure 4.8). The other start trinucleotides increased in frequency from shuffled to RefSeq 3’ UTRs; CUG increased the most from 1.32% to 3.20%. Mean start trinucleotide frequencies were reported and although the differences were not statistically significant (P>0.05), RefSeq 3’ UTRs start trinucleotide frequency (1.95%) increased compared to shuffled 3’ UTRs (1.56%) (Figure 4.8). RefSeq 3’ UTRs (1.95%) and Wu dORF-containing 3’ UTRs (1.97%) did not differ in mean start trinucleotide frequency (Figure 4.8). Potentially 3’ UTR start trinucleotides appeared slightly more frequently than expected, but Wu dORF-containing 3’ UTRs showed no preference for start codons. The dinucleotide and trinucleotide frequencies could influence each other, most dinucleotides in start codons had varied abundance between the different 3’ UTR sequences, such as UG, which had increased abundance in RefSeq 3’ UTRs compared to shuffled sequences (Figure 4.4).

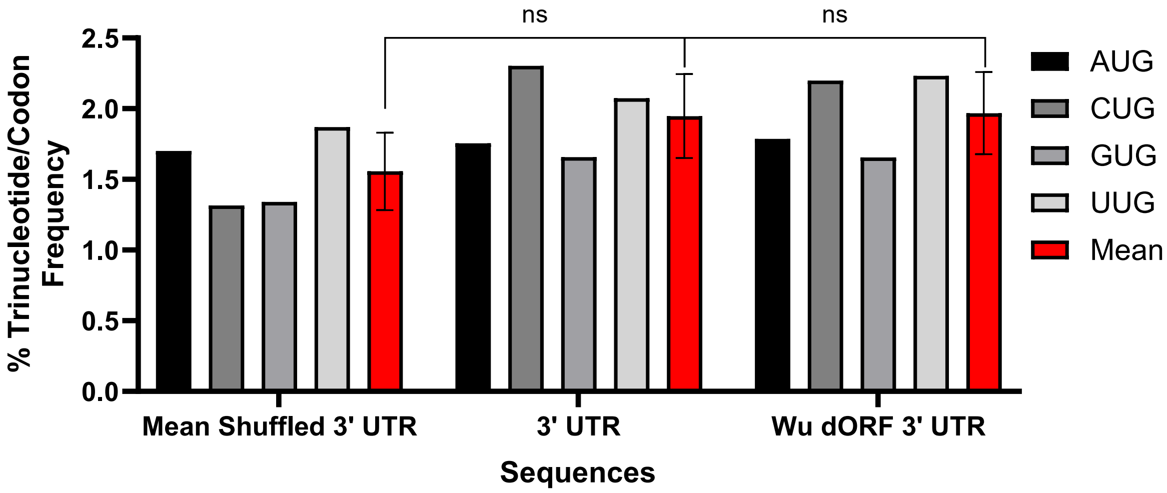

Figure 4.8: Start trinucleotides, including non-canonical start codons, occur at varied frequencies in 3' untranslated regions (UTRs), shuffled sequences, and Wu downstream open reading frame (dORF) containing 3' UTRs. *Comparing the frequency of start trinucleotides, represented as a percentage of all the trinucleotides, in different sequences, including: shuffled RefSeq mRNA ID 3' UTRs, 3' UTRs, and 3' UTRs containing Wu dORFs. For each group of sequences, the mean start trinucleotide frequency with standard deviation bars are included in red. The legend includes the which coloured bar represents which start codon. One-Way ANOVA multiple comparisons test with Šídák adjusted P values was used to compare the means. ns – P>0.05.*

The trinucleotide composition analysis data was used to investigate stop codon, or trinucleotide, frequency in RefSeq 3' UTRs, shuffled 3' UTRs, and those containing Wu dORFs. Using shuffled sequences helped explore whether stop trinucleotides appeared more or less frequently than expected in 3' UTRs, and whether Wu dORF presence altered stop codon frequency. Stop codons are required for dORFs, however reduced frequency of 3' UTR stop codons could extend dORF sequences, potentially promoting dORFs. UAA was the most common stop trinucleotide in shuffled 3' UTRs, followed by UAG and UGA (Figure 4.9). Stop trinucleotide frequency in RefSeq 3' UTRs and those containing Wu dORFs was similar, UAA was most common, followed closely by UGA, then UAG (Figure 4.9). The mean stop trinucleotide frequency difference between Wu dORF-containing 3' UTRs (1.80%) and RefSeq 3' UTRs (1.69%) was not statistically significant (P>0.05)(Figure 4.9). UAA and UAG stop trinucleotides occurred less frequently in RefSeq 3' UTRs compared to shuffled sequences, whereas UGA was slightly more frequent (Figure 4.9). UA dinucleotide frequencies shared the same pattern, less abundant in RefSeq 3' UTRs compared to shuffled sequences (Figure 4.4), and could influence stop trinucleotide occurrence. Although not statistically significant (P>0.05), mean stop trinucleotide frequency reduced slightly in RefSeq 3' UTRs (1.69%) compared to shuffled sequences (1.94%) (Figure 4.9). 3'UTR stop trinucleotides, particularly UAA and UAG, may appear slightly less frequently than expected in 3' UTRs. Wu dORF-containing 3' UTRs did not have fewer stop codons, they were slightly more abundant, suggesting there was no preference to extent dORF sequences.

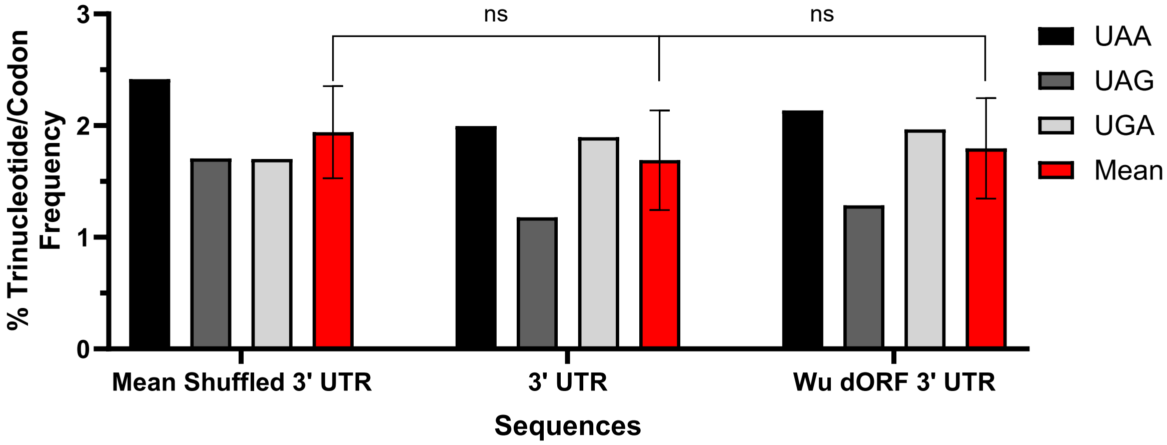

Figure 4.9: Stop trinucleotides have similar abundance in 3' untranslated regions (UTRs) and those containing Wu downstream open reading frame (dORF) shuffled sequences, with little change in containing 3' UTRs. *Comparing the frequency of stop trinucleotides, represented as a percentage of all the trinucleotides, in shuffled RefSeq mRNA ID 3' UTRs, 3' UTRs, and 3' UTRs containing Wu dORFs. For each group of sequences, the mean stop trinucleotide frequency with standard deviation bars are included in red. The legend includes the which coloured bar represents which stop codon. One-Way ANOVA multiple comparisons test with Šídák adjusted P values was used to compare the means. ns – P>0.05*

4.5 CDS stop codons used in transcripts containing Wu dORFs

One explanation for 3' UTR ribosome presence is stop codon readthrough, where the stop codon is ignored and ribosomes continue translation into the 3' UTR (Doronina and Brown, 2006; Namy and Rousset, 2010), even if it is a relatively uncommon occurrence. Stop codon use can influence readthrough, readthrough is most common with UGA (Howard *et al.*, 2000; Manuvakhova, Keeling and Bedwell, 2000; Bidou *et al.*, 2004; Floquet *et al.*, 2012; Wangen and Green, 2020). To investigate whether Wu dORF-containing transcripts were more likely to have 3' UTR RP reads found due to readthrough, the CDS stop codon preference genome-wide and in Wu dORF-containing transcripts were compared.

Table 4.5: Coding sequence (CDS) stop codon selection preceding Wu downstream open reading frames (dORFs) is similar to that seen across the human genome. *The table include the percentage of each stop codon ending the CDSs across the human genome and also in CDSs which precede Wu dORFs.*

	Percentage of each stop codon used at CDS end %	
CDS Stop Codon	**Genome Wide**	**Wu dORFs**
UAA	28.4%	31.4%
UAG	22.3%	20.6%
UGA	49.2%	48.0%

When comparing genome-wide CDS stop codon usage, UGA was most common (49.2%), UAG was least commonly used (22.3%), and UAA was used slightly more frequently (28.4%) (Table 4.5). In CDSs preceding Wu dORFs, UGA was again the most common (48.0%), followed by UAA (31.4%), and UAG remains the least common (20.6%) (Table 4.5). The stop codon usage genome-wide and in CDSs preceding Wu dORFs was similar, the small changes by a few percent are unlikely to affect the frequency of stop codon readthrough. Wu dORFs were not more common in transcripts where stop codon readthrough was more likely, due to the CDS stop codon used, meaning this is not part of the mechanism behind ribosomal association with Wu dORF-containing 3' UTRs.

4.6 Discussion

dORF sequences are common, like other sORFs (Slavoff *et al.*, 2013; Bazzini *et al.*, 2014; Couso and Patraquim, 2017; Lu *et al.*, 2019; van Heesch *et al.*, 2019; Chen *et al.*, 2020; Ouspenskaia *et al.*, 2020; Ruiz Cuevas *et al.*, 2021), covering a large proportion of 3' UTRs, and appearing more frequently than expected from shuffled sequences (Table 4.1). This suggests evolutionary selection for dORFs. The large number of dORF sequences found in shuffled sequences shows that these sequences could occur frequently through random generation (Guttman *et al.*, 2013; Couso and Patraquim, 2017). dORF function is suggested to depend on their translation (Wu *et* *al.*, 2020b), meaning dORFs could provide another explanation for 3’ UTR ribosomal association, suggesting dORFs would occur more frequently in 3’ UTRs with ribosomal association. However, when considering 3’ UTRs with RP alignment ‘Potential dORFs’ were not more abundant (Table 4.1). Around two thirds of 3’ UTR aligned RP reads were within ‘Potential dORF’ sequences, after removing RP reads that repeatedly aligned to the same region (Table 4.2), less than the proportion of 3’ UTRs covered by ‘Potential dORFs’. The 15% of 3’ UTRs not covered by dORFs contained around 33% of 3’ UTR aligned RP reads, indicating less ribosomal association, and potential translation, than expected with ‘Potential dORFs’, based on their coverage. Although ‘Potential dORFs’ could explain many 3’ UTR aligned ribosomes, only a small proportion of ‘Potential dORFs’ had ribosomal association and may be translated, known as ‘RP Potential dORFs’. This alongside the potential random generation of dORFs, could indicate that only some dORFs have translational regulator activity. The small proportion of 3’ UTRs without dORF coverage contained around one third of 3’ UTR ribosomal association, and the reason for this is unclear, one explanation could be stop codon readthrough (Doronina and Brown, 2006; Namy and Rousset, 2010), or ribosomes passing through 3’ UTRs without translating (Guydosh and Green, 2014; Miettinen and Björklund, 2015).

Transcripts containing ‘RP Potential dORFs’ underwent translational regulation analysis with comparison against transcripts without ‘RP Potential dORFs’. There was no evidence of translational regulation in ‘RP Potential dORF’ transcripts, in healthy or tumour samples (Figure 4.2). In tumour datasets there were more transcripts with ‘RP Potential dORFs’, and more ‘RP Potential dORFs’, potentially meaning more dORFs were being translated, however there was no change in regulatory activity (Figure 4.2). The increased ribosomal association with dORFs in tumour tissue, could relate to cellular conditions, which can influence 3’ UTR ribosomal presence (Hinnebusch, 1988, 2005; Ingolia *et al.*, 2009). Although many transcripts had ‘RP Potential dORFs’, RP read alignment to dORFs was not consistent across datasets. Some individual ‘RP Potential dORF’ transcripts, showed possible translational regulation, leading to investigation of whether regulatory activity correlated with ‘RP Potential dORF’ RP read density. Greater ‘RP Potential dORF’ RP read density, would be expected to correlate with greater relative translation, as dORF translation is suggested to increase CDS translation (Wu *et al.*, 2020b). The lack of overall regulatory activity could have been due to 'RP Potential dORFs' with low RP read density masking effects of those with greater RP read density. However, when considering RP read, or read location, density, there was likely no biologically meaningful correlation (Figure 4.3 and Table 4.3). Multiple dORF-containing transcripts had correlation data plotted for each dORF, meaning transcript differential expression values can be included several times. Most 'RP Potential dORFs' had low RP read density which increased in tumour datasets, but even with greater RP read density, no biologically meaningful correlation was present. Many dORF had little ribosomal association or regulatory activity, meaning some 'RP potential dORFs' may be transcribed sORFs without biological function (Guttman *et al.*, 2013). Rather than evidencing dORF translation, ribosomal association with these dORFs could occur by chance due to the extent of 3' UTR coverage by dORF sequences. These 3' UTR ribosomes may be from stop codon readthrough (Doronina and Brown, 2006; Namy and Rousset, 2010) or improper termination and recycling of ribosomes (Guydosh and Green, 2014; Miettinen and Björklund, 2015). It remains unclear why some dORFs, such as Wu dORFs, presented by Wu *et al.* (2020b), appear to be functional and others do not.

The nucleotide composition analysis compared Wu dORF sequences to 3' UTRs and CDSs, and the associated conclusions refer to consideration of groups of transcripts, meaning individual significance could be masked. Wu dORF translation could imply similarity between Wu dORF and CDS composition, as coding regions. The nucleotide, dinucleotide, and trinucleotide, or codon, composition results do not support this. Instead, Wu dORF sequences share more similarities with 3' UTRs than CDSs. The comparisons are supported by chi square values, which may gain significance due to large sizes of observations and vary with different sample sizes. The 3' UTR similarity and lack of CDS-like codon preferences supports suggestion that dORF function does not involve the translated product (Barbosa, Peixeiro and Romão, 2013; Couso and Patraquim, 2017). However, it could also suggest Wu dORFs are stochastically generated 3' UTR sORFs without function. However, Wu dORFs potentially have increased the GC base composition, and potential secondary structure, compared to the remaining 3' UTR (Shabalina, Ogurtsov and Spiridonov, 2006). Wu dORF-containing 3' UTR composition was hypothesised to differ from other 3’ UTRs, due to dORF translation, assisting in identification of other functional dORFs. However, 3’ UTRs with, and without, Wu dORFs, share very similar nucleotide, dinucleotide, and trinucleotide composition with other 3’ UTRs. The increased UU dinucleotide, and UUU trinucleotide, frequency in Wu dORFs, or the region upstream of dORFs could be explored as a potential opportunity to find other dORFs. 3’ UTR composition is not driven by chance, evidenced by comparisons with shuffled, essentially random, 3’ UTRs. The importance of 3’ UTR composition may relate to targets for post-transcriptional regulators, such as micro-RNAs (miRNA) (Ha and Kim, 2014; Bartel, 2018) or AU rich elements (AREs) (Eberhardt *et al.*, 2007; López De Silanes, Paz Quesada and Esteller, 2007; Harvey *et al.*, 2018). The composition of Wu dORFs and 3’ UTRs containing them is generally not useful when searching for other dORFs which share their activity (Wu *et al.*, 2020b).

The complexity of nucleotide, dinucleotide, trinucleotide, or codon, composition influences on each other makes investigating dinucleotides or trinucleotides difficult. C and G nucleotides are less common in 3’ UTRs, however, 3’ UTRs CG dinucleotide frequency was lower than expected, using shuffled sequences. This dinucleotide was least common in 3’ UTRs and Wu dORFs, with increased abundance in the more GC rich CDS, although still at lower frequency (Figure 4.5). This pattern continued with CG containing trinucleotides, although other dinucleotide or trinucleotide influences may be involved (Figure 4.7). 3’ UTRs have reduced CG dinucleotide frequency, especially compared to 5’ UTRs (Pesole *et al.*, 1997, 2001), possibly relating to DNA methylation of CpG sites (Portela and Esteller, 2010; Jang *et al.*, 2017). Start and stop codon frequency could be influenced by UG and UA dinucleotide frequencies. The start and stop codon abundance did not change in Wu dORF-containing 3’ UTRs, suggesting these sequences were not more likely to contain sORFs (Figures 4.8 and 4.9). Although not statistically significant, 3’ UTRs contained slightly more start codons and slightly less stop codons than expected from shuffled 3’ UTRs. This could indicate a slight preference towards dORF generation with increased sites to start 3’UTR translation and reduced termination sites, however this is unlikely due to very small overall changes.

This chapter raises questions about whether dORF and 3’ UTR ribosomal association is evidence of translation, and whether ribosomal association with some dORFs could occur by chance or other mechanisms. Although, the hypothesis for dORF translation suggests 3’ UTR reinitiation (Wu *et al.*, 2020b), stop codon readthrough could also explain dORF ribosomal association, or potentially translation if the dORF was in the CDS reading frame. CDS stop codons influence the likelihood of readthrough, UGA is most commonly involved in readthrough (Howard *et al.*, 2000; Manuvakhova, Keeling and Bedwell, 2000; Bidou *et al.*, 2004; Floquet *et al.*, 2012; Wangen and Green, 2020). CDS stop codon frequency in Wu dORF-containing transcripts was investigated to determine whether these transcripts were more likely to undergo stop codon readthrough, to increase 3’ UTR ribosomal presence. Wu dORF-containing transcripts and genome-wide CDS stop codon frequency was similar. UGA frequency slightly reduced and UAA frequency slightly increased, suggesting Wu dORF-containing transcripts did not have CDS stop codon frequencies that would make readthrough more common. Stop codons are a small part of the readthrough process and other factors can influence the process, such as 3’ UTR length or P site codon, where AUA, ACA, ACC, CUG, and GAC codons can make readthrough more efficient (Mangkalaphiban *et al.*, 2021). Future analysis could explore these other factors in transcripts with Wu dORFs, or other dORFs of interest, to explore readthrough further and its role in 3’ UTR associated ribosomes.

4.7 Conclusions

Wu dORFs shared more similarities with noncoding 3’ UTRs than CDSs when comparing nucleotide, dinucleotide and trinucleotide composition. The composition of Wu dORFs and their 3’ UTRs do not appear to be useful in identifying other similar functional dORF sequences. Potential dORF sequences appear slightly more frequently than expected by chance in 3’ UTRs. Potential dORFs cover a large proportion of the 3’ UTR; however, ribosomal association with these dORFs is lower than expected based on the coverage. Although potential dORF sequences are widespread and can associate with ribosomes, most dORFs have low ribosomal association and little evidence of regulatory activity. Again, there was increased dORF ribosomal association in tumour tissue compared to healthy tissue. Altered CDS stop codon frequency compared to CDSs across the genome, favouring stop codon readthrough to increase ribosomal presence in 3’ UTRs, was not found in Wu dORF-containing transcripts.

Chapter 5: dORFs are Conserved Across Species

5.1 Introduction

Downstream open reading frames (dORFs) are abundant in 3’ untranslated regions (UTRs), seen in the previous results chapter. Some sORFs, and encoded peptides, are conserved in other species, suggesting evolutionary pressure to maintain these sequences, implying a function (Martinez *et al.*, 2019). Results have varied so far, some support dORF regulatory function and ribosomal association, but others suggest little regulatory activity and translation. Through exploration of dORF conservation across species, with comparison to 3’ UTRs, this could show whether there is evolutionary pressure to conserve dORFs. Highly conserved 3’ UTR regions may have regulatory function (Bashirullah, Cooperstock and Lipshitz, 1998; Conne, Stutz and Vassalli, 2000; Grzybowska, Wilczynska and Siedlecki, 2001; Pesole *et al.*, 2001; Mignone *et al.*, 2002; Shabalina *et al.*, 2004). Wu dORFs, dORF presented by Wu *et al.* (2020b), in addition to having regulatory function, are suggested to be more conserved than the surrounding 3’ UTR in zebrafish and mice (Wu *et al.*, 2020b). Human dORF peptides have also been shown to be conserved in mice (Ji *et al.*, 2015). In 5’ UTRs, uORF presence is frequently conserved between species, often without sequence similarity (Chew, Pauli and Schier, 2016; Johnstone, Bazzini and Giraldez, 2016; Dumesic *et al.*, 2019). In vertebrates the most conserved UTR is the 3’ UTR, although it is not as conserved at the CDS (Siepel *et al.*, 2005; Litterman *et al.*, 2019), making it important to compare the dORFs and 3’ UTR. A fully conserved 3’ UTR conserves a dORF too, however it would be difficult to determine whether the regulatory potential of the dORF, or something else, had driven the 3’ UTR conservation. The previous results chapters have shown that dORF activity and ribosomal association with dORFs and 3’ UTRs can vary between cancer and healthy tissue. Under different cellular conditions, particularly under stressful conditions, 3’ UTR ribosomal presence can change (Hinnebusch, 1988, 2005; Ingolia *et al.*, 2009). The regulatory activity of dORFs could change in different disease states or cellular conditions. These changes could also relate to changes in the relative ribosomal association, and possible translation of dORFs and 3’ UTRs. This could suggest dORFs are more like uORFs. uORF translation can be affected by local conditions and disease states, such as cancer or inflammation (Young and Wek, 2016; Sendoel *et al.*, 2017; Renz, Valdivia Francia and Sendoel, 2020). If dORF activity or translation varies across conditions and disease states, this is important for understanding how dORFs function, and it could be exploited in disease states for potential future treatment options. The hypotheses for this chapter are that dORFs, potentially due to some regulatory function, will be more conserved than 3' UTRs across species, and that with differing cell types, treatments or conditions the relative ribosomal association with dORFs and 3' UTRs will change. This is expected to be most apparent in stressful cellular conditions.

Bioinformatic analyses were used to explore dORF and 3' UTR conservation in different species and the influence of different cell types and conditions on ribosomal association with dORFs and 3' UTRs. Section 2.9 described the methods used to run conservation analysis of human dORFs, starting with an AUG start codon (referred to as AUG dORFs), and the 3' UTRs containing these dORFs, in other species ranging from primates to plants. To validate the findings, control sequences were also developed and analysed to provide a comparison for the dORF and 3' UTR conservation results. These analyses investigated whether dORFs were more conserved than the 3' UTR that contained them. A shortlist of highly conserved dORFs across species, known as HC dORFs, from this analysis were used in the subsequent analysis. Analysis then turned to ribosome profiling (RP) read alignments to dORF shortlists, MSVW dORFs (Wu dORFs with MS validation of translation) and HC dORFs, and the 3' UTRs containing these, relative to the RP reads aligned to the gene, using methods described in section 2.10. This analysis used RP datasets from various cell types, treatments, conditions and disease states. This analysis investigated whether ribosomal association with dORFs, possibly indicating translation, varies in different conditions and disease states to further dORF understanding and potentially highlight dORF importance in different diseases.

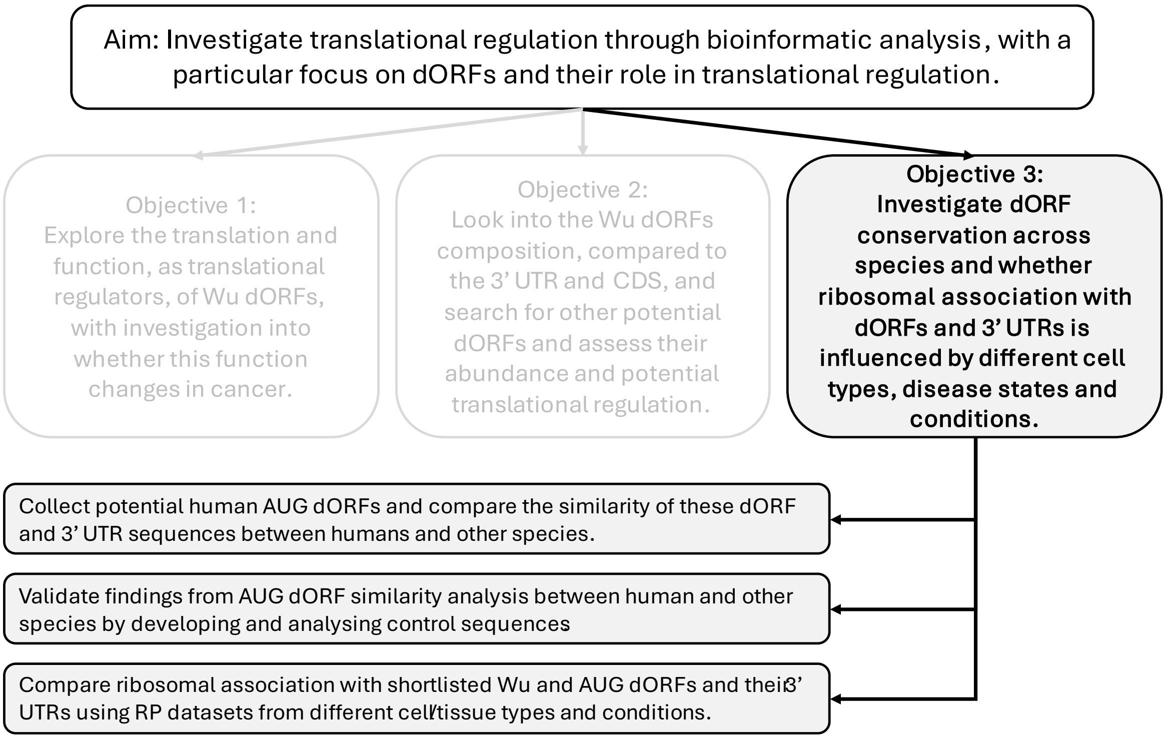
Aim: Investigate translational regulation through bioinformatic analysis, with a particular focus on dORFs and their role in translational regulation.
Objective 1: Explore the translation and function, as translational regulators, of Wu dORFs, with investigation into whether this function changes in cancer.
Objective 2: Look into the Wu dORFs composition, compared to the 3' UTR and CDS, and search for other potential dORFs and assess their abundance and potential translational regulation.
Objective 3: Investigate dORF conservation across species and whether ribosomal association with dORFs and 3' UTRs is influenced by different cell types, disease states and conditions.
Collect potential human AUG dORFs and compare the similarity of these dORF and 3' UTR sequences between humans and other species.
Validate findings from AUG dORF similarity analysis between human and other species by developing and analysing control sequences
Compare ribosomal association with shortlisted Wu and AUG dORFs and their 3' UTRs using RP datasets from different cell/tissue types and conditions.

Figure 5.1: Summary of approaches used in section 5 to meet the chapter objectives and overall study aim. *dORF – downstream open reading frame, 3' UTR – 3' untranslated region, CDS – Coding Sequence, RP – Ribosome Profiling, AUG dORF – downstream open reading frame using an AUG start codon.*

5.2 Conservation of human dORFs and 3' UTRs in other species

Conservation of sORFs, and the encoded peptides, suggests importance and function (Martinez *et al.*, 2019). In 3' UTRs, highly conserved regions are suggested to have regulatory function (Bashirullah, Cooperstock and Lipshitz, 1998; Conne, Stutz and Vassalli, 2000; Grzybowska, Wilczynska and Siedlecki, 2001; Pesole *et al.*, 2001; Mignone *et al.*, 2002; Shabalina *et al.*, 2004), and Wu dORFs appear to be more conserved than surrounding regions in a few species (Wu *et al.*, 2020b). These results sought to expand this knowledge by investigating conservation of dORFs with AUG start codons (AUG dORFs), and their 3' UTRs, across a wide range of species. Greater conservation of dORFs compared to their 3' UTRs could suggest these dORFs have function. Human 3' UTRs with RefSeq mRNA IDs that contained AUG dORFs were used. In other species the homologous genes of human AUG dORF-containing genes were used to find RefSeq 3' UTRs, and those containing AUG dORFs, in other species. This allowed AUG dORF conservation between humans and other species to be investigated by comparing AUG dORF sequence similarity in homologous genes. The number of 3' UTR sequences gathered for each species varies depending on the number of genes, available transcripts, and RefSeq annotation. Generally, the number of 3' UTR sequences gathered reduced in species that are more distantly related to humans (Table 5.1), in all vertebrates, up to *D. rerio*, there are at least 20000 3' UTRs in homologous genes to human 3' UTRs with AUG dORFs. It was not possible to collect 3' UTR sequences in three species using these methods, caused by only partial mRNA sequences being available in the NCBI nucleotide database (Sayers *et al.*, 2022). Fewer 3' UTR sequences meant that considerably fewer AUG dORF comparisons can take place, meaning the conservation analysis was more limited in more distantly related species. The proportion of collected 3' UTRs in each species which do not contain AUG dORFs was fairly consistent across vertebrates, ranging from 7.8% to 13.0%, suggesting most 3' UTRs contain AUG dORFs (Table 5.1). This proportion increased in invertebrates, where over one third of 3' UTRs can be without an AUG dORF (Table 5.1). Then in some yeast and plant species this proportion reduced again (Table 5.1). Most species had homolog 3' UTRs, and most 3' UTRs contained AUG dORFs. The frequency of AUG dORFs varied between species, ranging from 0.75 in *C. elegans* to 1.32 dORFs per 100 nucleotides in *N. crassa*. Generally, in 3' UTRs containing AUG dORFs, there was around one dORF per 100 nucleotides of 3' UTR (Table 5.1).

The mean lengths of 3' UTRs which did, and did not, contain AUG dORFs were compared. Differing 3' UTR lengths could influence the likelihood of AUG dORFs being found, shorter 3' UTRs could be less likely to contain dORFs. The mean length of AUG dORF-containing 3' UTRs varied across species, and generally the mean 3' UTR length reduced in species more distantly related to humans, particularly yeast and plant species (Figure 5.2). The mean length of 3' UTRs without AUG dORFs was consistent across species, usually between 100 and 200 nucleotides (Figure 5.2). Across all species, the mean length of AUG dORF-containing 3' UTRs was at least twice as long as 3' UTRs without AUG dORFs (Figure 5.2). In vertebrates, the mean AUG dORF-containing 3' UTR length can be over ten times longer than the mean of 3' UTRs without (Figure 5.2). In all species between humans and *C. elegans* the differences between the mean length of 3' UTRs with, and without, AUG dORFs were statistically significant (P<0.05), and in all vertebrates the P value was less than 0.0001 (Figure 5.2). The shorter 3' UTRs in the more distantly related species suggests fewer AUG dORFs could be found in these 3' UTRs. There were also fewer 3' UTRs collected for these species which should be considered in the subsequent analysis. Generally, 3' UTRs containing an AUG dORF are longer than 3' UTRs without, potentially due to shorter sequences of around 100 to 200 nucleotides being much less likely to contain a dORF which could be 30 to 300 nucleotides long. This is supported by previous findings where around one AUG dORF was found per 100 nucleotides of 3' UTR across species (Table 5.1).

Table 5.1: Number of 3' untranslated region (UTR) sequences with RefSeq annotation with, and without, downstream open reading frames (dORFs) starting with an AUG start codon across various species. *This table includes data for the species used in the AUG dORF conservation analysis, reporting the number of 3' UTR sequences, once duplicate 3' UTR sequences from different transcripts were removed, which do, and do not, contain AUG dORFs and the percentage of collected 3' UTRs that contain AUG dORFs. The frequency of AUG dORFs in the 3' UTRs that contain them was also reported per 100 nucleotides.*

Species	**Homolog RefSeq 3' UTRs of human genes with AUG dORF**	**Number of AUG dORFs per 100nt**	**3' UTRs without AUG dORFs**	**% 3' UTRs without AUG dORFs**
Homo sapiens	174518	1.07	23514	11.9
Pan troglodytes	63388	1.07	6837	10.8
Macaca mulatta	40008	1.05	4177	10.4
Canis lupus	142416	1.03	18511	13.0
Bos taurus	46405	1.02	5878	12.7
Mus musculus	63667	1.08	4949	7.8
Rattus norvegicus	50677	1.05	5874	11.6
Gallus gallus	70789	1.13	5603	7.9
Xenopus tropicalis	25488	1.21	2243	8.8
Danio rerio	21555	1.23	2472	11.5
Anopheles gambiae	1864	0.96	723	38.8
Drosophila melanogaster	7637	1.00	1731	22.7
Caenorhabditis elegans	3453	0.75	1322	38.3
Saccharomyces cerevisiae	0	NA	NA	NA
Kluyvermyces lactis	0	NA	NA	NA
Eremothecium gossypii	0	NA	NA	NA
Schizosaccharomyces pombe	1149	0.99	290	25.2
Magnaporthe oryzae	1203	1.16	159	13.2
Neurospora crassa	1045	1.32	95	9.1
Arabidopsis thaliana	5276	1.03	832	15.8
Oryza sativa	1928	1.27	210	10.9

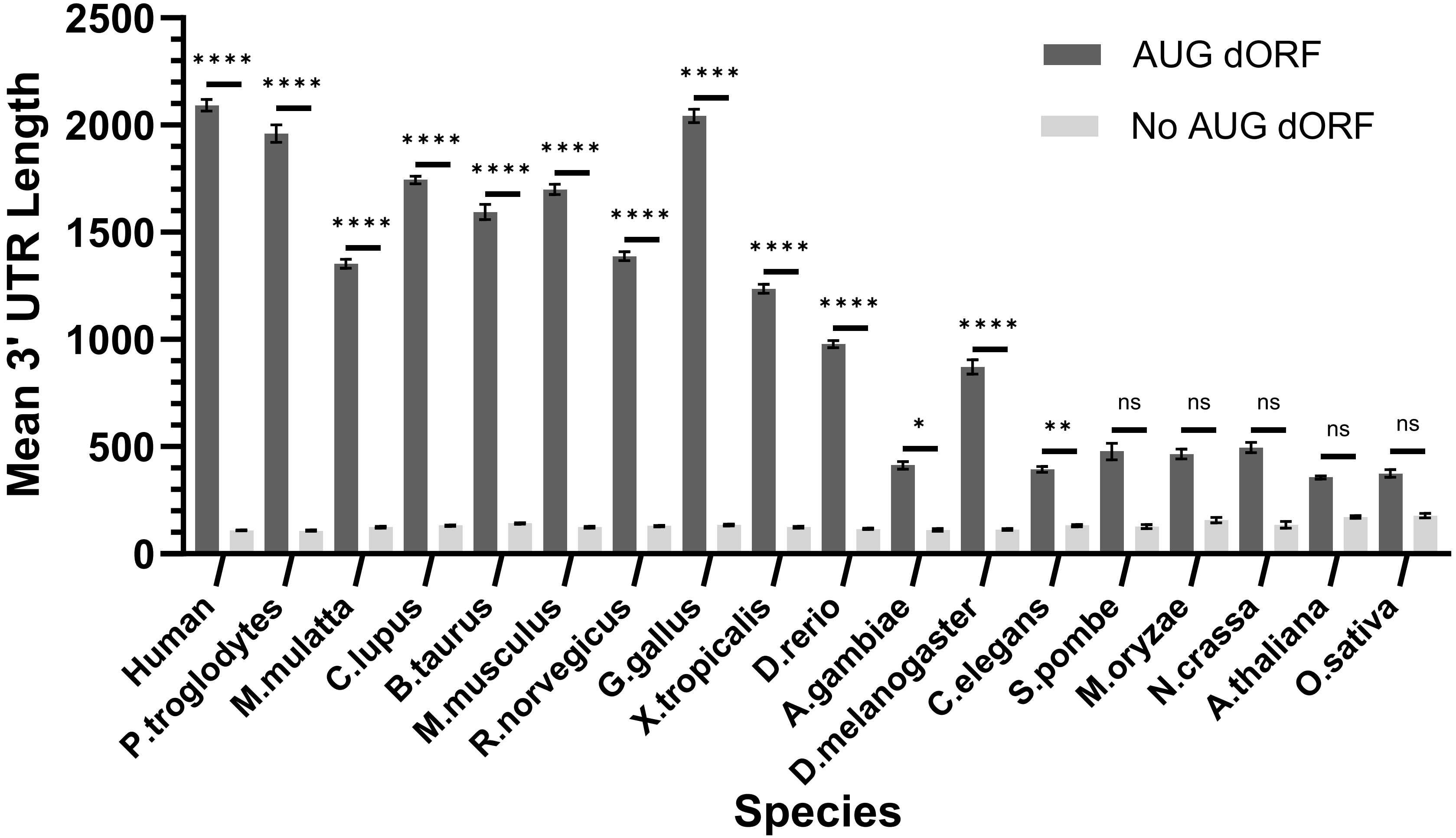

Figure 5.2: Across different species, the mean 3' untranslated region (UTR) length of sequences without downstream open reading frames (dORFs) which begin with an AUG start codon is often much shorter than 3' UTRs that contain AUG dORFs. *The length is reported in nucleotides. The 3' UTR sequences included have RefSeq annotation and duplicate sequences from different transcripts have been removed. Each species has two bars, the dark grey represents the mean 3' UTR length of sequences with AUG dORFs and the light grey represents the 3' UTRs without AUG dORFs. All bars include error bars which represent the standard deviation. One-Way ANOVA multiple comparisons test with Šídák adjusted P values was used to compare the means. * - P<0.05, ** - P<0.005, **** - P<0.0001.*

The conservation analysis is based on comparisons of AUG dORF similarity in human 3' UTRs and 3' UTRs of homologous genes, of those containing the human AUG dORF, in other species. However, in some species there is no homologous gene or no 3' UTR gathered, meaning each human AUG dORF can't be analysed against all species. The similarity analysis used the EMBOSS WATER tool (version – 6.6.0.0) which reported the similarity of the sequences if the alignment score exceeds zero, meaning some low similarity AUG dORF comparisons were not reported (Smith and Waterman, 1981; Peter Rice, Longden and Bleasby, 2000). Results were excluded if the alignment length was smaller than the shortest sequence compared. Figure 5.3 shows how many human AUG dORFs had comparisons, similarity results, in different numbers of species. 193437 human AUG dORFs had comparisons in one other species, and most, 494099, had comparisons in up to three species (Figure 5.3). In contrast, only one human AUG dORF had comparisons in 15 species, and 1270 in more than 10 species (Figure 5.3). Most human AUG dORFs had similarity comparison results in a few species and a minority had similarity results in a wider range of species, likely due to reduced homologous 3' UTR availability or low alignment scores in more distantly related species (Table 5.1).

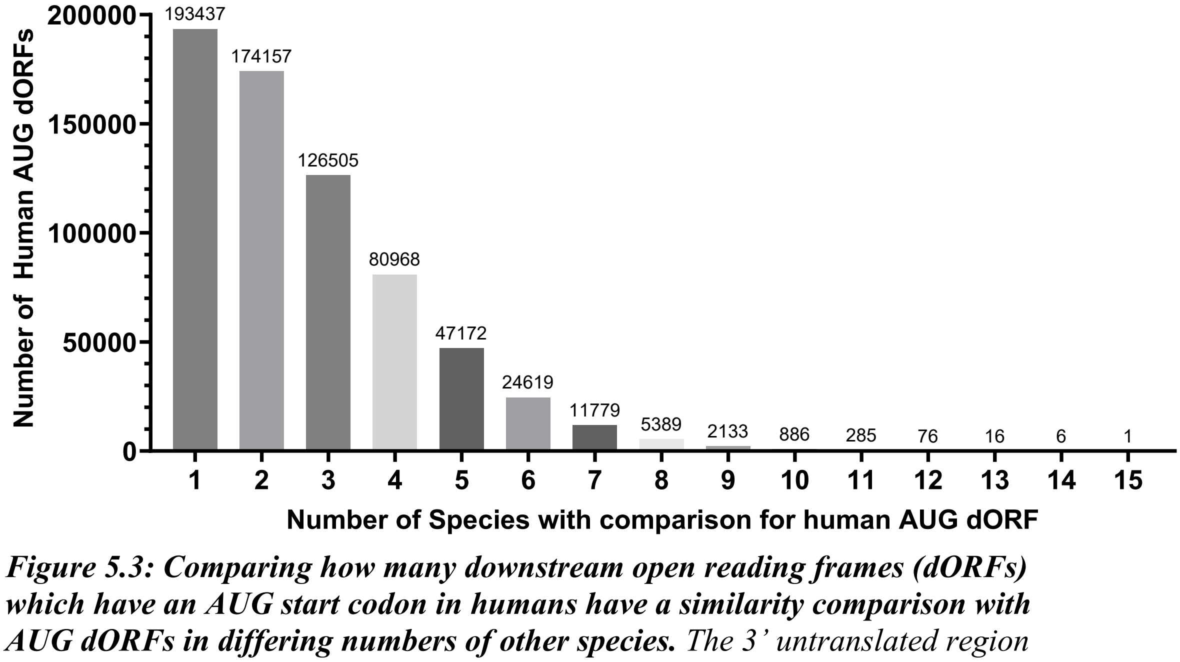

Figure 5.3: Comparing how many downstream open reading frames (dORFs) which have an AUG start codon in humans have a similarity comparison with AUG dORFs in differing numbers of other species. *The 3' untranslated region (UTR) sequences for each species have had duplicate sequences from different transcripts removed. The bars represent the number of human AUG dORFs with a comparison, with different bars for the number of species the human AUG dORF has comparisons with. The number of human AUG dORFs is included above each bar.*

As discussed previously, conservation of sORFs and 3’ UTR regions can indicate evolutionary pressure to maintain these, implying function (Bashirullah, Cooperstock and Lipshitz, 1998; Conne, Stutz and Vassalli, 2000; Grzybowska, Wilczynska and Siedlecki, 2001; Pesole *et al.*, 2001; Mignone *et al.*, 2002; Shabalina *et al.*, 2004; Martinez *et al.*, 2019). dORFs are presented as translational regulators (Wu *et al.*, 2020b), and if this function extends beyond Wu dORFs, many dORFs could be conserved across species due to this regulatory function. Although more conserved than 5’ UTRs, generally full 3’ UTRs are not associated with conservation, instead particular regions are (Siepel *et al.*, 2005; Litterman *et al.*, 2019). The similarity percentage represents the proportion of the aligned sequences that match nucleotides, and the alignment length was no shorter than the shorter of the two AUG dORFs, or 3’ UTRs, compared. Comparing dORF and 3’ UTR similarity indicated whether dORFs were more conserved than their 3’ UTRs, and whether 3' UTR conservation explained dORF conservation. The violin plots show the distribution of similarity percentages of AUG dORFs and 3’ UTRs between humans and each species, with the median, 25th, and 75th percentiles included. The violin plots suggest that across the species AUG dORF similarity was greater than their 3’ UTRs when comparing the human sequences with those of the other species (Figure 5.4). In all species, AUG dORF violins were shifted towards greater similarity percentages, also shown by shifted median and percentile lines (Figure 5.4). Although not seen in overall trends, the violins indicated the range of similarity percentages, suggesting individual dORF similarity can be reduced compared to its 3’ UTR. In all vertebrate species, up to *D. rerio,* some AUG dORFs had 100% similarity, also seen in 3’ UTRs up to *G. gallus* (Figure 5.4). Across the species, the smallest similarity percentages were generally between 15% and 30% (Figure 5.4). The range of AUG dORF similarities reduced as species were more distantly related to humans. The species from invertebrates to plants show consistent dORF similarity, with the median around 50% (Figure 5.4). The violins show most AUG dORFs had similarity of around 100% in *P. troglodytes*, which reduced slightly in *M. mulatta* (Figure 5.4). Once outside of primates, in the remaining vertebrates, the proportion of AUG dORFs with greater similarity percentages reduced with most similarity around 60% (Figure 5.4). 3’ UTR similarity shared a similar pattern to AUG dORFs, 3’ UTR similarity was greatest in primates, before reducing across the remaining mammals, to then become relatively consistent in remaining species with a median around 40% similarity (Figure 5.4). Across the species, violin plots suggest the distribution of 3’ UTR similarity percentages reduced compared to AUG dORFs (Figure 5.4). In mammals, the binomial distribution, shown with two bulges on some violin plots (Figure 5.4), showed two potential groups of 3’ UTRs or AUG dORFs with greater, or reduced, similarity percentages. AUG dORFs appeared to be more conserved than their 3’ UTRs in all species considered. Both AUG dORF and 3’ UTR sequence similarity reduced between humans and homolog species as homolog species became more distantly related, but dORFs still had increased similarity.

To provide additional context to the results in Figure 5.4, the mean 3’ UTR and AUG dORF similarity percentage between humans and each homolog species were compared. The results from the mean comparisons in Table 5.2, support the Figure 5.4 results. The mean AUG dORF similarity percentage between the human and homolog sequences was greater than the mean 3’ UTR similarity percentage across all species analysed, and in all cases this difference was statistically significant (P<0.0001) (Table 5.2). As species become less closely related to humans, moving down the table, the mean AUG dORF and 3’ UTR similarity percentage reduced alongside the number of comparisons (Table 5.2). The difference between AUG dORF and 3’ UTR mean similarity was relatively consistent across species, only *M. mulatta* (4.035%), *B. taurus* (7.787%), *R. norvegicus* (9.246%), and *C. elegans* (9.931%), had a difference below ten percent (Table 5.2). There were more AUG dORF comparisons than 3’ UTR comparisons (Table 5.2) as multiple AUG dORFs were often found in human and homolog 3 UTRs. AUG dORF and 3’ UTR conservation were shown to reduce as species are more distantly related to humans. The mean similarities suggest that dORFs are more conserved than their 3’ UTRs.

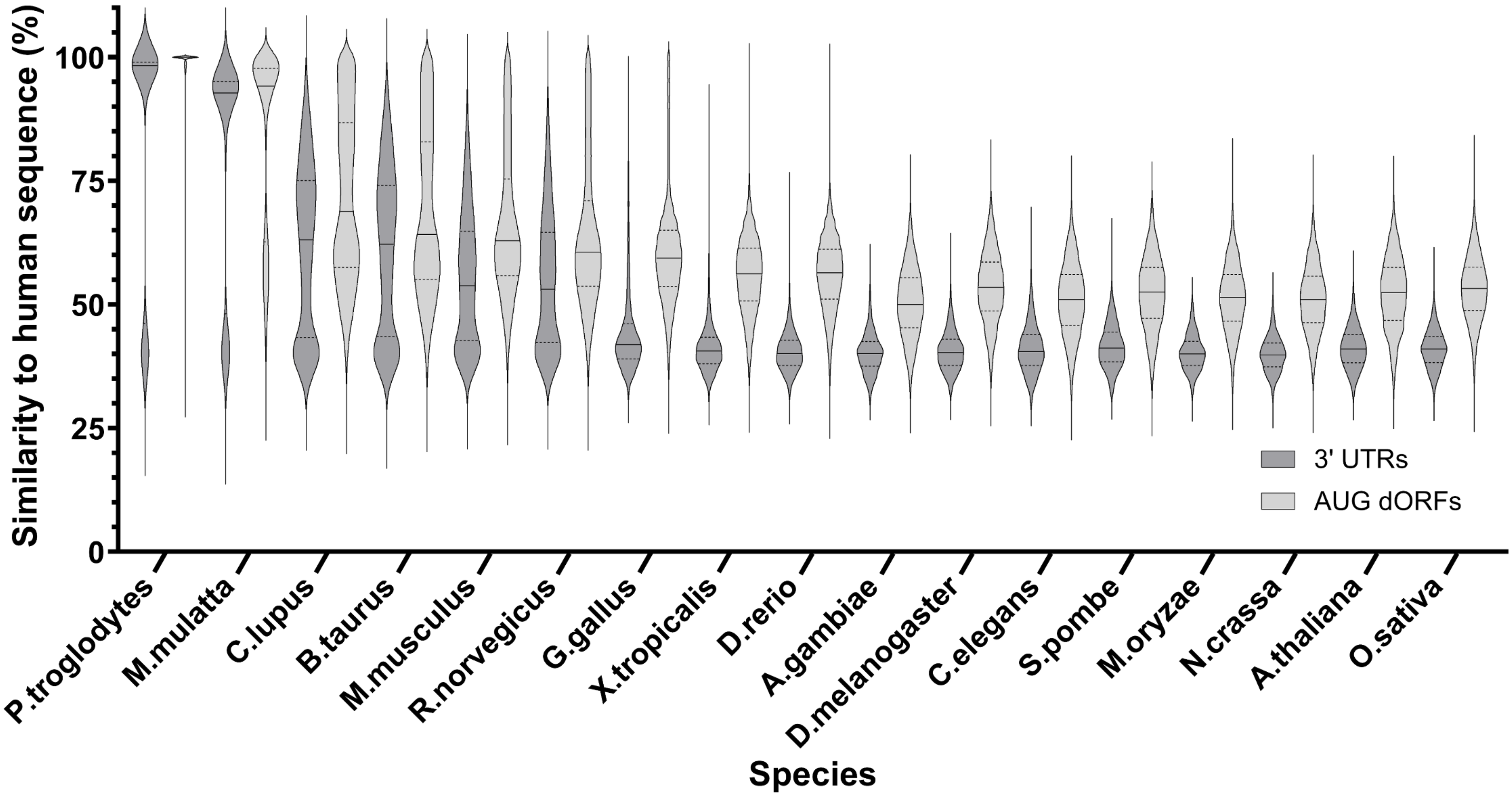

Figure 5.4: The similarity of downstream open reading frame (dORF) sequences which have an AUG start codon between humans and other species appears to be greater than the similarity of the 3' untranslated region (UTR) sequences containing the AUG dORFs. *The similarity of human and homolog AUG dORF and 3' UTR sequences is included for a range of species. Violin plots, generated in GraphPad Prism (10.2.3), were used to show the distribution of the similarity percentages, the horizontal line within the violin marks the median and the dashed lines mark the 25th to the 75th percentile of the values distribution. Violin plots in darker grey show the similarity percentages in 3' UTRs and the lighter grey for AUG dORFs.*

Table 5.2: The mean similarity of downstream open reading frame (dORF) sequences which start with an AUG start codon is increased compared to the 3' untranslated region (UTR) sequences containing the AUG dORFs between humans and various species. *One-Way ANOVA multiple comparisons test with Šídák adjusted P values was used to compare the means and the data generated by this test are included in the table.*

Species	Mean 3' UTR similarity to human	Mean AUG dORF similarity to human	Mean Difference	Standard Error of Difference	95% Confidence Intervals of difference	3' UTR comparisons	AUG dORF comparisons	Adjusted P Value
P. troglodytes	82.28	92.46	10.18	0.09182	10.45 to 9.904	27564	350026	<0.0001
M. mulatta	79.44	83.47	4.035	0.1029	4.340 to 3.729	22190	241772	<0.0001
C. lupus	61.16	71.54	10.38	0.1005	10.68 to 10.08	24454	167113	<0.0001
B. taurus	60.57	68.36	7.787	0.1043	8.096 to 7.477	22589	160201	<0.0001
M. musculus	54.96	66.2	11.24	0.1032	11.55 to 10.94	23718	138168	<0.0001
R. norvegicus	54.58	63.83	9.246	0.103	9.552 to 8.941	23464	150024	<0.0001
G. gallus	44.4	60.69	16.29	0.1204	16.65 to 15.94	17459	100081	<0.0001
X. tropicalis	41.42	56.14	14.72	0.1179	15.07 to 14.37	17514	135091	<0.0001
D. rerio	40.5	56.01	15.51	0.1253	15.88 to 15.14	15513	119397	<0.0001
A. gambiae	40.21	50.43	10.22	0.3312	11.21 to 9.242	2146	23178	<0.0001
D. melanogaster	40.63	53.51	12.88	0.2555	13.64 to 12.13	3640	35381	<0.0001
C. elegans	41.03	50.96	9.931	0.2911	10.80 to 9.068	2740	35079	<0.0001
S. pombe	41.57	52.28	10.71	0.3886	11.86 to 9.559	1539	19479	<0.0001
M. oryzae	40.31	51.41	11.1	0.3516	12.14 to 10.06	1909	19929	<0.0001
N. crassa	39.97	51.09	11.12	0.3889	12.27 to 9.962	1547	17998	<0.0001
A. thaliana	41.18	52.18	10.99	0.2762	11.81 to 10.18	3041	39584	<0.0001
O. sativa	41.01	53.14	12.13	0.2896	12.99 to 11.27	2778	34187	<0.0001

As seen previously, longer 3’ UTRs are associated with more dORF sequences. To investigate whether human 3’ UTR length impacts the mean similarity results, analysis was repeated with human 3’ UTR sequences restricted to maximum lengths of 2500, 5000 and 10000 nucleotides (Appendix 3 in Tables 8.6-8.8). Other than changing the number of AUG dORF and 3’ UTR comparisons, the results were similar to Table 5.2. In all species, the mean AUG dORF similarity was greater than the 3’ UTR, and these differences between means were statistically significant (P<0.0001) (Tables 8.6-8.8). Restricting the length of human 3’ UTRs did not influence the mean similarity results obtained for 3’ UTRs and AUG dORFs.

The consistency of differences between mean 3’ UTR and AUG dORF similarity (Table 5.2) and shifted distributions in Figure 5.4 led the investigation into whether these differences showed a normal distribution or were skewed by populations of very high or low differences. Instead of comparing 3’ UTR and AUG dORF similarity, between human sequences and other species, separately, the differences between the similarity of each AUG dORF and the AUG-dORF containing 3’ UTRs were plotted as violin plots. The violin plots showed normal distributions across all species, and other than *P. troglodytes*, and *M. mulatta*, the distributions were similar (Figure 5.5). The positive and negative values show that AUG dORF similarity can be much greater, or smaller, than the AUG dORF-containing 3’ UTR similarity (Figure 5.5). In all species, only a small minority of comparisons had increased 3’ UTR similarity compared to the AUG dORFs. Across the species most AUG dORFs are more conserved than the 3’ UTRs containing them. The median vales in Figure 5.5 were similar to the mean difference values in Table 5.2, except for *P. troglodytes* where 3’ UTR and AUG dORF similarity percentages were very similar (Figure 5.5). The difference between AUG dORF and 3’ UTR similarity is consistent and had normal distribution across most species.

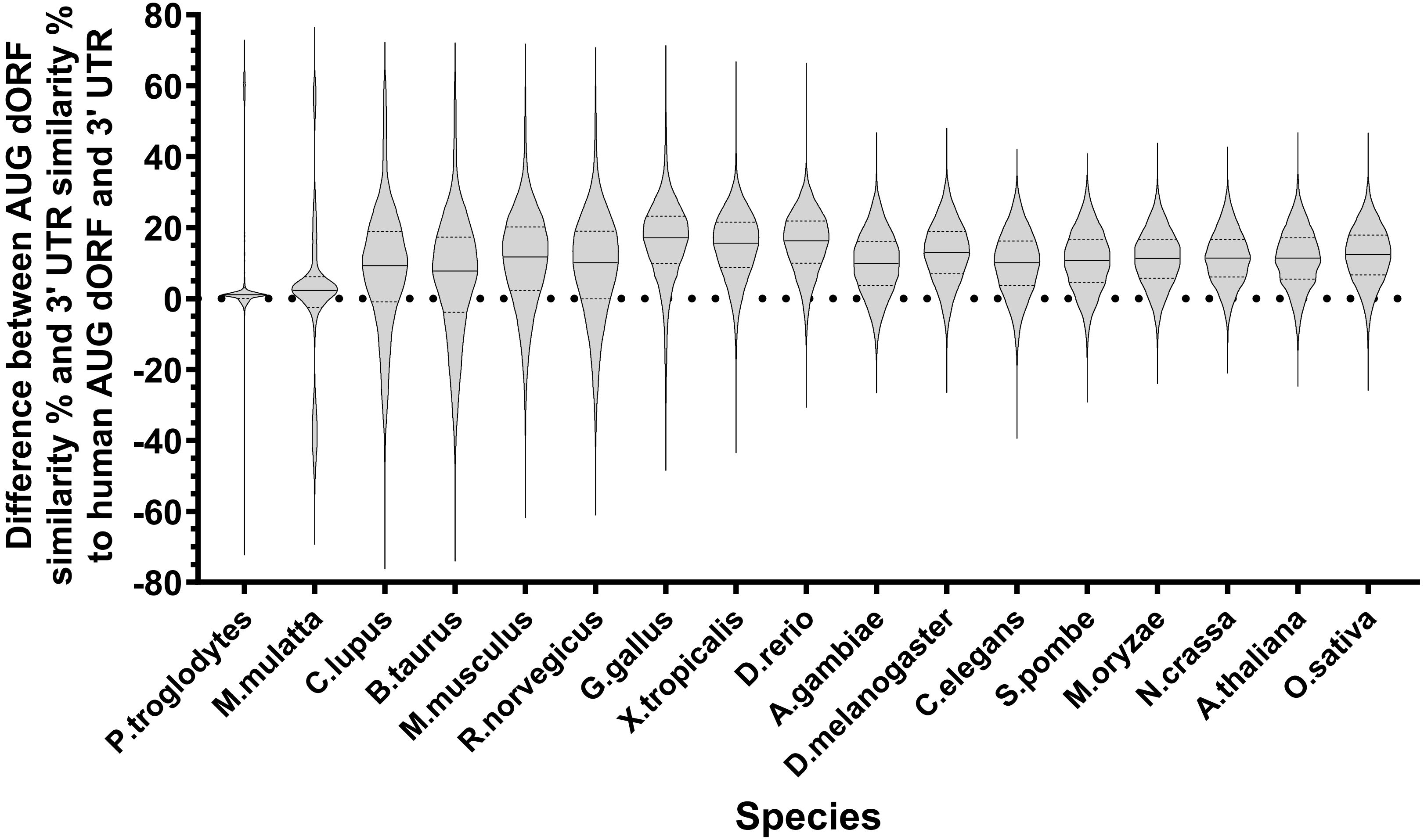

Figure 5.5: Comparing the difference between AUG start codon downstream open reading frame (dORF) and 3' untranslated region (UTR) sequence similarity across species. *For each AUG dORF similarity comparison in each species the difference between this similarity percentage and the 3' UTR sequence similarity containing this AUG dORF is included. Violin plots, generated in GraphPad Prism (10.2.3), were used to show the distribution of the difference in similarity percentages, the horizontal line within the violin marks the median and the dashed lines mark the 25th to the 75th percentile of the values distribution.*

5.3 Validation of conservation results with control sequences

dORFs appear to be more conserved than 3' UTRs across a wide range of species. Although this is a useful comparison, the short nature of dORFs within 3' UTRs could make these regions more likely to be conserved. Conservation in 3' UTRs is suggested to usually be short regions rather than the whole sequence (Siepel *et al.*, 2005; Litterman *et al.*, 2019). Aside from being longer, 3' UTRs vary in length between humans and other species and 3' UTR conservation could also be influenced by other regulatory regions. To overcome these considerations, several control sequences and adjustments were made to validate the dORF conservation findings. Initially, it was considered whether the AUG start codon present in every dORF being compared could influence the overall dORF similarity. In short dORFs of 30 to 60 nucleotides the start codon would make up five to ten percent of the whole sequence, potentially explaining some of the difference between AUG dORF and 3' UTR similarity. Analysis was repeated with AUG start codons discounted from similarity comparisons. Removing the AUGs from the analysis produced fewer comparisons (Table 5.3) when contrasted with Table 5.2, where alignment length, or score, fell below required levels. In most species the differences between mean AUG dORF similarity with and without the AUG start codon were statistically significant ($P<0.05$) (Table 5.3). However, these differences were very small, across the species the reduction in AUG dORF mean similarity percentage, when AUG start codons were removed, ranged from 0.173% to 1.384% (Table 5.3). Although including the AUG start codon in the similarity analysis slightly increased the similarity, this was a very small effect and does not explain the difference between AUG dORF and 3' UTR conservation.

Table 5.3: Comparing the mean similarity percentages of AUG start codon downstream open reading frame (dORF) sequences when the AUG start codon is included, or excluded, from the similarity analysis between humans and various species. *One-Way ANOVA multiple comparisons test with Šídák adjusted P values was used to compare the means and the data generated by this test are included in the table.*

Species	**Mean AUG dORF similarity to human (AUG excluded)**	**Mean AUG dORF similarity to human (AUG included)**	**Mean Difference**	**Standard Error of Difference**	**95% Confidence Intervals of difference**	**AUG dORFs compared**	**Adjusted P Value**
P. troglodytes	93.45	93.62	0.1734	0.03599	0.06664 to 0.2802	336699	<0.0001
M. mulatta	84.77	85.17	0.4042	0.04448	0.2722 to 0.5361	220426	<0.0001
C. lupus	71.82	72.6	0.7802	0.0555	0.6156 to 0.9449	141596	<0.0001
B. taurus	68.49	69.28	0.786	0.05703	0.6168 to 0.9552	134094	<0.0001
M. musculus	65.85	66.89	1.034	0.06201	0.8500 to 1.218	113443	<0.0001
R. norvegicus	63.45	64.41	0.9547	0.05974	0.7774 to 1.132	122203	<0.0001
G. gallus	59.37	60.76	1.384	0.07411	1.165 to 1.604	79414	<0.0001
X. tropicalis	54.73	55.83	1.109	0.06428	0.9179 to 1.299	105574	<0.0001
D. rerio	54.52	55.68	1.155	0.06885	0.9512 to 1.360	92005	<0.0001
A. gambiae	49.69	50.04	0.3513	0.1566	-0.1134 to 0.8160	17777	0.3488
D. melanogaster	52.63	53.18	0.5484	0.1277	0.1697 to 0.9271	26765	0.0003
C. elegans	50.31	50.6	0.2892	0.1254	-0.08276 to 0.6611	27749	0.304
S. pombe	51.69	52.02	0.3339	0.1724	-0.1776 to 0.8455	14669	0.6024
M. oryzae	50.42	50.97	0.5521	0.1655	0.06110 to 1.043	15918	0.0144
N. crassa	50.21	50.7	0.487	0.1736	-0.02788 to 1.002	14481	0.082
A. thaliana	51.27	51.87	0.6063	0.1189	0.2535 to 0.9591	30837	<0.0001
O. sativa	52.09	52.81	0.7174	0.1287	0.3357 to 1.099	26349	<0.0001

To validate the 3’ UTR and dORF conservation findings, control sequences were used to investigate the inter-species conservation of short regions of human 3’ UTRs, of AUG dORF length, rather than whole 3’ UTRs. In human 3’ UTRs reported in the AUG dORFs and 3’ UTRs similarity analysis, the AUG dORF with the greatest similarity percentage was selected and a control sequence matching the length of that AUG dORF was used from the centre of the 3’ UTR. Then similarity analysis comparing the control sequence against the whole homologous 3’ UTR was done, investigating the control sequence conservation anywhere in the homologous 3’ UTR. The control sequence similarity was then compared to 3’ UTR similarity and the most similar AUG dORF. The violin plots across the species showed that control sequence similarity was reduced compared to AUG dORFs (Figure 5.6). This was supported by the shifted median and percentile lines (Figure 5.6). The violin plots also suggest that in more closely related species, up to *R. norvegicus*, where AUG dORF and 3’ UTR similarity were higher, control sequence similarity was reduced compared to 3’ UTRs (Figure 5.6). However, in all other species, control sequence similarity was increased compared to 3’ UTRs but reduced compared to AUG dORFs (Figure 5.6). In all species, all the differences between 3’ UTR, control sequence and AUG dORF mean similarity were statistically significant (P<0.0001) (Table 5.4). In mammalian species the control sequence mean similarity reduced, as seen with 3’ UTRs and AUG dORFs, as the species become more distantly related (Figure 5.6 and Table 5.4). Then results for the mean and distributions remained more similar in remaining species (Figure 5.6 and Table 5.4). The mean control similarity in mammals was reduced compared to 3’ UTRs and AUG dORFs, and in all other species the mean control sequence similarity was reduced compared to AUG dORFs, but greater than 3’ UTRs (Table 5.4). Conservation of short, AUG dORF length, 3’ UTR regions in homolog species appears to be reduced compared to AUG dORFs, and 3’ UTRs in some species, suggesting that increased AUG dORF conservation, and increased conservation compared to 3’ UTRs, is not explained by AUG dORFs being short 3’ UTR regions that may be more likely to be conserved.

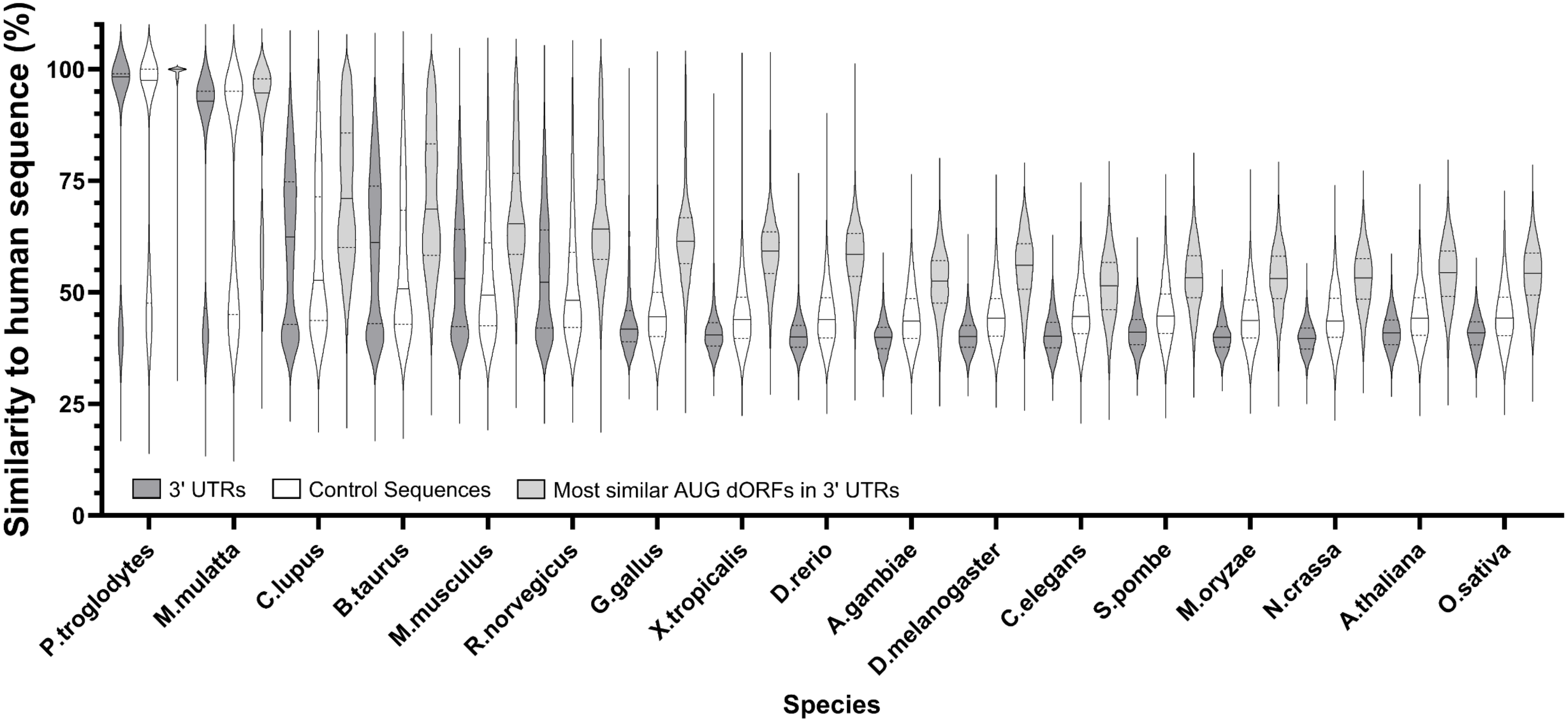

Figure 5.6: Downstream open reading frame (dORF) sequences which use an AUG start codon are more similar than 3' untranslated region (UTR) sequences containing these dORFs and control sequences, when compared between humans and other species. *Control sequences are generated in human 3' UTR sequences containing AUG dORFs and are sequences of the same length as the AUG dORF taken from the centre of the 3' UTR. The control sequence similarity percentage for each species is the similarity of the human control sequence compared against the whole 3' UTR of the homolog species, originally compared against for dORF and 3' UTR similarity. Then the similarity percentages between the human and homolog species of the 3' UTR sequence, containing the human control sequence, and the most similar AUG dORF in the 3' UTR are also included for comparison. Violin plots, generated in GraphPad Prism (10.2.3), were used to show the distribution of the similarity percentages, the horizontal line within the violin marks the median and the dashed lines mark the 25th to the 75th percentile of the values distribution. Violin plots in darker grey show the similarity percentages in the 3' UTRs, lighter grey for the AUG dORFs, and white for controls.*

Table 5.4: The mean similarity percentage of downstream open reading frame (dORF) sequences which use an AUG start codon is greater than the 3' untranslated region (UTR) sequences containing these dORFs, and control sequences between humans and other species. *Control sequences are generated in the human 3' UTR sequences containing AUG dORFs and are sequences of the same length as the AUG dORF taken from the centre of the 3' UTR. The control sequence similarity percentage for each species is the similarity of the human control sequence compared against the whole 3' UTR of the homolog species. Then the mean similarity percentages between the human and homolog species of the 3' UTR sequence, containing the human control sequence, and the most similar AUG dORF in the 3' UTR are also included for comparison. One-Way ANOVA multiple comparisons test with Šídák adjusted P values was used to compare the means and the data generated by this test are included in the table.*

Species	Mean 3' UTR similarity to human	Mean Control similarity to human	Mean Top AUG dORF similarity to human	Number of each sequence	Adjusted P Value (3' UTR vs Control)	Adjusted P Value (3' UTR vs dORF)	Adjusted P Value (Control vs dORF)
P. troglodytes	82.61	77.93	92.59	25103	<0.0001	<0.0001	<0.0001
M. mulatta	79.19	70.33	85.81	19443	<0.0001	<0.0001	<0.0001
C. lupus	60.7	57.86	72.44	21130	<0.0001	<0.0001	<0.0001
B. taurus	60.1	56.05	70.72	19656	<0.0001	<0.0001	<0.0001
M. musculus	54.44	53.11	68.04	21192	<0.0001	<0.0001	<0.0001
R. norvegicus	54.1	52.01	66.94	20673	<0.0001	<0.0001	<0.0001
G. gallus	44.24	46.34	62.98	16210	<0.0001	<0.0001	<0.0001
X. tropicalis	41.3	44.92	59.28	16011	<0.0001	<0.0001	<0.0001
D. rerio	40.38	44.7	58.15	13979	<0.0001	<0.0001	<0.0001
A. gambiae	39.95	44.42	52.24	1747	<0.0001	<0.0001	<0.0001
D. melanogaster	40.31	44.82	55.58	3104	<0.0001	<0.0001	<0.0001
C. elegans	40.66	45.29	51.34	2100	<0.0001	<0.0001	<0.0001
S. pombe	41.29	45.63	53.41	1275	<0.0001	<0.0001	<0.0001
M. oryzae	40.16	44.3	52.97	1683	<0.0001	<0.0001	<0.0001
N. crassa	39.76	44.42	53.07	1365	<0.0001	<0.0001	<0.0001
A. thaliana	41.16	44.86	54.15	2642	<0.0001	<0.0001	<0.0001
O. sativa	40.86	44.89	54.01	2462	<0.0001	<0.0001	<0.0001

The final validation of dORF conservation results also used control sequences to investigate the conservation of regions surrounding AUG dORFs. Conservation of surrounding regions could indicate that dORFs were part of larger conserved regions, or that the surrounding regions are important in dORF function. An example could be a region upstream of dORFs to help recruit ribosomes for dORF translation (Wu *et al.*, 2020b). For each AUG dORF with similarity analysis results, a control sequence the same length as the dORF was collected in the human and homolog species either ending 100, 200 or 500 bases upstream of the AUG dORF start, or starting 100, 200 or 500 nucleotides downstream of the dORF stop codon (Figure 2.8). Then human and homolog control sequence similarity was analysed with comparison to AUG dORF and 3' UTR similarity. Except for varying numbers of results, the results for the different control sequence 3' UTR locations were consistent (Tables 5.5 and 5.6). In mammalian species, the mean control sequence similarities were reduced compared to 3' UTR and AUG dORF similarities (Tables 5.4 and 5.6). In all other species the mean control sequence similarities were increased compared to 3' UTRs, but were reduced compared to AUG dORFs (Tables 5.5 and 5.6). All differences between control sequence and 3' UTR, or AUG dORF, means were statistically significant (P<0.0001). These results were supported by violin plots of the different control sequences relative to AUG dORF and 3' UTR similarity in Appendix 4. These violin plots also show that control sequence similarity was reduced compared to 3' UTRs and AUG dORFs in mammalian species (Figures 8.1 and 8.2), then increased compared to 3' UTRs and still reduced compared to AUG dORFs in all other species (Figures 8.3-8.6). Similar to 3' UTR and AUG dORF similarity, control sequence similarity reduced across vertebrate species as they become more distantly related, then remained more consistent in the other species (Tables 5.5 and 5.6, Figures 8.1-8.6). AUG dORFs do not appear to be part of larger conserved 3' UTR regions. In mammalian species, sequences surrounding dORFs were less conserved than the 3' UTR and AUG dORFs. In all other species dORFs were more conserved than the surrounding regions, although the surrounding regions were slightly more conserved than 3' UTRs.

Table 5.5: The mean similarity percentage of downstream open reading frame (dORF) sequences which use an AUG start codon is increased compared to the 3' untranslated region (UTR) sequences containing these dORFs, and upstream control sequences between humans and other species. *Control sequences are generated in the human and homolog 3' UTR sequences containing AUG dORFs and are sequences of the same length as the AUG dORF which end either, 500, 200, or 100 nucleotides upstream of the AUG dORF start codon in the 3' UTRs. The table is split to include the results for the different control sequence locations relative to the AUG dORF. The control sequence similarity percentage for each species is the similarity of the human control sequence compared against the homolog species control sequence. For comparison the mean similarity percentage between human and homolog species is included for the 3' UTR sequences containing the control sequence and the AUG dORF downstream of the control sequences. One-Way ANOVA multiple comparisons test with Šídák adjusted P values was used to compare the means and the data generated by this test are included in the table.*

	500 Upstream							200 Upstream							100 Upstream						
	Mean similarity to human			Number of each sequence	Adjusted P Value			Mean similarity to human			Number of each sequence	Adjusted P Value			Mean similarity to human			Number of each sequence	Adjusted P Value		
Species	3' UTR	Control	AUG dORF		3' UTR vs Control	3' UTR vs dORF	Control vs dORF	3' UTR	Control	AUG dORF		3' UTR vs Control	3' UTR vs dORF	Control vs dORF	3' UTR	Control	AUG dORF		3' UTR vs Control	3' UTR vs dORF	Control vs dORF
P. troglodytes	90.98	74.47	96.12	89387	<0.0001	<0.0001	<0.0001	91.57	77.81	95.22	135947	<0.0001	<0.0001	<0.0001	91.65	78.72	94.67	160063	<0.0001	<0.0001	<0.0001
M. mulatta	85.41	62.88	87.75	39168	<0.0001	<0.0001	<0.0001	84.7	65.85	85.05	67008	<0.0001	0.0251	<0.0001	85.05	68.13	84.75	84654	<0.0001	0.0382	<0.0001
C. lupus	65.22	54.45	76.39	32867	<0.0001	<0.0001	<0.0001	64.28	54.95	73.5	48608	<0.0001	<0.0001	<0.0001	63.95	55.88	72.45	58764	<0.0001	<0.0001	<0.0001
B. taurus	64.08	53.95	74.1	26358	<0.0001	<0.0001	<0.0001	62.32	53.78	70.33	42238	<0.0001	<0.0001	<0.0001	62.3	54.51	69.25	51369	<0.0001	<0.0001	<0.0001
M. musculus	57.37	52.02	71.37	27679	<0.0001	<0.0001	<0.0001	56.11	52.15	68.51	42802	<0.0001	<0.0001	<0.0001	55.88	52.68	67.62	51259	<0.0001	<0.0001	<0.0001
R. norvegicus	56.61	51.71	70.26	24280	<0.0001	<0.0001	<0.0001	56.06	51.8	67.34	39513	<0.0001	<0.0001	<0.0001	55.63	52	65.94	48896	<0.0001	<0.0001	<0.0001
G. gallus	45.34	49.49	65.71	17365	<0.0001	<0.0001	<0.0001	44.37	49.65	63.23	30483	<0.0001	<0.0001	<0.0001	44.36	49.64	62.36	37833	<0.0001	<0.0001	<0.0001
X. tropicalis	40.85	48.91	62.21	13241	<0.0001	<0.0001	<0.0001	40.87	48.84	60.12	28096	<0.0001	<0.0001	<0.0001	40.99	48.85	59.07	37630	<0.0001	<0.0001	<0.0001
D. rerio	39.37	48.91	61.67	9637	<0.0001	<0.0001	<0.0001	39.5	48.87	59.42	23409	<0.0001	<0.0001	<0.0001	39.65	48.85	58.41	33223	<0.0001	<0.0001	<0.0001
A. gambiae	38.09	47.15	58.1	229	<0.0001	<0.0001	<0.0001	38.96	47.41	54.44	1513	<0.0001	<0.0001	<0.0001	39.03	47.7	53.39	2700	<0.0001	<0.0001	<0.0001
D. melanogaster	38.9	49.46	60.27	1266	<0.0001	<0.0001	<0.0001	39.24	48.53	57.33	4163	<0.0001	<0.0001	<0.0001	39.44	48.58	56.01	6614	<0.0001	<0.0001	<0.0001
C. elegans	37.82	47.79	56.12	670	<0.0001	<0.0001	<0.0001	38.59	47.59	53.39	3197	<0.0001	<0.0001	<0.0001	39.2	48.4	53	5878	<0.0001	<0.0001	<0.0001
S. pombe	38.68	48.43	55.51	404	<0.0001	<0.0001	<0.0001	39.8	49.06	55.8	1208	<0.0001	<0.0001	<0.0001	40.07	49.2	54.97	2510	<0.0001	<0.0001	<0.0001
M. oryzae	37.8	48.11	56.61	236	<0.0001	<0.0001	<0.0001	39.03	47.51	54.92	1369	<0.0001	<0.0001	<0.0001	39.31	47.29	53.95	2823	<0.0001	<0.0001	<0.0001
N. crassa	38.27	47.68	56.94	148	<0.0001	<0.0001	<0.0001	38.52	47.83	55.24	1186	<0.0001	<0.0001	<0.0001	39.27	47.6	53.64	2596	<0.0001	<0.0001	<0.0001
A. thaliana	39.56	48.48	58.63	235	<0.0001	<0.0001	<0.0001	39.72	48.59	56.51	1659	<0.0001	<0.0001	<0.0001	40.79	48.52	55.56	4516	<0.0001	<0.0001	<0.0001
O. sativa	38.57	48.96	59.6	191	<0.0001	<0.0001	<0.0001	39.11	48.1	56.52	1967	<0.0001	<0.0001	<0.0001	39.61	48.44	55.63	4511	<0.0001	<0.0001	<0.0001

Table 5.6: The mean similarity percentage of downstream open reading frame (dORF) sequences which use an AUG start codon was increased compared to the 3' untranslated region (UTR) sequences containing these dORFs, and downstream control sequences between humans and other species. *Control sequences are generated in the human and homolog 3' UTR sequences containing AUG dORFs and are sequences of the same length as the AUG dORF which start either, 100, 200, or 500 nucleotides downstream of the AUG dORF stop codon in the 3' UTRs. The table is split to include the results for the different control sequence locations relative to the AUG dORF. The control sequence similarity percentage for each species is the similarity of the human control sequence compared against the homolog species control sequence. For comparison the mean similarity percentage between human and homolog species is included for the 3' UTR sequences containing the control sequence and the AUG dORF upstream of the control sequences. One-Way ANOVA multiple comparisons test with Šídák adjusted P values was used to compare the means and the data generated by this test are included in the table.*

	100 Downstream							**200 Downstream**							**500 Downstream**						
	Mean similarity to human			**Number of each sequence**	**Adjusted P Value**			**Mean similarity to human**			**Number of each sequence**	**Adjusted P Value**			**Mean similarity to human**			**Number of each sequence**	**Adjusted P Value**		
Species	**3' UTR**	**Control**	**AUG dORF**		**3' UTR vs Control**	**3' UTR vs dORF**	**Control vs dORF**	**3' UTR**	**Control**	**AUG dORF**		**3' UTR vs Control**	**3' UTR vs dORF**	**Control vs dORF**	**3' UTR**	**Control**	**AUG dORF**		**3' UTR vs Control**	**3' UTR vs dORF**	**Control vs dORF**
P. troglodytes	92.82	86.89	95.75	181390	<0.0001	<0.0001	<0.0001	92.87	86.76	96.3	153811	<0.0001	<0.0001	<0.0001	92.84	85.46	97.17	99211	<0.0001	<0.0001	<0.0001
M. mulatta	86.31	74.13	86.38	83400	<0.0001	>0.9999	<0.0001	86.54	72.89	87.38	62874	<0.0001	<0.0001	<0.0001	86.97	70.26	89.25	35942	<0.0001	<0.0001	<0.0001
C. lupus	63.74	55.68	72.65	53644	<0.0001	<0.0001	<0.0001	64.29	55.14	73.92	43548	<0.0001	<0.0001	<0.0001	65.3	54.58	76.12	29921	<0.0001	<0.0001	<0.0001
B. taurus	62.28	54.36	69.51	47513	<0.0001	<0.0001	<0.0001	62.35	53.97	70.91	37409	<0.0001	<0.0001	<0.0001	63.65	54.25	74.56	22829	<0.0001	<0.0001	<0.0001
M. musculus	55.68	52.41	67.87	47974	<0.0001	<0.0001	<0.0001	56.25	51.99	68.93	40111	<0.0001	<0.0001	<0.0001	57.48	51.91	71.61	26318	<0.0001	<0.0001	<0.0001
R. norvegicus	55.47	51.52	66.29	45302	<0.0001	<0.0001	<0.0001	55.81	51.61	67.55	36462	<0.0001	<0.0001	<0.0001	56.92	51.43	70.74	22638	<0.0001	<0.0001	<0.0001
G. gallus	44.37	49.82	62.55	34592	<0.0001	<0.0001	<0.0001	44.32	49.38	63.24	28270	<0.0001	<0.0001	<0.0001	44.55	49.26	65.34	17115	<0.0001	<0.0001	<0.0001
X. tropicalis	40.88	48.85	58.47	37291	<0.0001	<0.0001	<0.0001	40.7	48.87	59.79	25735	<0.0001	<0.0001	<0.0001	40.91	49	62.4	11471	<0.0001	<0.0001	<0.0001
D. rerio	39.64	48.99	58.26	31670	<0.0001	<0.0001	<0.0001	39.41	48.93	59.25	20980	<0.0001	<0.0001	<0.0001	39.21	49.2	61.12	7960	<0.0001	<0.0001	<0.0001
A. gambiae	39.05	48.21	53.14	3300	<0.0001	<0.0001	<0.0001	38.99	48.3	54.06	1783	<0.0001	<0.0001	<0.0001	37.25	49.33	55.92	176	<0.0001	<0.0001	<0.0001
D. melanogaster	39.65	48.39	55.74	6553	<0.0001	<0.0001	<0.0001	38.99	48.59	57.48	3572	<0.0001	<0.0001	<0.0001	38.53	49.31	60.22	1213	<0.0001	<0.0001	<0.0001
C. elegans	39.01	48.02	52.94	4801	<0.0001	<0.0001	<0.0001	38.99	47.53	53.02	2840	<0.0001	<0.0001	<0.0001	37.92	47.83	54.53	531	<0.0001	<0.0001	<0.0001
S. pombe	41.15	49.63	53.5	4113	<0.0001	<0.0001	<0.0001	41.29	50.19	53.68	2818	<0.0001	<0.0001	<0.0001	39.45	49.26	58.34	279	<0.0001	<0.0001	<0.0001
M. oryzae	39.34	47.44	53.52	3417	<0.0001	<0.0001	<0.0001	39.29	48.13	54.65	1512	<0.0001	<0.0001	<0.0001	38.25	47.69	56.24	246	<0.0001	<0.0001	<0.0001
N. crassa	38.78	47.72	53.77	2278	<0.0001	<0.0001	<0.0001	38.36	46.99	54.91	1146	<0.0001	<0.0001	<0.0001	38.34	46.71	57.77	175	<0.0001	<0.0001	<0.0001
A. thaliana	40.31	49.17	55.61	4833	<0.0001	<0.0001	<0.0001	39.55	48.87	56.2	2128	<0.0001	<0.0001	<0.0001	38.75	48.95	57.78	253	<0.0001	<0.0001	<0.0001
O. sativa	38.99	48.63	54.63	6357	<0.0001	<0.0001	<0.0001	37.42	47.96	54.85	2527	<0.0001	<0.0001	<0.0001	36.28	47.39	57.42	453	<0.0001	<0.0001	<0.0001

5.4 Conservation analysis dORF shortlist

To focus in on highly conserved AUG dORFs as candidates for further future analysis, a shortlist was developed known as HC dORFs (Highly conserved dORFs). The shortlist contains human AUG dORFs which were conserved with at least 90% sequence similarity in at least 5 homolog species. In all species, the dORF similarity also had to be greater than their 3' UTR similarity, and any available control sequence similarity for that dORF. This produced a shortlist of 122 AUG dORFs in 102 different genes. No HC dORFs were found to match MSVW dORFs, a shortlist of MS validated Wu dORFs developed from results in section 3.8, when aligned in Galaxy using the NCBI BLAST+ blastn tool (Altschul *et al.*, 1997; Camacho *et al.*, 2009; Cock *et al.*, 2015).

The HC dORFs, much like uORFs (Chew, Pauli and Schier, 2016; Johnstone, Bazzini and Giraldez, 2016; Dumesic *et al.*, 2019), other conserved sORFs, or short 3' UTR regions, could be conserved due to a regulatory function (Bashirullah, Cooperstock and Lipshitz, 1998; Conne, Stutz and Vassalli, 2000; Grzybowska, Wilczynska and Siedlecki, 2001; Pesole *et al.*, 2001; Mignone *et al.*, 2002; Shabalina *et al.*, 2004; Martinez *et al.*, 2019). This function could be translational regulation, which is proposed for Wu dORFs (Wu *et al.*, 2020b). The high sequence similarity could suggest that peptide function may also be conserved, seen previously with some dORF encoded peptides (Ji *et al.*, 2015). The potential conservation and implied function of HC dORFs led investigation into the 102 dORF-containing genes, looking for enrichment of gene ontologies, using methods described in section 2.9.3. Did these HC dORFs appear in particular genes, potentially genes with particular functions, or do they occur in a range of genes with various functions? To do this, the DAVID functional annotation tool (Sherman *et al.*, 2022), the g:Profiler g:GOSt functional profiling tool (Raudvere *et al.*, 2019), and the Gene Ontology Enrichment analysis and visualization tool (GOrilla) (Eden *et al.*, 2009) were used with varied results. When using the GOrilla tool no enrichment results were produced, suggesting no ontologies were enriched. The DAVID functional annotation tool creates functional annotation clusters, pulling together similar functional annotations, then providing an enrichment score for that annotation cluster, a higher score indicates greater enrichment (Sherman *et al.*, 2022). All functional annotations in Table 5.7 were enriched and had statistically significant P values (<0.05), and all but three annotations were statistically significant when adjusted with the Bonferroni test for multiple comparisons (<0.05). These results suggested there was enrichment of genes associated with transcription and transcriptional regulation, including around one third of the genes (Table 5.7). The second cluster suggested the genes were also enriched in terms of protein modifications to the encoded proteins (Table 5.7). Otherwise, there was suggested to be enrichment of genes associated with translational regulation and autism (Table 5.7). Although enriched, a maximum of eight genes were associated with these annotations (Table 5.7). Due to the small number of genes in the HC dORF shortlist, this meant only a few genes were enough suggest enrichment, however the biological relevance of that enrichment is potentially reduced. There was some similarity with functional annotation results using the g:Profiler g:GOSt functional profiling tool (Raudvere *et al.*, 2019) (Table 5.8), although only 99 of the genes were recognised by this tool. In all cases there was statistically significant enrichment in the abundance of molecular functions (Bonferroni adjusted P values <0.05) (Table 5.8). There was overlap and similarity between the functions reported, but Table 5.8 also suggested association with transcriptional regulation and DNA transcription. There was also association with mRNA binding (Table 5.8), which may relate to the translational regulation suggested by the DAVID tool (Table 5.7). The only other function relates to voltage-gated calcium channel activity, however only two genes were associated with this function (Table 5.7). There is some discrepancy between gene ontology tools regarding whether there was gene enrichment in the HC dORF shortlist. However, there was suggestion that there is an enrichment of genes associated with transcriptional and translational regulation.

Table 5.7: Functional annotation clusters for genes containing the highly conserved shortlist of downstream open reading frames (HC dORFs) which have an AUG start codon. *The DAVID functional annotation tool (Sherman et al., 2022) was used to generate this data, using a list of genes containing the HC dORF shortlist. Only the four clusters with the greatest enrichment score are reported. The clusters group similar functional annotations and report the relative enrichment of these functional annotations within the list of genes provided with associated proportion of genes matching the annotation, P values, Bonferroni adjusted P values and false discovery rate (FDR) values.*

Annotation Cluster 1 Enrichment Score: 5.68

Term	Genes matching term	Total Genes	Matching %	Fold Enrichment	P Value	Bonferroni	FDR
GO:0006357~regulation of transcription from RNA polymerase II promoter	29	102	28.43	3.53	3.67E-09	2.99E-06	2.97E-06
KW-0805~Transcription regulation	33	102	32.35	2.44	1.74E-07	5.21E-06	5.19E-06
KW-0804~Transcription	33	102	32.35	2.37	3.46E-07	1.04E-05	5.19E-06
GO:0005634~nucleus	53	102	51.96	1.81	1.09E-06	2.11E-04	2.02E-04
KW-0238~DNA-binding	29	102	28.43	2.33	7.00E-06	2.03E-04	1.01E-04
KW-0539~Nucleus	50	102	49.02	1.69	1.48E-05	3.85E-04	3.85E-04
GO:0000785~chromatin	17	102	16.67	3.18	6.83E-05	0.013168	0.00316
GO:0000981~RNA polymerase II transcription factor activity, sequence-specific DNA binding	18	102	17.65	2.76	2.05E-04	0.040376	0.009995

Annotation Cluster 2 Enrichment Score: 4.75

Term	Genes matching term	Total Genes	Matching %	Fold Enrichment	P Value	Bonferroni	FDR
CROSSLNK:Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in SUMO2)	22	102	21.57	3.47	8.29E-07	8.43E-04	2.09E-04
KW-1017~Isopeptide bond	28	102	27.45	2.25	4.90E-05	7.35E-04	3.92E-04
KW-0832~Ubl conjugation	34	102	33.33	1.90	1.43E-04	0.002136	6.44E-04

Annotation Cluster 3 Enrichment Score: 3.54

Term	Genes matching term	Total Genes	Matching %	Fold Enrichment	P Value	Bonferroni	FDR
GO:0010494~cytoplasmic stress granule	7	102	6.86	14.81	7.50E-06	0.001454	4.63E-04
GO:0000932~P-body	6	102	5.88	12.31	1.24E-04	0.023708	0.004133
KW-0810~Translation regulation	6	102	5.88	7.89	8.90E-04	0.026347	0.008896
GO:0017148~negative regulation of translation	5	102	4.90	10.03	0.001522	0.71099	0.183982
GO:0003729~mRNA binding	7	102	6.86	5.51	0.001658	0.283619	0.040415

Annotation Cluster 4 Enrichment Score: 3.15

Term	Genes matching term	Total Genes	Matching %	Fold Enrichment	P Value	Bonferroni	FDR
KW-1268~Autism spectrum disorder	6	102	5.88	11.64	1.21E-04	0.002773	0.002656
GO:0014069~postsynaptic density	8	102	7.84	6.87	1.60E-04	0.030492	0.004218
KW-1269~Autism	3	102	2.94	13.65	0.018915	0.355449	0.138707

Table 5.8: The molecular functions associated with genes containing the highly conserved shortlist of downstream open reading frames (HC dORFs) which start with an AUG start codon. *The g:Profiler g:GOSt functional profiling tool (Raudvere et al., 2019) was used to generate this data, using a list of genes containing the HC dORF shortlist. This tool reported the enriched molecular function of the genes in the list provided with the number of genes associated with the function and the Bonferroni adjusted P values associated with the enrichment. Only statistically significant (P<0.05) molecular functions were reported.*

Source	Molecular Function	Genes Matching Function	Total Genes	Adjusted P value
GO:MF	DNA binding	39	99	4.62E-09
GO:MF	protein binding	94	99	9.13E-06
GO:MF	transcription regulator activity	29	99	1.14E-05
GO:MF	transcription regulatory region nucleic acid binding	24	99	5.02E-05
GO:MF	transcription cis-regulatory region binding	23	99	0.000201
GO:MF	sequence-specific DNA binding	24	99	0.000371
GO:MF	nucleic acid binding	49	99	0.000658
GO:MF	RNA polymerase II transcription regulatory region sequence-specific DNA binding	20	99	0.003744
GO:MF	mRNA binding	10	99	0.00487
GO:MF	DNA-binding transcription factor activity	20	99	0.006683
GO:MF	DNA-binding transcription factor activity, RNA polymerase II-specific	19	99	0.009547
GO:MF	miRNA binding	4	99	0.011467
GO:MF	RNA polymerase II cis-regulatory region sequence-specific DNA binding	17	99	0.02009
GO:MF	cis-regulatory region sequence-specific DNA binding	17	99	0.025507
GO:MF	voltage-gated calcium channel activity involved in AV node cell action potential	2	99	0.025781
GO:MF	regulatory RNA binding	4	99	0.035167

The same approach to investigate translational regulation used in chapter 3 was applied to transcripts containing HC dORFs, investigating whether the translational regulation suggested for Wu dORFs, increasing translation of CDSs in transcripts with dORFs (Wu *et al.*, 2020b), was seen with HC dORF transcripts. Using the healthy and cancerous datasets meant the reduced Wu dORF activity seen in cancer could be investigated with these HC dORFs. The modelled translational regulation of 31 transcripts containing shortlisted HC dORFs was reported. The translational regulation varied across the different cell and tissue types with relative translation increased and decreased (Table 5.9). Out of the 31 transcripts, 13 were suggested to have increased relative translation in all datasets where data is available (Table 5.9). 25 transcripts had results from the healthy kidney tissue, where transcripts with Wu dORFs showed regulatory activity (Figure 3.2), however, only 10 of these transcripts suggested similar activity (Table 5.9). Out of 24 HC dORF transcripts with data for healthy and tumour datasets, 12 transcripts suggested increased relative translation in cancer datasets compared to healthy. Six transcripts suggested reduced relative translation in cancer datasets compared to healthy, matching the trend seen in previous data (Section 3.2), and six transcripts had varying changes in relative translation (Table 5.9). There was no consistent change in relative translation of transcripts in datasets from fed and starved conditions and the control and FKBP10 knockdown cancer cell datasets (Table 5.9). HC dORFs generally did not appear to follow the translational regulation seen previously with Wu dORFs. There were a few transcripts which show potential translational upregulation, and some that do follow the previously described reduced relative translation in cancer datasets in section 3.2. However, most of HC dORF transcripts showed varied and previously unseen changes in tumour tissue with reduced relative translation in healthy tissue which increased in cancer datasets. This could suggest that some HC dORFs lack translational regulation function and may instead have different functions.

Table 5.9: Summary of the mean relative translation data for the highly conserved shortlist of downstream open reading frames (HC dORFs) which use an AUG start codon. *Data from the PRJNA256316, PRJNA880902 and PRJNA532400 datasets was used. Relative translation of transcript calculated as Log2 fold change in expression from RNAseq to ribosome profiling (RP) dataset for each transcript. For each transcript the mean (reported to two decimal places) for the datasets from different treatments/conditions in each Sequence Read Archive (SRA) project are included. The colours of the table entries ranges from the lowest value in the darkest red, becoming more yellow as values approach zero, and then from zero positive values become greener as values approach the highest value in the darkest green.*

SRA Project		**PRJNA256316**		**PRJNA880902**		**PRJNA532400**	
Transcript containing Shortlisted HC dORF	**Gene**	**Relative Translation Healthy Kidney Tissue Mean**	**Relative Translation Kidney Tumour Tissue Mean**	**Relative Translation RKO cells Fed Mean**	**Relative Translation RKO cells Starved Mean**	**Relative Translation A549 Control Mean**	**Relative Translation A549 FKBP10 KD Mean**
ENST00000223145.10	*GLCCI1*	-1.16	0.01	-0.49	-0.54	-1.16	-1.06
ENST00000239938.5	*EGR1*	2.02	1.63	1.26	1.27	1.74	1.34
ENST00000265070.7	*GOLPH3*	1.55	0.86	1.37	1.46	1.85	1.97
ENST00000268154.9	*ZNF710*	-1.85	-1.27	0.56	0.39	-0.74	-0.95
ENST00000275034.5	*PHIP*	-0.87	0.26	1.51	1.59	0.93	0.49
ENST00000310015.12	*SP3*	-0.43	0.41	-3.35	-2.82	-4.12	-2.13
ENST00000332556.5	*LAMP1*	4.35	2.52	1.49	1.49	2.20	2.76
ENST00000352645.5	*ZC3H7B*	-0.76	-0.57	1.31	1.35	-0.82	-0.62
ENST00000374580.10	*BMPR2*	0.01	0.14	2.95	3.00	0.24	0.63
ENST00000378827.5	*BMP2*	-0.71	-1.41	2.81	1.87	2.45	1.36
ENST00000393203.3	*PTGFRN*	2.33	1.41	1.20	1.38	0.22	0.68
ENST00000443185.7	*ZNF516*	-0.88	-0.11	0.86	0.98	1.06	0.33
ENST00000486442.6	*KLHL29*	-3.25	-2.43	0.50	0.39	-3.24	-2.76
ENST00000296452.5	*BSN*	-3.38	-1.76	N/A	-0.70	-1.55	-0.16
ENST00000342232.5	*JPH1*	-0.99	0.50	1.23	1.16	-0.39	N/A

SRA Project		PRJNA256316		PRJNA880902		PRJNA532400	
Transcript containing Shortlisted HC dORF	Gene	Relative Translation Healthy Kidney Tissue Mean	Relative Translation Kidney Tumour Tissue Mean	Relative Translation RKO cells Fed Mean	Relative Translation RKO cells Starved Mean	Relative Translation A549 Control Mean	Relative Translation A549 FKBP10 KD Mean
ENST00000434382.2	VKORC1L1	N/A	-0.72	3.23	3.23	-3.26	-3.48
ENST00000231061.9	SPARC	5.23	1.64	N/A	N/A	1.26	1.56
ENST00000310624.7	NEFH	1.41	1.21	1.05	1.58	N/A	N/A
ENST00000491695.2	AKAP9	1.51	-1.22	0.17	0.90	N/A	N/A
ENST00000566936.5	DOK4	2.62	2.82	N/A	N/A	7.97	6.51
ENST00000260653.5	SIX3	2.37	N/A	N/A	N/A	2.84	4.48
ENST00000369701.8	SORCS3	-1.22	-0.24	1.84	N/A	N/A	N/A
ENST00000370768.7	FUBP1	N/A	-3.44	-3.21	-1.35	N/A	N/A
ENST00000263923.5	KDR	1.84	0.71	N/A	N/A	N/A	N/A
ENST00000278550.12	TENM4	-1.20	-0.43	N/A	N/A	N/A	N/A
ENST00000298282.14	PKNOX2	-0.75	-0.29	N/A	N/A	N/A	N/A
ENST00000394524.7	CAMK2D	N/A	N/A	2.12	2.16	N/A	N/A
ENST00000467011.6	STIM2	N/A	N/A	1.70	2.38	N/A	N/A
ENST00000417717.6	SATB1	N/A	1.69	N/A	N/A	N/A	N/A
ENST00000449047.8	MAPK10	-0.49	N/A	N/A	N/A	N/A	N/A
ENST00000458046.6	SYNE2	N/A	2.36	N/A	N/A	N/A	N/A

5.5 Alignments of RP reads to dORF and 3’ UTRs from RP datasets with various cell types, treatments and disease states

dORF activity and ribosomal association with dORFs and 3’ UTRs has been shown to change in healthy and cancer datasets. 3’ UTR ribosomal presence also changes when cells are exposed to stressful conditions (Hinnebusch, 1988, 2005; Ingolia *et al.*, 2009). Local conditions and disease states, including cancer and inflammation, can affect uORF translation (Young and Wek, 2016; Sendoel *et al.*, 2017; Renz, Valdivia Francia and Sendoel, 2020). This means dORF activity, or translation, could change in different disease states or conditions, potentially influenced by changes in dORF, and 3’ UTR, ribosomal association. Ribosome profiling (RP) datasets from various cell, and tissue types, in varied cellular conditions were used to compare RP read alignments to two dORF shortlists, the dORF-containing 3’ UTRs, and 3’ UTRs of three housekeeping genes (ACTB, TUBB and GAPDH). The dORF shortlists were MSVW dORFs, Wu dORFs with MS validation of translation (Section 3.8), and HC dORFs, highly conserved AUG dORFs (Section 5.4). Investigating RP alignment to regions relative to overall gene alignment allowed comparisons between datasets of differing sizes, where larger datasets could have more alignments. It also means differences are specific to 3’ UTRs or dORFs, rather than being caused by overall gene changes in RP read association. This analysis investigated the enrichment of RP reads in 3’ UTRs or dORFs in different cell states or diseases.

Each shortlist was analysed separately, initially the MSVW dORF shortlist was used. In housekeeping genes, except for RKO cells under fed and starved arginine conditions, HEK293 cells treated with cycloheximide or harringtonine, HeLa cells with *ABCE1* knockdown, and with thapsigargin treatment, RP read alignments to 3’ UTRs was almost zero relative to the gene RP read alignments (Figures 5.7-5.12). In all RP datasets, other than RKO cells under fed and starved arginine conditions, the mean ribosomal association with 3’ UTRs and dORFs, relative to the gene, of MSVW dORFs was increased compared to housekeeping 3’ UTRs, although these differenced were not statistically significant ($P<0.05$) (Figures 5.7-5.12). To establish whether these differences may be biologically relevant further investigation will be needed. It could be that more ribosomes associated with the MSVW dORF-containing 3' UTRs and dORFs than housekeeping 3' UTRs, due to increased translation in these regions. Some individual dORFs and 3' UTRs contained the only RP reads associated with the gene, this was relatively rare, but it could be biologically meaningful, or a consequence of very low RP read alignment to the gene. Generally, around 10% or fewer of the RP reads aligned to the gene aligned to 3' UTRs and dORFs (Figures 5.7-5.12). Across different RP datasets from various cell types, conditions and disease states, there was varied ribosomal association with MSVW dORFs and dORF-containing 3' UTRs relative to the gene (Figures 5.7-5.12). This suggests different conditions and cell types can potentially influence dORF and 3' UTR ribosomal association and possible translation. Although not statistically significant, when comparing healthy kidney and kidney tumour tissue, and oral cavity carcinoma (OCC) tissue and paracancerous tissue, dORF RP read association relative to the gene increased in cancerous tissues, by contrast the RP read association with the remaining 3' UTR was reduced or similar (Figure 5.7). Proliferative BJ-Ras-ER cells had increased 3' UTR and dORF RP read association, relative to the gene, compared to the senescent state, and GC7 treatment in proliferative cells appeared to reduce 3' UTR RP read association (Figure 5.7). Cycloheximide and harringtonine treatment of HEK293 cells also increased RP read association with MSVW dORF 3' UTRs and dORFs, relative to the gene, compared to other RP dataset results (Figure 5.8). The cycloheximide and harringtonine treatments only had one RP dataset, meaning further validation of these results would be needed. Different human tissues and cell types had varied RP read association with 3' UTRs and dORFs relative to the gene, in some tissues, or cells, 3' UTR aligned RP reads generally align in dORFs, whereas others align elsewhere in 3' UTRs (Figure 5.9). Exposure to ionizing radiation appeared to have little impact on 3' UTR and dORF ribosomal association in U251 and U343 cells (Figure 5.10). HeLa cells with *ABCE1* knockdown showed increased RP read association with MSVW dORF 3' UTRs in particular, but also dORFs, relative to the gene, compared to control HeLa cells (Figure 5.11). When compared to DMSO treatment, thapsigargin appeared to increase MSVW dORF shortlist 3' UTR and dORF RP read association relative to the gene (Figure 5.12). These results suggest that ribosomal association with dORFs and 3' UTRs does change with different cell types, disease states, and cellular conditions. 3’ UTR and dORF ribosomal association, although often being increased or decreased together, can be affected separately. Some cancer tissues and cells show increased ribosomal association, and possible translation, in dORFs and 3’ UTRs of the MSVW dORF shortlist. Thapsigargin, cycloheximide, and harringtonine treatments appear to increase ribosomal association with 3’ UTRs, including housekeeping genes, and dORFs relative to the gene and could indicate the effects of cell stress. A similar, and potentially exaggerated effect, also influencing ribosomal association with housekeeping 3’ UTRs, was seen in HeLa cells with *ABCE1* knockdown.

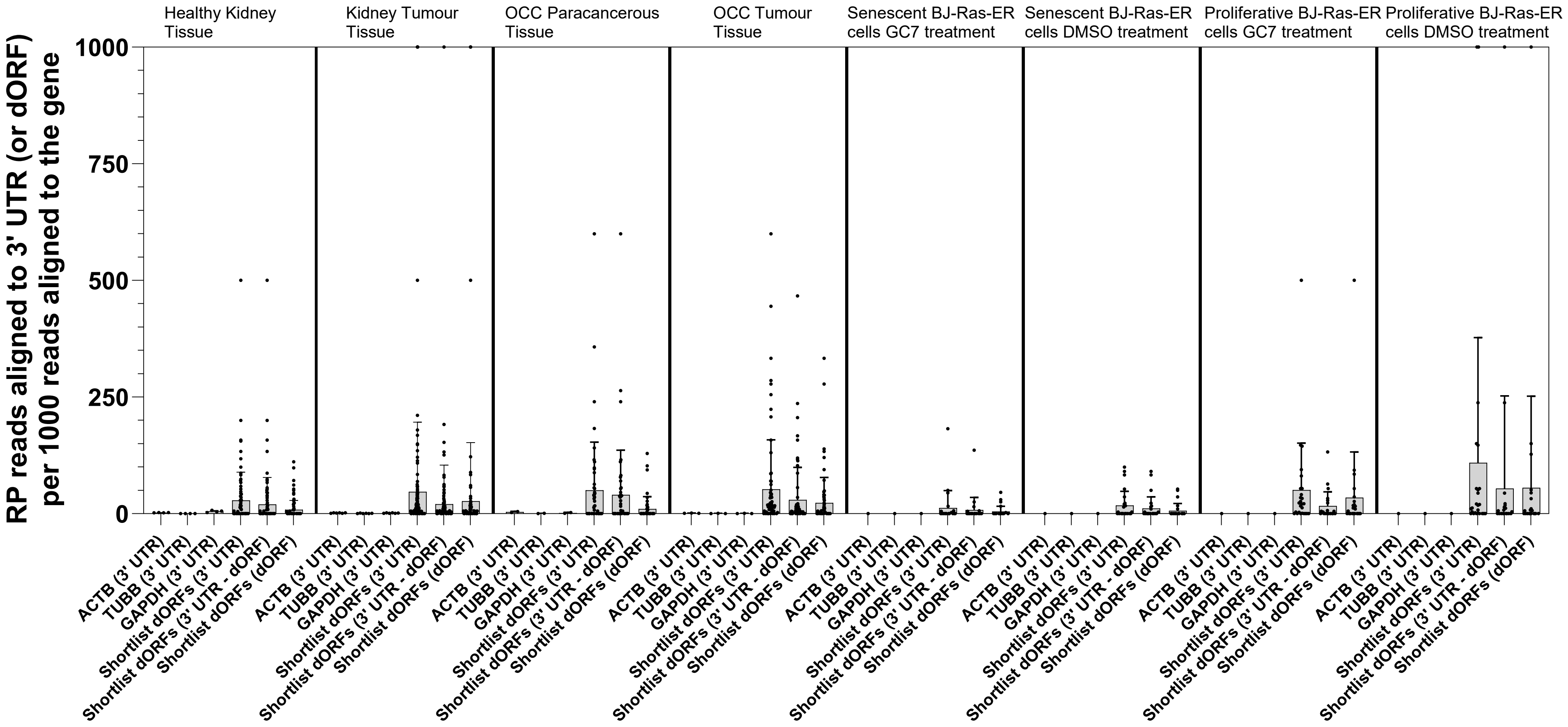

Figure 5.7: ***Ribosome profiling (RP) read alignment to 3' untranslated regions (UTRs), of housekeeping genes and transcripts containing shortlisted downstream open reading frames (dORFs), and shortlisted dORFs relative to the gene in RP datasets from differing cell types and cellular conditions (PRJNA256316, PRJNA768399, and PRJNA982716).*** *The number of RP reads aligned to the 3' UTRs or dORFs are reported per 1000 RP reads aligned to the gene. ACTB, TUBB and GAPDH are the housekeeping genes used. The shortlisted dORFs are the 26 Wu dORFs with mass spectroscopy (MS) validation of the encoded protein, known as MSVW dORFs. RP read alignments are reported for the 3' UTR of the housekeeping genes and the 3' UTR as a whole, 3' UTR without the dORF, and the dORF of the shortlisted dORFs. The different charts are for the cell/tissue type and treatment/condition, if used, included above each chart. Each chart includes all of the datapoints for each dORF or 3' UTR in each dataset, a bar representing the mean with error bars marking the standard deviation. One-Way ANOVA multiple comparisons test with Šidák adjusted P values was used to compare the means. Sequency Read Archive accessions included. From left to right in the charts number of RP datasets = 4, 6, 2, 3, 1, 1, 1, 1.*

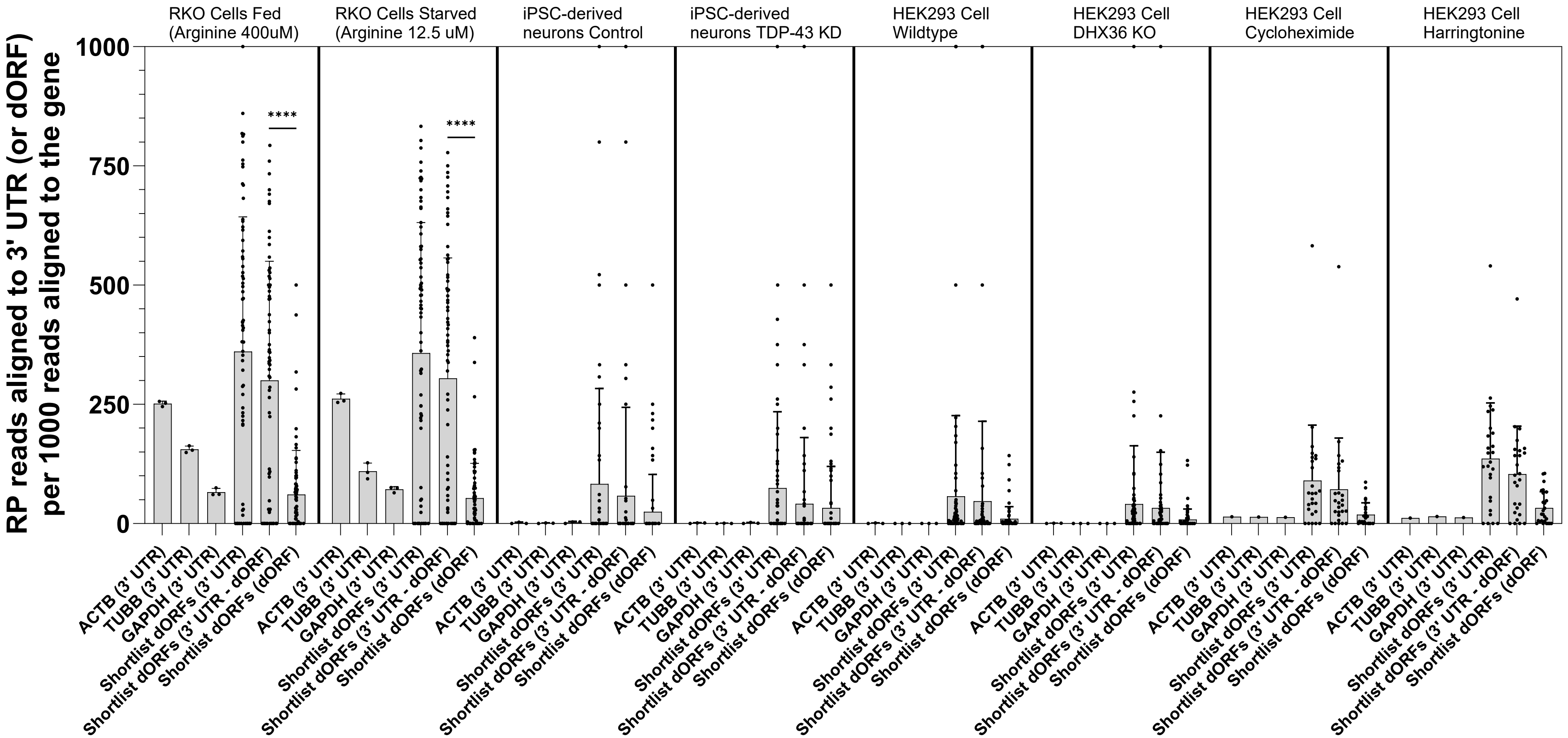

Figure 5.8: Ribosome profiling (RP) read alignment to 3' untranslated regions (UTRs), of housekeeping genes and transcripts containing shortlisted downstream open reading frames (dORFs), and shortlisted dORFs relative to the gene in RP datasets from differing cell types and cellular conditions (PRJNA880902, PRJEB43705, PRJNA415033, and PRJNA369552). *The number of RP reads aligned to the 3' UTRs or dORFs are reported per 1000 RP reads aligned to the gene. ACTB, TUBB and GAPDH are the housekeeping genes used. The shortlisted dORFs are the 26 Wu dORFs with mass spectroscopy (MS) validation of the encoded protein, known as MSVW dORFs. RP read alignments are reported for the 3' UTR of the housekeeping genes and the 3' UTR as a whole, 3' UTR without the dORF, and the dORF of the shortlisted dORFs. The different charts are for the cell/tissue type and treatment/condition, if used, included above each chart. Each chart includes all of the datapoints for each dORF or 3' UTR in each dataset, a bar representing the mean with error bars marking the standard deviation. One-Way ANOVA multiple comparisons test with Šidák adjusted P values was used to compare the means. **** - P<0.0001. Sequency Read Archive accessions included. From left to right in the charts number of RP datasets = 3, 3, 3, 3, 3, 3, 1, 1.*

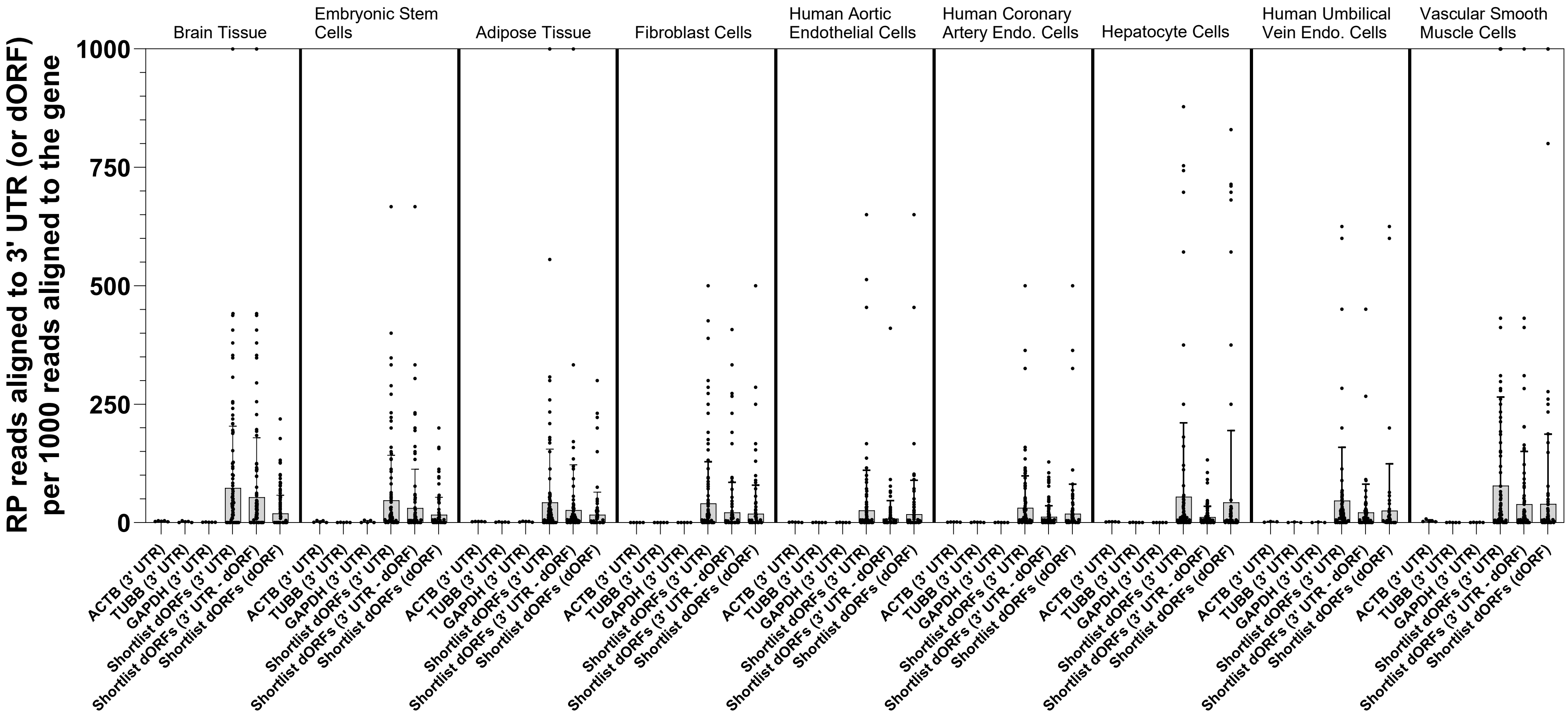

Figure 5.9: Ribosome profiling (RP) read alignment to 3' untranslated regions (UTRs), of housekeeping genes and transcripts containing shortlisted downstream open reading frames (dORFs), and shortlisted dORFs relative to the gene in RP datasets from differing cell types and cellular conditions (PRJNA756018). *The number of RP reads aligned to the 3' UTRs or dORFs are reported per 1000 RP reads aligned to the gene. ACTB, TUBB and GAPDH are the housekeeping genes used. The shortlisted dORFs are the 26 Wu dORFs with mass spectroscopy (MS) validation of the encoded protein, known as MSVW dORFs. RP read alignments are reported for the 3' UTR of the housekeeping genes and the 3' UTR as a whole, 3' UTR without the dORF, and the dORF of the shortlisted dORFs. The different charts are for the cell/tissue type and treatment/condition, if used, included above each chart. Each chart includes all of the datapoints for each dORF or 3' UTR in each dataset, a bar representing the mean with error bars marking the standard deviation. One-Way ANOVA multiple comparisons test with Šídák adjusted P values was used to compare the means. Sequency Read Archive accession included. From left to right in the charts number of RP datasets = 5, 5, 5, 5, 5, 5, 5, 3, 5.*

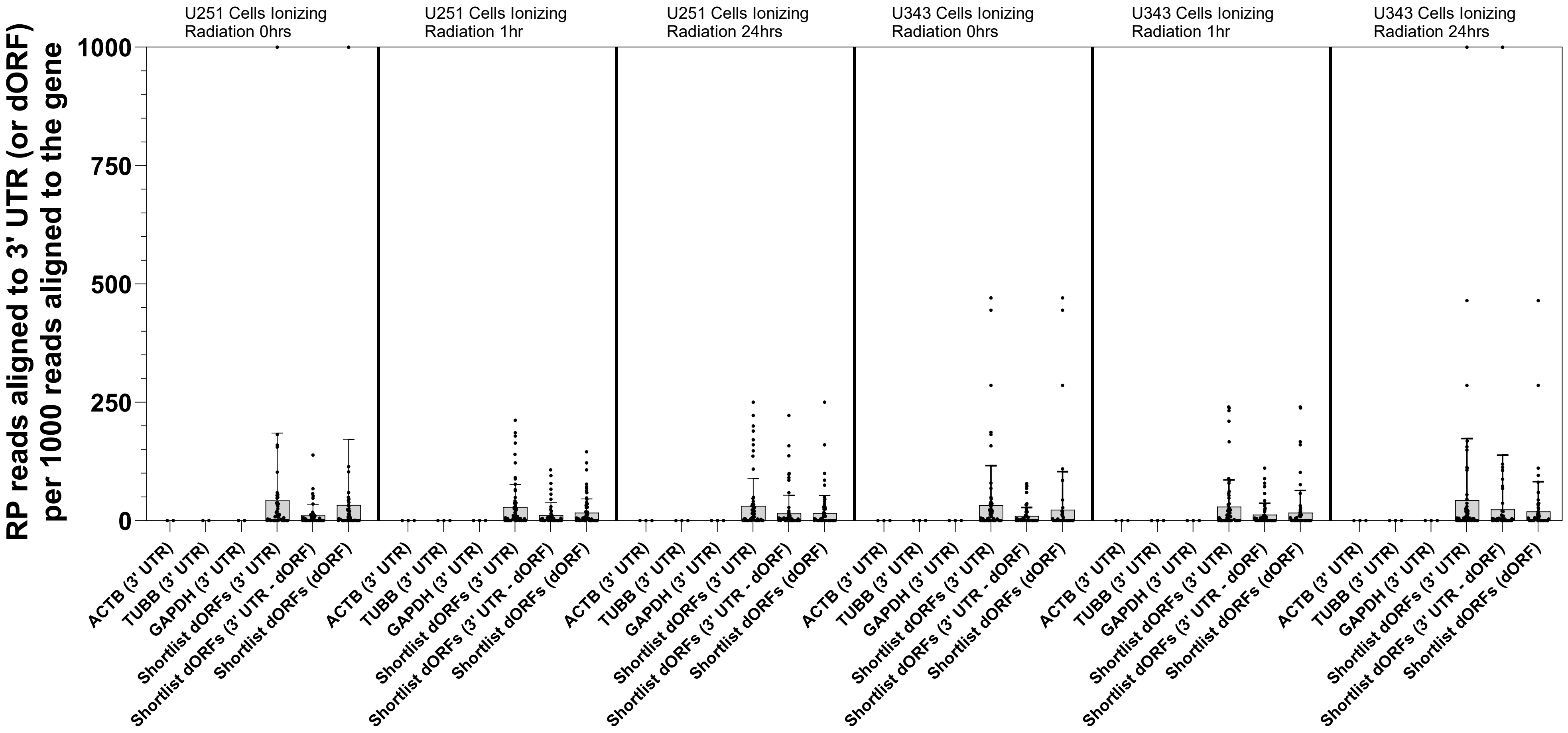

Figure 5.10: Ribosome profiling (RP) read alignment to 3' untranslated regions (UTRs), of housekeeping genes and transcripts containing shortlisted downstream open reading frames (dORFs), and shortlisted dORFs relative to the gene in RP datasets from differing cell types and cellular conditions (PRJNA591767). *The number of RP reads aligned to the 3' UTRs or dORFs are reported per 1000 RP reads aligned to the gene. ACTB, TUBB and GAPDH are the housekeeping genes used. The shortlisted dORFs are the 26 Wu dORFs with mass spectroscopy (MS) validation of the encoded protein, known as MSVW dORFs. RP read alignments are reported for the 3' UTR of the housekeeping genes and the 3' UTR as a whole, 3' UTR without the dORF, and the dORF of the shortlisted dORFs. The different charts are for the cell/tissue type and treatment/condition, if used, included above each chart. Each chart includes all of the datapoints for each dORF or 3' UTR in each dataset, a bar representing the mean with error bars marking the standard deviation. One-Way ANOVA multiple comparisons test with Šídák adjusted P values was used to compare the means. Sequency Read Archive accession included. From left to right in the charts number of RP datasets = 2, 3, 3, 3, 3, 3.*

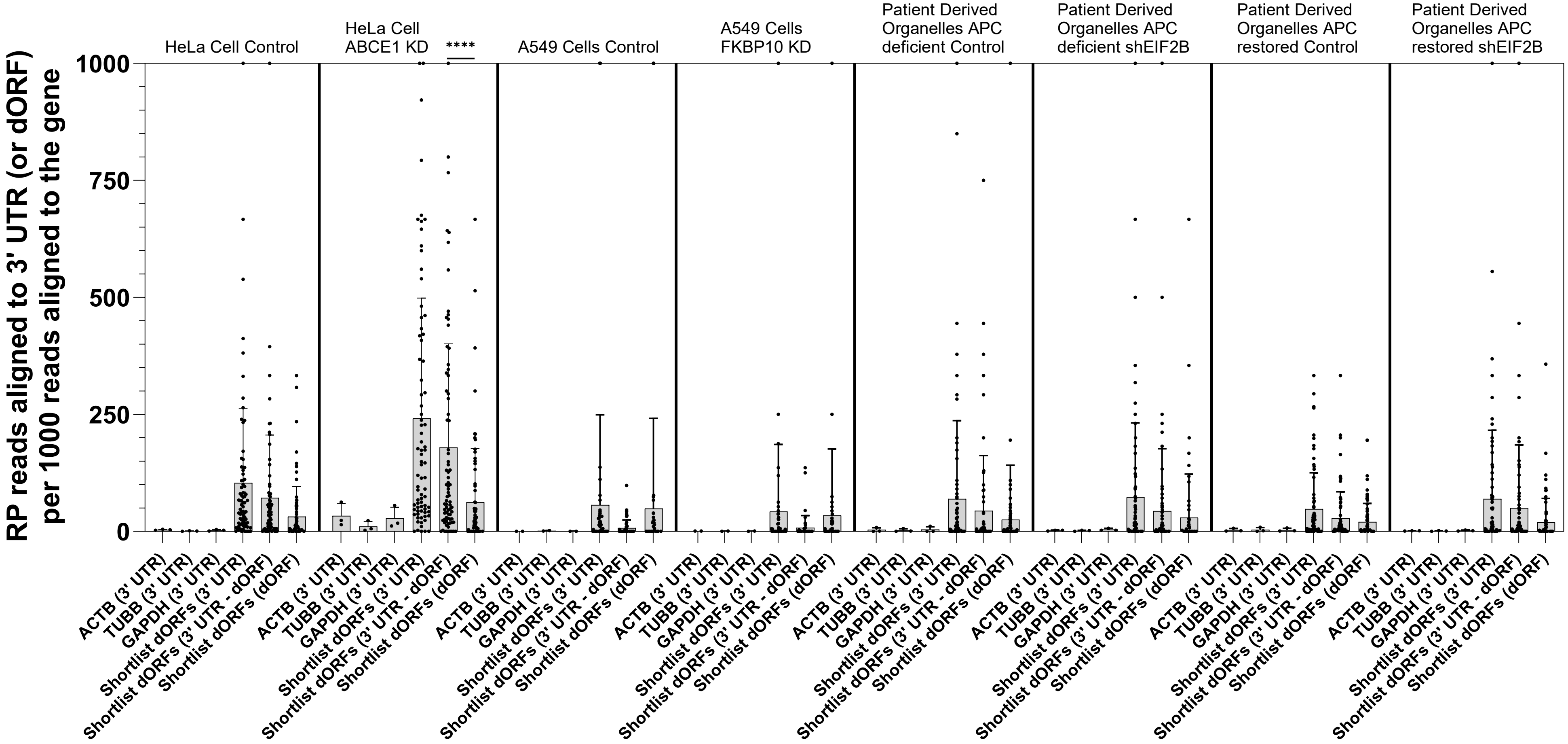

Figure 5.11: Ribosome profiling (RP) read alignment to 3' untranslated regions (UTRs), of housekeeping genes and transcripts containing shortlisted downstream open reading frames (dORFs), and shortlisted dORFs relative to the gene in RP datasets from differing cell types and cellular conditions (PRJNA599943, PRJNA532400, and PRJNA418238). *The number of RP reads aligned to the 3' UTRs or dORFs are reported per 1000 RP reads aligned to the gene. ACTB, TUBB and GAPDH are the housekeeping genes used. The shortlisted dORFs are the 26 Wu dORFs with mass spectroscopy (MS) validation of the encoded protein, known as MSVW dORFs. RP read alignments are reported for the 3' UTR of the housekeeping genes and the 3' UTR as a whole, 3' UTR without the dORF, and the dORF of the shortlisted dORFs. The different charts are for the cell/tissue type and treatment/condition, if used, included above each chart. Each chart includes all of the datapoints for each dORF or 3' UTR in each dataset, a bar representing the mean with error bars marking the standard deviation. One-Way ANOVA multiple comparisons test with Šídák adjusted P values was used to compare the means. **** - p<0.0001. Sequency Read Archive accessions included. From left to right in the charts number of RP datasets = 3, 3, 2, 2, 3, 3, 3, 3.*

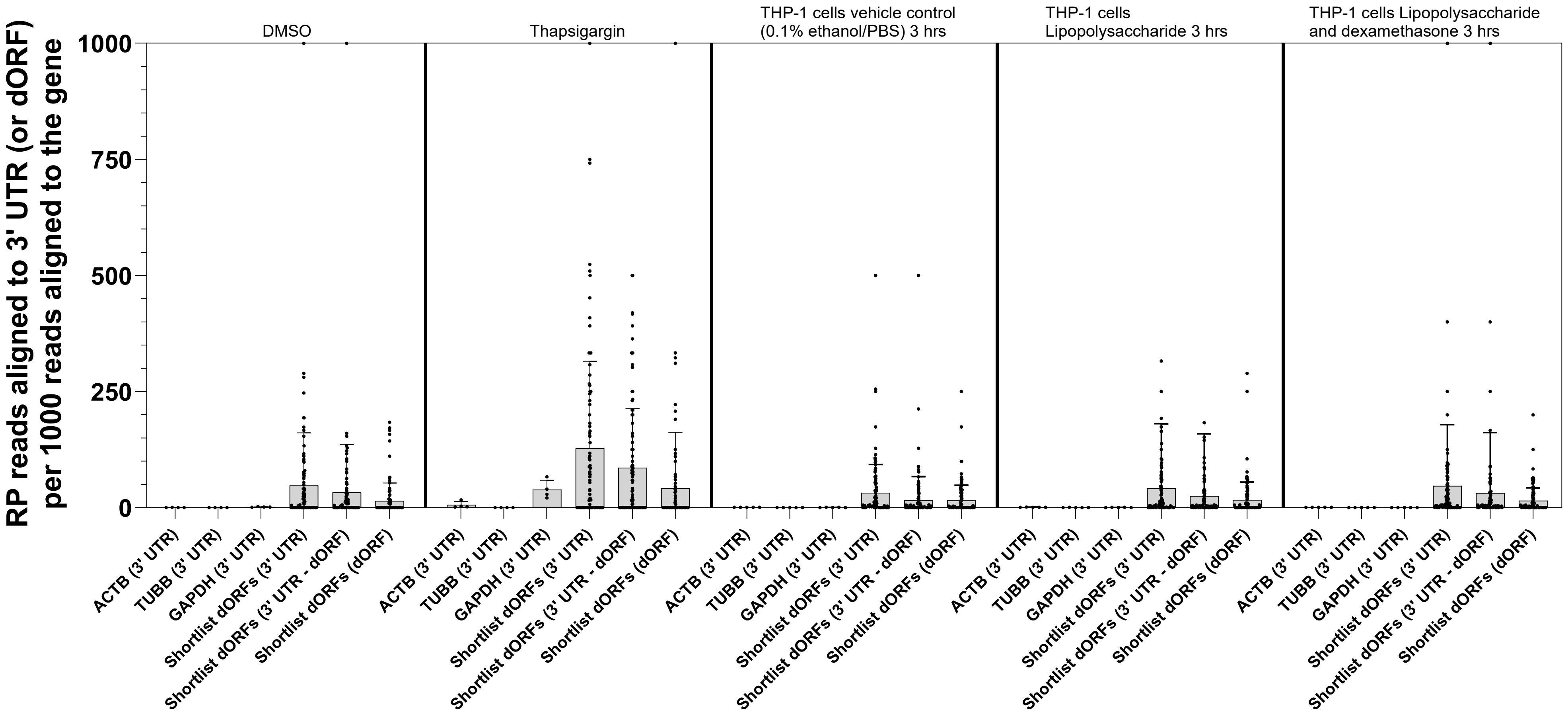

Figure 5.12: Ribosome profiling (RP) read alignment to 3' untranslated regions (UTRs), of housekeeping genes and transcripts containing shortlisted downstream open reading frames (dORFs), and shortlisted dORFs relative to the gene in RP datasets from differing cell types and cellular conditions (PRJNA406823, PRJNA858047). *The number of RP reads aligned to the 3' UTRs or dORFs are reported per 1000 RP reads aligned to the gene. ACTB, TUBB and GAPDH are the housekeeping genes used. The shortlisted dORFs are the 26 Wu dORFs with mass spectroscopy (MS) validation of the encoded protein, known as MSVW dORFs. RP read alignments are reported for the 3' UTR of the housekeeping genes and the 3' UTR as a whole, 3' UTR without the dORF, and the dORF of the shortlisted dORFs. The different charts are for the cell/tissue type and treatment/condition, if used, included above each chart. Each chart includes all of the datapoints for each dORF or 3' UTR in each dataset, a bar representing the mean with error bars marking the standard deviation. One-Way ANOVA multiple comparisons test with Šídák adjusted P values was used to compare the means. Sequency Read Archive accessions included. From left to right in the charts number of RP datasets = 4, 4, 5, 5, 5.*

The HC dORF shortlist, highly conserved dORFs from the conservation analysis, was then used with the same datasets in the same analysis, investigating RP read association with 3’ UTRs and dORFs of the shortlist, with comparison to housekeeping gene 3’ UTRs. Housekeeping gene results were the same as those described previously, they are included for comparison with HC dORFs. In most RP datasets the mean ribosomal association with HC dORF shortlist 3’ UTRs and dORFs, relative to the gene, was increased compared to housekeeping 3’ UTRs (Figures 5.13-5.18). Again, these mean differences were not statistically significant ($P<0.05$), with smaller differences than the MSVW dORF shortlist. Ribosomal association with 3’ UTRs and dORFs of the HC dORF shortlist appeared to be greatly reduced when compared to the MSVW dORF shortlist. There was very little ribosomal association relative to the gene in HC dORFs, even compared to housekeeping 3’ UTRs (Figures 5.13 – 5.18), potentially suggesting there was little translation. There was variation in RP read alignment to 3’ UTRs, relative to the gene, of the HC dORFs across the different RP datasets (Figures 5.7-5.12). There was some slight variation for dORFs, but the general low RP read association meant these differences were small. In many different cell, or treatment, types the RP read association with 3’ UTRs, excluding the dORFs, was greater than that of dORFs, and the mean difference was often statistically significant ($P<0.05$) (Figures 5.13-5.18). Although not as pronounced as the MSVW dORF shortlist, different conditions and cell types can still potentially influence dORF and 3’ UTR ribosomal association in the HC dORF shortlist. In cancer tissues there was little change, or a possible slight reduction, in RP read association, relative to the gene, in dORFs and 3’ UTRs of the HC dORF shortlist when comparing healthy kidney and kidney tumour tissue, and OCC tissue and paracancerous tissue (Figure 5.13). In contrast to MSVW dORF shortlist, DMSO treated senescent BJ-Ras-ER cells had greater RP read association with 3’ UTRs of the HC dORF shortlist compared to the other BJ-Ras-ER cells, however there is low RP alignment to dORFs and 3’ UTRs (Figure 5.13). Cycloheximide and harringtonine treatment of HEK293 cells also increased RP read association with 3’ UTRs of the HC dORF shortlist, relative to the gene, however, there was little change in dORF RP read association, when compared to other RP dataset results (Figure 5.14). In different human tissues and cell types, RP read association to 3’ UTRs and dORFs relative to the gene, varied slightly, however there was little RP read alignment (Figure 5.15). Similar to MSVW dORFs, ionizing radiation exposure had little effect on 3’ UTR and dORF ribosomal association in U251 and U343 cells for HC dORFs (Figure 5.16). Increased RP read association with 3’ UTRs in particular, but also dORFs, relative to gene, for the HC dORF shortlist was seen when comparing control and *ABCE1* knockdown HeLa cells (Figure 5.17). Thapsigargin treatment compared to DMSO also appeared to increase 3’ UTR RP read association relative to the gene (Figure 5.18). Although small changes, lipopolysaccharide treatment of THP-1 cells, compared to the vehicle, increased RP read association with HC dORF shortlist 3’ UTRs and dORFs, relative to the gene, an effect which reduced when dexamethasone was added (Figure 5.18). Increases in HC dORFs mean RP read association was often due to increases in a small number of dORFs, rather than across the shortlist. There were some similarities to the MSVW dORF shortlist, different cell types, disease states, and cellular conditions did change ribosomal association with dORFs and 3’ UTRs. Treatment with thapsigargin, cycloheximide, and harringtonine, similar to the MSVW dORF shortlist, seem to increase ribosomal association with the HC dORF shortlist 3’ UTRs and dORFs, and a similar effect was seen with *ABCE1* knockdown. The differences seen when comparing healthy and cancer tissue in the MSVW dORF shortlist were not present for the HC dORF shortlist. The influence of lipopolysaccharide and dexamethasone treatment, suggest inflammation may affect ribosomal association with the HC dORF shortlist 3’ UTRs and dORFs. The results for the HC dORFs, particularly when contrasted with the MSVW dORF shortlist, show how different they appear. The general lack of ribosomal association with, and potential translation of, HC dORFs alongside these differences to MSVW dORFs, highlights that HC dORFs and Wu dORFs may not share the same regulatory effect.

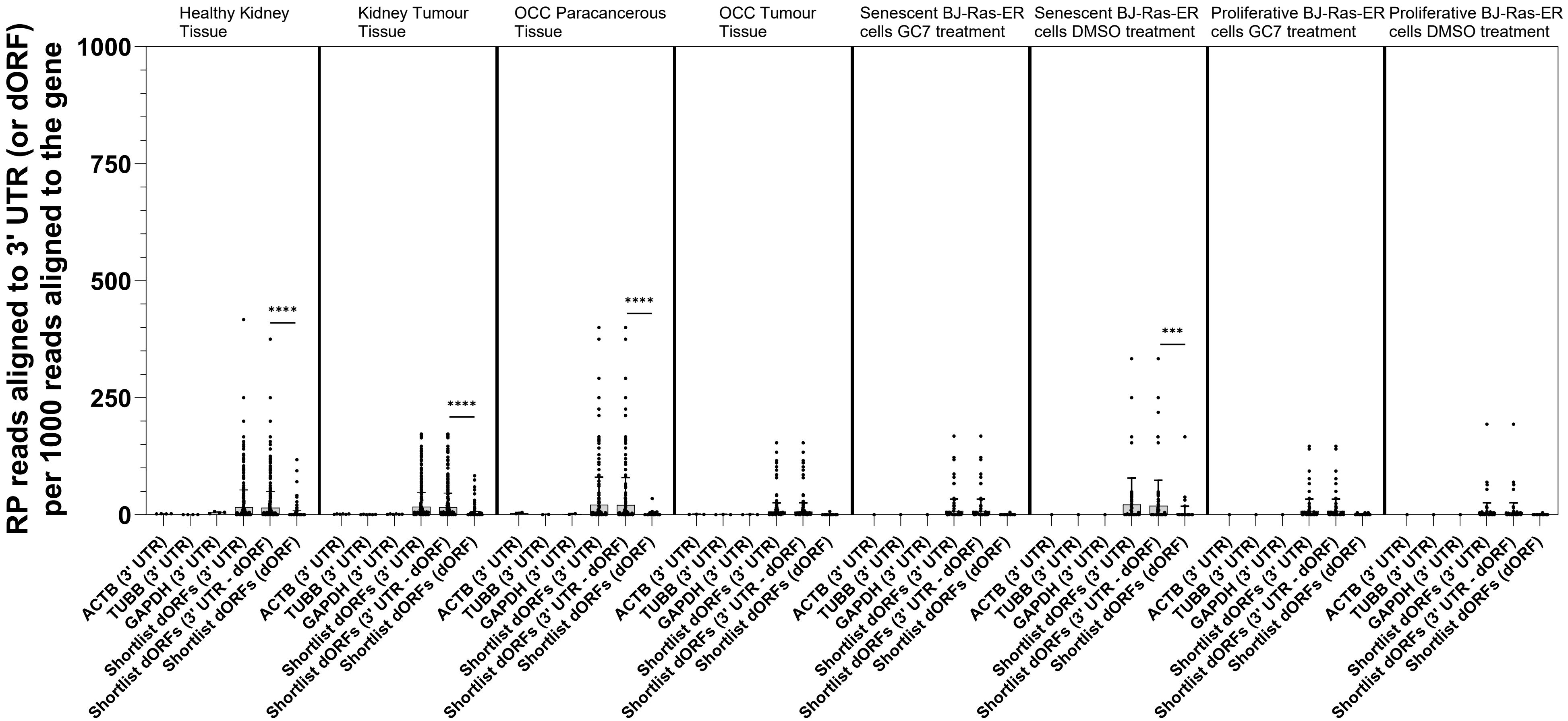

Figure 5.13: Ribosome profiling (RP) read alignment to 3' untranslated regions (UTRs), of housekeeping genes and transcripts containing conservation analysis shortlisted downstream open reading frames (dORFs), and shortlisted dORFs relative to the to the gene in RP datasets from differing cell types and cellular conditions (PRJNA256316, PRJNA768399, and PRJNA982716). *The number of RP reads aligned to the 3' UTRs or dORFs are reported per 1000 RP reads aligned to the gene. ACTB, TUBB and GAPDH are the housekeeping genes used. The shortlisted dORFs are the 122 dORFs identified in the conservation analysis, known as HC dORFs. RP read alignments are reported for the 3' UTR of the housekeeping genes and the 3' UTR as a whole, 3' UTR without the dORF, and the dORF of the shortlisted dORFs. The different charts are for the cell/tissue type and treatment/condition, if used, included above each chart. Each chart includes all of the datapoints for each dORF or 3' UTR in each dataset, a bar representing the mean with error bars marking the standard deviation. One-Way ANOVA multiple comparisons test with Šídák adjusted P values was used to compare the means. **** - P<0.0001, *** - P<0.0005. Sequency Read Archive accessions included. From left to right in the charts number of RP datasets = 4, 6, 2, 3, 1, 1, 1, 1.*

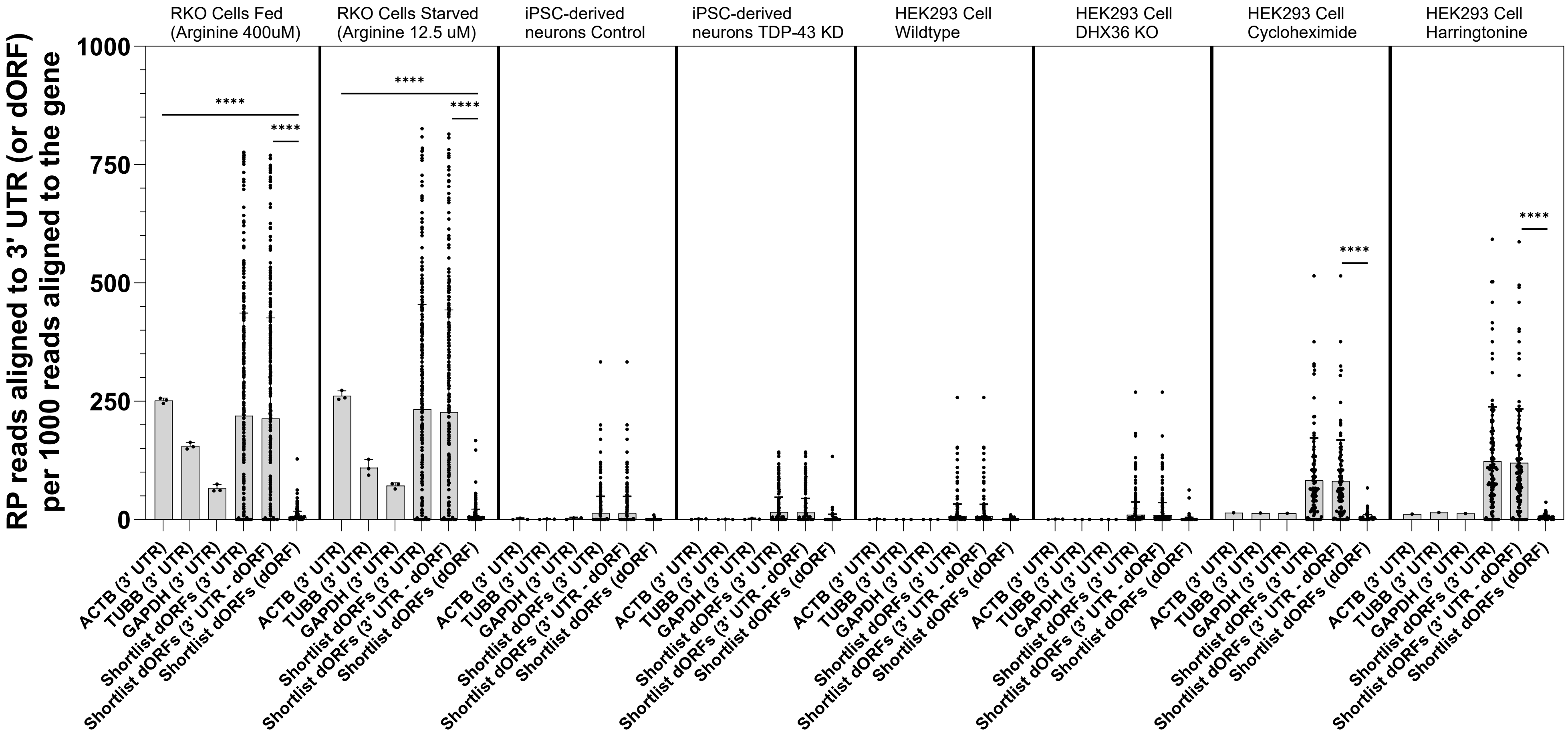

Figure 5.14: Ribosome profiling (RP) read alignment to 3' untranslated regions (UTRs), of housekeeping genes and transcripts containing conservation analysis shortlisted downstream open reading frames (dORFs), and shortlisted dORFs relative to the gene in RP datasets from differing cell types and cellular conditions (PRJNA880902, PRJEB43705, PRJNA415033, and PRJNA369552). *The number of RP reads aligned to the 3' UTRs or dORFs are reported per 1000 RP reads aligned to the gene. ACTB, TUBB and GAPDH are the housekeeping genes used. The shortlisted dORFs are the 122 dORFs identified in the conservation analysis, known as HC dORFs. RP read alignments are reported for the 3' UTR of the housekeeping genes and the 3' UTR as a whole, 3' UTR without the dORF, and the dORF of the shortlisted dORFs. The different charts are for the cell/tissue type and treatment/condition, if used, included above each chart. Each chart includes all of the datapoints for each dORF or 3' UTR in each dataset, a bar representing the mean with error bars marking the standard deviation. One-Way ANOVA multiple comparisons test with Šidák adjusted P values was used to compare the means. **** - P<0.0001. Sequency Read Archive accessions included. From left to right in the charts n = 3, 3, 3, 3, 3, 3, 1, 1.*

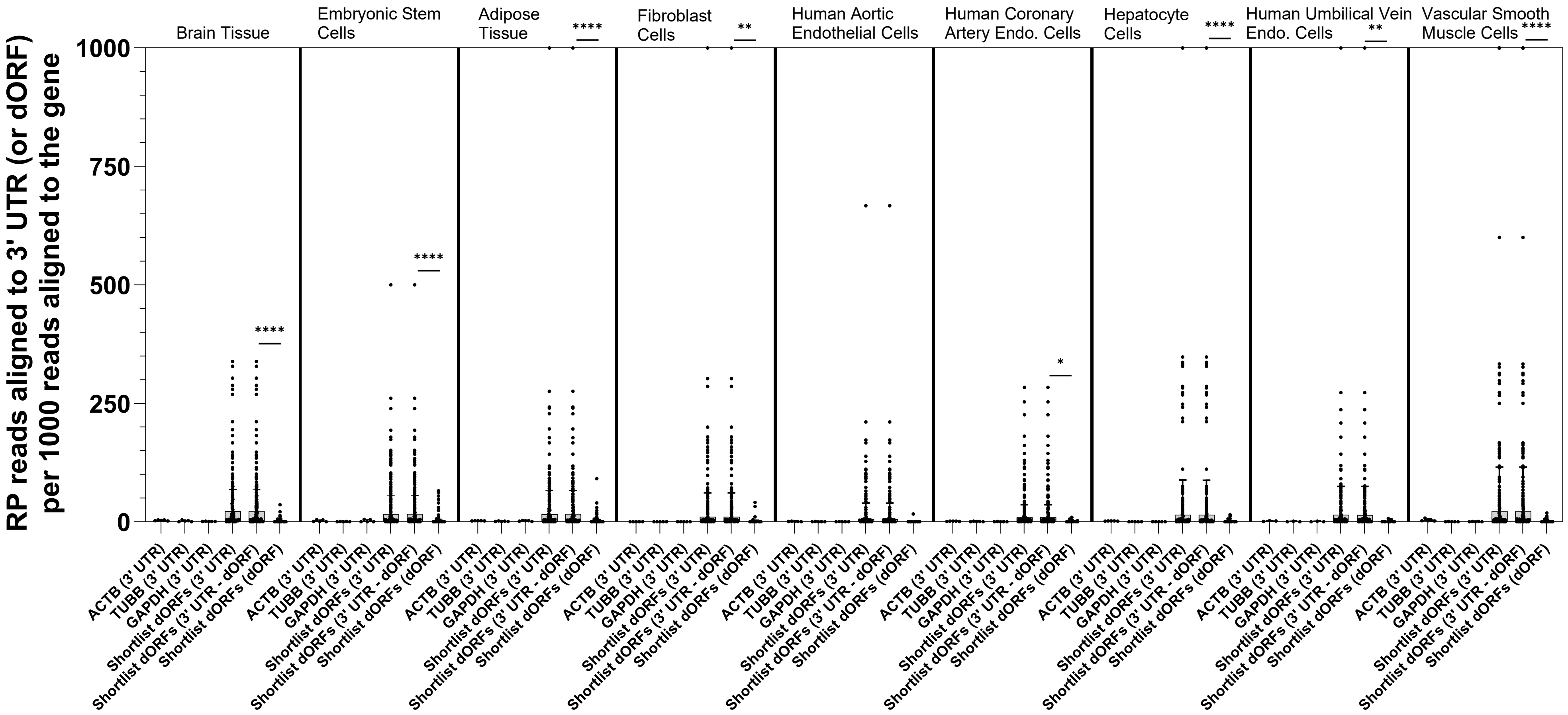

Figure 5.15: Ribosome profiling (RP) read alignment to 3' untranslated regions (UTRs), of housekeeping genes and transcripts containing conservation analysis shortlisted downstream open reading frames (dORFs), and shortlisted dORFs relative to the gene in RP datasets from differing cell types and cellular conditions (PRJNA756018).** The number of RP reads aligned to the 3' UTRs or dORFs are reported per 1000 RP reads aligned to the gene. ACTB, TUBB and GAPDH are the housekeeping genes used. The shortlisted dORFs are the 122 dORFs identified in the conservation analysis, known as HC dORFs. RP read alignments are reported for the 3' UTR of the housekeeping genes and the 3' UTR as a whole, 3' UTR without the dORF, and the dORF of the shortlisted dORFs. The different charts are for the cell/tissue type and treatment/condition, if used, included above each chart. Each chart includes all of the datapoints for each dORF or 3' UTR in each dataset, a bar representing the mean with error bars marking the standard deviation. One-Way ANOVA multiple comparisons test with Šídák adjusted P values was used to compare the means. * - P<0.0001, ** - P<0.005, * - P<0.05. Sequency Read Archive accession included. From left to right in the charts number of RP datasets = 5, 5, 5, 5, 5, 5, 5, 3, 5.*

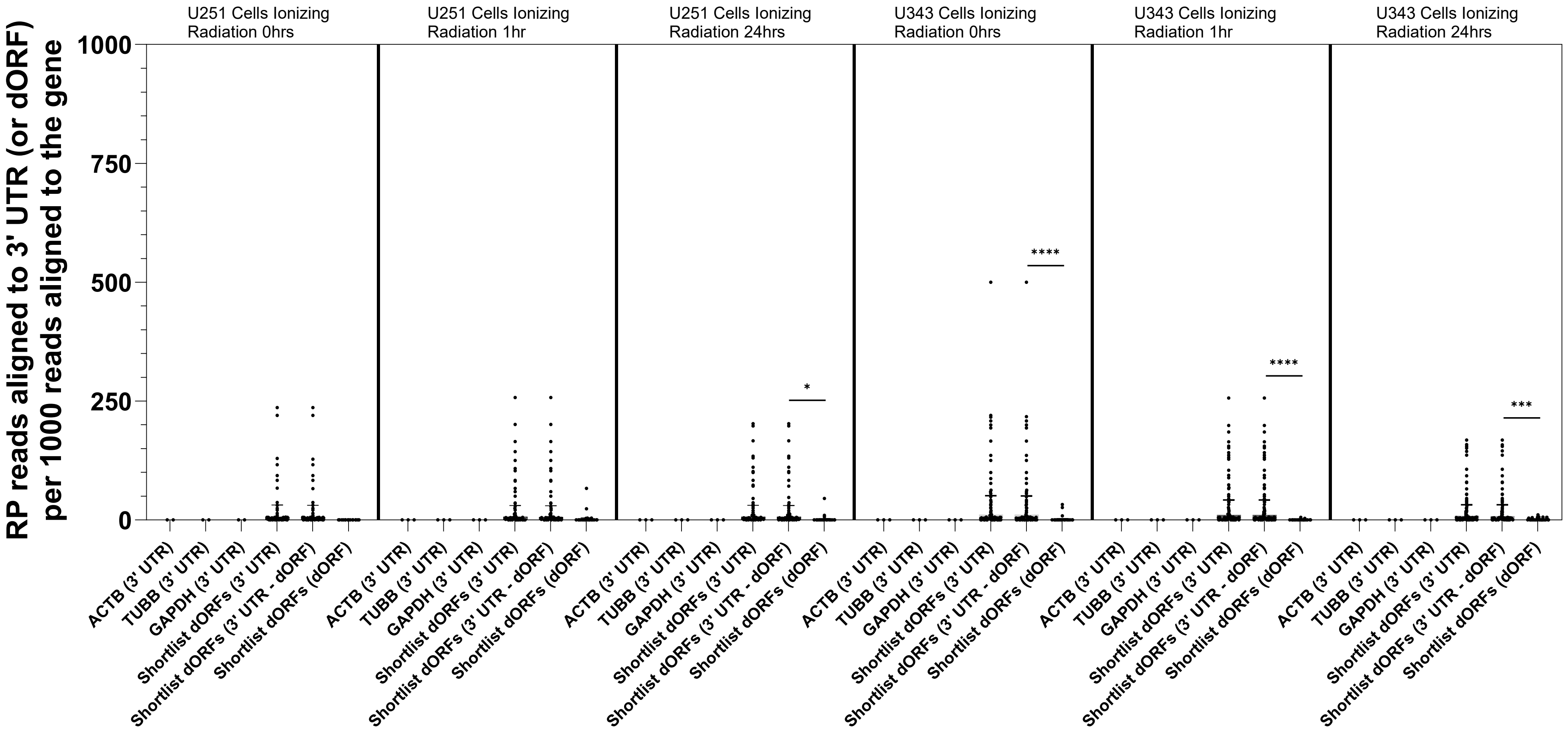

Figure 5.16: Ribosome profiling (RP) read alignment to 3' untranslated regions (UTRs), of housekeeping genes and transcripts containing conservation analysis shortlisted downstream open reading frames (dORFs), and shortlisted dORFs relative to the gene in RP datasets from differing cell types and cellular conditions (PRJNA591767). *The number of RP reads aligned to the 3' UTRs or dORFs are reported per 1000 RP reads aligned to the gene. ACTB, TUBB and GAPDH are the housekeeping genes used. The shortlisted dORFs are the 122 dORFs identified in the conservation analysis, known as HC dORFs. RP read alignments are reported for the 3' UTR of the housekeeping genes and the 3' UTR as a whole, 3' UTR without the dORF, and the dORF of the shortlisted dORFs. The different charts are for the cell/tissue type and treatment/condition, if used, included above each chart. Each chart includes all of the datapoints for each dORF or 3' UTR in each dataset, a bar representing the mean with error bars marking the standard deviation. One-Way ANOVA multiple comparisons test with Šídák adjusted P values was used to compare the means. **** - P<0.0001, *** - P<0.0005, * - P<0.05. Sequency Read Archive accession included. From left to right in the charts number of RP datasets = 2, 3, 3, 3, 3, 3.*

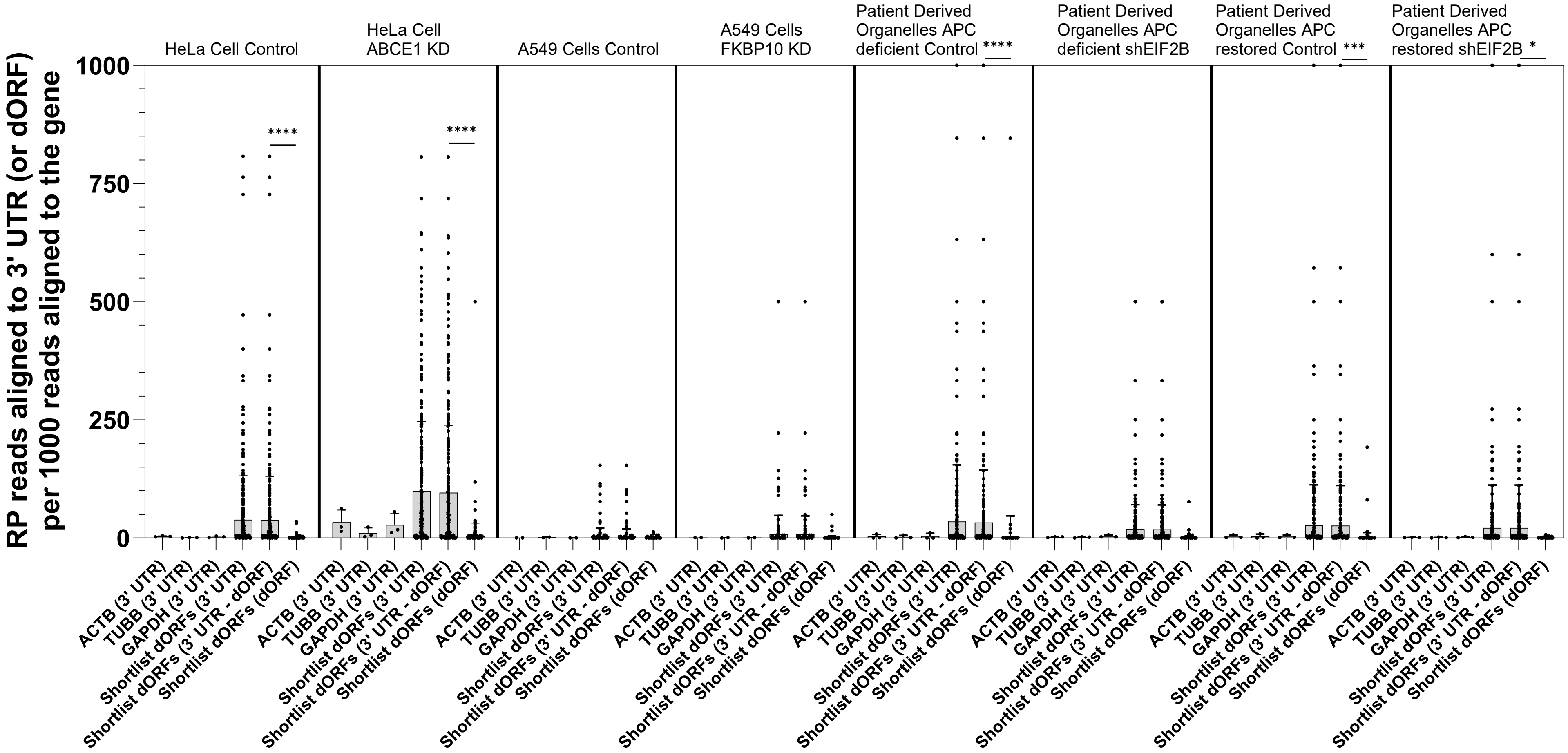

Figure 5.17: Ribosome profiling (RP) read alignment to 3' untranslated regions (UTRs), of housekeeping genes and transcripts containing conservation analysis shortlisted downstream open reading frames (dORFs), and shortlisted dORFs relative to the gene in RP datasets from differing cell types and cellular conditions (PRJNA599943, PRJNA532400, and PRJNA418238). *The number of RP reads aligned to the 3' UTRs or dORFs are reported per 1000 RP reads aligned to the gene. ACTB, TUBB and GAPDH are the housekeeping genes used. The shortlisted dORFs are the 122 dORFs identified in the conservation analysis, known as HC dORFs. RP read alignments are reported for the 3' UTR of the housekeeping genes and the 3' UTR as a whole, 3' UTR without the dORF, and the dORF of the shortlisted dORFs. The different charts are for the cell/tissue type and treatment/condition, if used, included above each chart. Each chart includes all of the datapoints for each dORF or 3' UTR in each dataset, a bar representing the mean with error bars marking the standard deviation. One-Way ANOVA multiple comparisons test with Šídák adjusted P values was used to compare the means. **** - p<0.0001, *** - P<0.0005, * - P<0.05. Sequency Read Archive accessions included. From left to right in the charts number of RP datasets = 3, 3, 2, 2, 3, 3, 3, 3.*

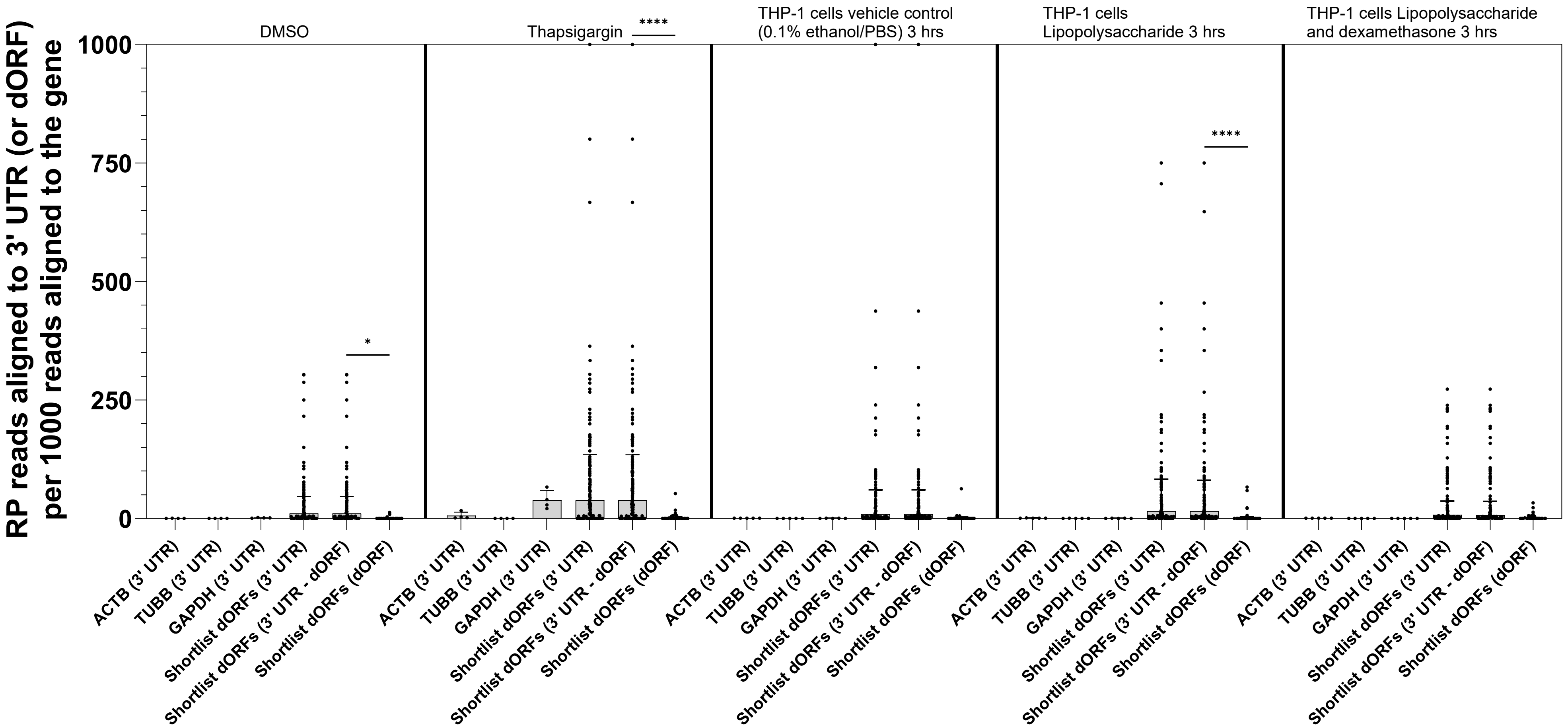

Figure 5.18: Ribosome profiling (RP) read alignment to 3' untranslated regions (UTRs), of housekeeping genes and transcripts containing conservation analysis shortlisted downstream open reading frames (dORFs), and shortlisted dORFs relative to the gene in RP datasets from differing cell types and cellular conditions (PRJNA406823, PRJNA858047). *The number of RP reads aligned to the 3' UTRs or dORFs are reported per 1000 RP reads aligned to the gene. ACTB, TUBB and GAPDH are the housekeeping genes used. The shortlisted dORFs are the 122 dORFs identified in the conservation analysis, known as HC dORFs. RP read alignments are reported for the 3' UTR of the housekeeping genes and the 3' UTR as a whole, 3' UTR without the dORF, and the dORF of the shortlisted dORFs. The different charts are for the cell/tissue type and treatment/condition, if used, included above each chart. Each chart includes all of the datapoints for each dORF or 3' UTR in each dataset, a bar representing the mean with error bars marking the standard deviation. One-Way ANOVA multiple comparisons test with Šidák adjusted P values was used to compare the means. **** - P<0.0001, * - P<0.05. Sequency Read Archive accessions included. From left to right in the charts number of RP datasets = 4, 4, 5, 5, 5.*

5.6 Discussion

Investigating the conservation of dORFs and the 3' UTR sequences containing them across a range of species, required complex analysis and the development of custom Python scripts. One challenge was 3' UTR sequence availability, especially for species more distantly related to humans. There were some species where no 3' UTRs could be collected (Table 5.1). The differing size of genomes across species means that, especially for invertebrates, yeasts, and plants, there are not homologous genes for many human genes, meaning conservation of many human dORFs cannot be compared in these species. The sequence similarity analysis indicated conservation. The EMBOSS WATER tool (version – 6.6.0.0) locally aligned 3' UTRs and dORFs with AUG start codons (AUG dORFs) between species, using the Smith-Waterman local alignment algorithm (Smith and Waterman, 1981; Peter Rice, Longden and Bleasby, 2000). Although developed in the 1980s, this algorithm is still regarded as one of the best methods for local alignments between sequences (Wise, 2003; Rucci *et al.*, 2018; Xia *et al.*, 2022). This analysis does not report alignment results for every human and homolog dORF compared, alignment scores must exceed zero to be reported (Smith and Waterman, 1981; Peter Rice, Longden and Bleasby, 2000). Additionally, instead of including partial short alignments, only those where the alignment length was at least as long as the shorter of the two sequences compared were included. The effect of this alignment length was investigated to ensure it was not biasing the results. Table 8.5 (Appendix 2) shows that relaxing the alignment length to 90% or 80% of the shorter dORF being aligned had very little impact on the similarity results for each species. Although, more comparisons were made when the stringency of alignment length filtering was reduced (Table 8.5). The human AUG dORF was compared against all AUG dORFs in the homologous 3' UTRs in each species, meaning each human dORF was potentially compared with multiple homolog dORFs. The reported similarity for each human dORF in each species was the alignment with the greatest similarity percentage, suggesting the most conserved dORF.

RP read alignment to dORFs and 3' UTRs across various RP datasets with different cell types, disease states, and conditions was investigated to explore whether ribosomal association with, and potential translation of, these regions changes across these different datasets. All bioinformatic tools used in this analysis are widely used in preprocessing, read alignments, and alignment processing (Weeks and Luecke, 2017; Jiang *et al.*, 2024). The selection of tools, used in methods described in section 2.10, was driven by their regular use in bioinformatics, familiarity to me, and accessibility and ease of use within Python scripts on a UNIX network. The tool used to quantify RP read alignments, reports alignments if one RP read nucleotide overlaps a region. In RP reads the translated codon is towards the centre, the precise location is often estimated (Ahmed *et al.*, 2019), meaning to have confidence that the aligned RP read suggests translation the central portion of the RP read should align. This was done as described in section 2.10. Selection of RP datasets, from different cell types, conditions, and disease states, was based on availability of human RP datasets in the SRA database (Leinonen, Sugawara and Shumway, 2011), Some datasets were removed due to poor human genome alignment, potentially due to issues with dataset production or differences in methodologies or read processing, making them unsuitable with the analysis without adjustment. Some results are only based on one or two biological repeats because the number of repeats for each treatment or cell type was solely based on availability. This was important to investigate the broadest range of conditions and cell types, but further validation would be needed with more datasets in the future. Although 3' UTR or dORFs ribosomal presence could indicate translation (Ingolia *et al.*, 2009; Guydosh and Green, 2014; Ji *et al.*, 2015; Young *et al.*, 2015; Hsu *et al.*, 2016), ribosomes could also associate with these dORFs if they form part of the CDS in alternative transcripts. When investigated, only one dORFs from either shortlist had a small proportion which aligned to the CDS of an alternative transcript. This method used ribosomal alignment relative to the gene, meaning the focus was on changing proportions of gene aligned RP reads in dORFs and 3' UTRs. The raw data generally suggested more RP reads were aligned to housekeeping genes when compared to genes with the two dORF shortlists, made up of HC dORFs, highly conserved dORFs from the conservation analysis, and MSVW dORFs, Wu dORFs (dORFs presented by Wu *et al.* (2020b)) with validation from MS.

The conservation analysis showed AUG dORFs were abundant across a range of species, supporting previous findings and suggestion that sORF presence can be underestimated (Slavoff *et al.*, 2013; Bazzini *et al.*, 2014; Couso and Patraquim, 2017; Lu *et al.*, 2019; van Heesch *et al.*, 2019; Chen *et al.*, 2020; Ouspenskaia *et al.*, 2020; Ruiz Cuevas *et al.*, 2021). AUG dORFs were less common in shorter 3' UTRs (only a few hundred nucleotides), potentially due to dORF and 3' UTR lengths being similar. The huge number of AUG dORFs makes it more likely that some may be randomly generated, especially considering dORFs also use non-AUG start codons (Guttman *et al.*, 2013; Couso and Patraquim, 2017). Conservation of AUG dORFs was increased compared to their 3' UTRs in all species considered (Figure 5.4 and Table 5.2). AUG dORF and 3' UTR conservation reduced as species became more distantly related. In *P.troglodytes* and *M.mulatta*, closely related species, human 3' UTRs were often highly conserved, supporting suggestion that 3' UTRs are the most conserved UTR (Siepel *et al.*, 2005; Litterman *et al.*, 2019). In some mammal species binomial distribution of similarity percentages could suggest differing conservation of groups of dORFs, which could relate to dORF group functions, or an influence from increased 3' UTR conservation in some genes, such as those associated with development and cell survival (Duret, Dorkeld and Gautier, 1993; Shabalina *et al.*, 2003; Litterman *et al.*, 2019). Reduced conservation in more distantly related species could be due to reduced 3' UTR availability, meaning there were no homolog 3' UTRs available for similarity analysis with highly conserved human AUG dORFs in mammalian species. The potential conservation of such a large number of AUG dORFs relative to their 3' UTRs, could suggest they have regulatory function (Bashirullah, Cooperstock and Lipshitz, 1998; Conne, Stutz and Vassalli, 2000; Grzybowska, Wilczynska and Siedlecki, 2001; Pesole *et al.*, 2001; Mignone *et al.*, 2002; Shabalina *et al.*, 2004; Martinez *et al.*, 2019), potentially the translational regulation described for Wu dORFs (Wu *et al.*, 2020b). Unlike uORFs, where their presence rather than sequence is often conserved (Chew, Pauli and Schier, 2016; Johnstone, Bazzini and Giraldez, 2016; Dumesic *et al.*, 2019), the sequence similarity appears to be conserved for some dORFs.

Control sequences were used to investigate conservation of dORF-like regions around AUG dORFs, and to explore whether conservation short human 3' UTR sequences in other species could explain the findings. The short human 3' UTR control sequences had reduced conservation compared to AUG dORFs in all species, and full 3' UTRs, in some species (Figure 5.6 and Table 5.4). The conservation of dORFs could not be explained by chance conservation of short human 3' UTR regions in other species, suggesting there is a biological reason for dORF conservation. Using the control sequences surrounding dORFs, AUG dORF were found to be more conserved than the surrounding 3’ UTR (Tables 5.5 and 5.6). These control sequences were less conserved than the full 3’ UTRs in mammalian species, whereas in other species they were slightly more conserved than 3’ UTRs. Most AUG dORFs were not part of larger conserved 3’ UTR regions. The increased conservation, compared to full 3’ UTRs, of 3’ UTR regions around AUG dORFs in some species could suggest some of these regions may be biologically important, perhaps for ribosomal recruitment if AUG dORFs are translated (Wu *et al.*, 2020b). Reduced conservation compared to AUG dORFs could mean these regions are less biologically important than dORFs, or they may only be associated with some AUG dORFs.

The highly conserved AUG dORF were combined into the HC dORF shortlist. The data from gene ontology tools highlighted some issues with ontology analysis of short gene lists. The GOrilla tool (Eden *et al.*, 2009) reported there was no enriched gene function, whereas the DAVID functional annotation tool (Sherman *et al.*, 2022), and g:Profiler g:GOSt functional profiling tool (Raudvere *et al.*, 2019) did. The DAVID functional annotation tool reported clusters of functions for the genes (Sherman *et al.*, 2022). However, only the first four clusters were reported, remaining clusters lacked statistical significance or reported enrichment for very few, one or two, genes. Functional annotation tools can report very similar functions separately, making it appear that more functions are enriched. There was potentially an enrichment of genes associated with transcriptional, and potentially translational, regulation. The proposed translational regulation by Wu dORFs (Wu *et al.*, 2020b) in these genes, associated with transcriptional or translational regulation, could mean dORF regulation has downstream effects on other transcriptionally or translationally regulated genes. If dORFs are regulators in these pathways, with effects on a wider range of genes, dORFs would be an interesting treatment target.

However, the HC dORF shortlist do not generally share the translational regulation seen with Wu dORFs (Wu *et al.*, 2020b) (Table 5.9). A small group of transcripts with HC dORFs potentially had similar regulatory activity, and reduced activity in cancer datasets, to Wu dORFs. Some HC dORFs without regulatory activity may have different functions. The high sequence similarity across species could mean HC dORF sequences are biologically important, which could indicate a function through the encoded protein, such as the dORF encoded protein in *ABCB5* with suggested immunogenic activity (Chong *et al.*, 2020). Proteins encoded in sORFs can be conserved across species (Martinez *et al.*, 2019), and can regulate a range of biological processes (Couso and Patraquim, 2017; Renz, Valdivia Francia and Sendoel, 2020), including different cancers (Sendoel *et al.*, 2017; Sriram, Bohlen and Teleman, 2018; Lei *et al.*, 2020; Wu *et al.*, 2020a; Bakhti and Latifi-Navid, 2021; He *et al.*, 2021; Zhou *et al.*, 2021; Zhou *et al.*, 2022b). sORFs and their encoded proteins can have different, or even complementary functions, such as proteins encoded by uORFs (Andrews and Rothnagel, 2014; Zhou *et al.*, 2022b) or lncRNAs (Lee *et al.*, 2021). Different dORFs with different functions could be interesting to investigate in the future.

The 3’ UTRs and dORFs of the MSVW dORF and HC dORF shortlists generally had greater relative ribosomal association than housekeeping genes 3’ UTRs. This suggests that generally, where dORFs aren’t present, 3’ UTR ribosomal association is low, especially compared to the gene, which has been suggested in the literature (Ingolia *et al.*, 2009). 3’ UTRs from the dORF shortlists could have increase 3’ UTR and dORFs ribosomal association due to dORF translation (Wu *et al.*, 2020b). The relative dORF and 3’ UTR ribosomal association varied with different cell types, tissues and disease states, such as human brain tissue appearing to have increased association compared to human aortic endothelial cells (Figure 5.9), or increased association with dORFs and 3’ UTRs in cancerous tissue compared to healthy tissue (Figure 3.10). Cell stress also appeared to have an influence and is discussed below. However, using relative 3’ UTR or dORF ribosomal association to the whole gene could mean that altered relative ribosomal association was reported when 3’ UTR or dORF ribosomal association stayed the same, and instead ribosomal association with other gene regions changed under different conditions. If dORF ribosomal association, and potential translation, changes in differing cellular conditions, or disease states, this could explain why dORF activity changes in different condition, such as the cancer datasets. This may be a feature they share with uORFs, where disease states, such as cancer or inflammation, and local conditions can alter uORF translation (Young and Wek, 2016; Sendoel *et al.*, 2017; Renz, Valdivia Francia and

Sendoel, 2020). Although changes in ribosomal association often affected both, some changes were specific to dORFs, or 3' UTRs excluding dORFs, alone. Another explanation for changes in 3' UTR ribosomal association in regions without shortlisted dORFs could be other dORF sequences. Although it seems likely that ribosomes associate with 3' UTRs for other reasons, whether through stop codon readthrough (Doronina and Brown, 2006; Namy and Rousset, 2010), or processes unrelated to translation (Guydosh and Green, 2014; Miettinen and Björklund, 2015).

In both shortlists when thapsigargin, an endoplasmic reticulum stress inducer through inhibition of the sarcoplasmic/endoplasmic reticulum calcium ATPase pump (Michelangeli and East, 2011; Peterková *et al.*, 2020), cycloheximide, or harringtonine, both of which are inhibitors of the elongation step of translation (Fresno, Jiménez and Vázquez, 1977; Schneider-Poetsch *et al.*, 2010), treatment was used, 3' UTR, including in housekeeping genes, and dORF ribosomal association increased. This could suggest that under stressful conditions, which can reduce cap-dependent translation, a greater proportion of ribosomes are found, or maintained, in 3' UTRs and dORFs compared to other gene regions. Stressful conditions could increase stop codon readthrough, or disrupt ribosomal recycling, increasing, or maintaining, 3' UTR ribosome presence. This could explain greater 3' UTR and dORF ribosome presence in cancer, potentially through a stress-driven response increasing ribosome entry into 3' UTRs. These 3' UTR ribosomes may not translate dORFs, especially if dORF translation was reduced by the stressful conditions (Wu *et al.*, 2020b). Increased dORF ribosomal association under conditions that inhibit translation is interesting, dORF translation seems unlikely, although some sORF translation is suggested to potentially differ from CDS cap-dependent translation (Zhou *et al.*, 2022b). When translation is inhibited, ribosomes could still be recruited to dORFs or 3' UTRs, by a currently unknown mechanism. Increased 3' UTR and dORF ribosome association with *ABCE1* knockdown can be explained more easily. ABCE1 is essential for recycling of ribosomes following translation termination (Pisarev *et al.*, 2010; Franckenberg, Becker and Beckmann, 2012; Jackson, Hellen and Pestova, 2012). Reduced *ABCE1* expression can lead to post-termination ribosomes remaining attached to mRNA which can pass into 3' UTRs and reinitiate (Skabkin *et al.*, 2013; Young *et al.*, 2015; Zinoviev, Hellen and Pestova, 2015; Mills *et al.*, 2016), potentially translating dORFs or 3' UTRs.

The two shortlists had differing results when investigating dORF and 3’ UTR ribosomal association, such as increased dORF and 3’ UTR ribosomal association for MSVW dORFs in cancerous tissue being lost in HC dORFs. Compared to MSVW dORFs, ribosomal association, particularly with dORFs, reduced with the HC dORF shortlist across the RP datasets. The dORFs in the two shortlists may have different functions, and this function could relate to the genes they appear in. Even if HC dORFs had functional encoded proteins, they would still require translating. Alternatively, HC dORFs could be a 3’ UTR motif, with another function, unrelated to translation, which could explain their low ribosomal association. The varying dORF and 3’ UTR ribosomal association across different conditions, could mean that the conditions, cell types and disease states used in this analysis were not the conditions where HC dORFs were most active, however, this seems unlikely.

5.7 Conclusions

dORFs are abundant in humans and other species, and dORFs show greater conservation across a range of species than surrounding regions of the 3’ UTR and full 3’ UTRs. This could indicate an evolutionary pressure to maintain these sequences, indicating a biological importance, or function for these dORFs. When comparing the most highly conserved dORFs across multiple species, these dORFs may occur more often in genes associated with transcriptional or translational regulation and often did not share the regulatory activity of Wu dORFs. With changing cell types, cellular conditions, and disease states, relative ribosomal association with, and potentially translation of, 3’ UTRs and dORFs changes. Stressful cellular conditions appear to increase relative 3’ UTR and dORF ribosomal association compared with the gene. The HC dORF shortlist differed from the MSVW dORF shortlist, with reduced ribosomal association, and some inconsistency in results across the different cell types and conditions. This chapter raises interesting questions about whether highly conserved dORFs and Wu dORFs have different regulatory functions, and how dORF and 3’ UTR ribosomal association may change in states of cell stress and whether these ribosomal associations are indicative of translation.

Chapter 6: Discussion

The overarching aim of this research was to investigate translational regulation of gene expression, particularly focussing on downstream open reading frames (dORFs) and their involvement in translational regulation. The presence of dORFs in 3' UTRs is not debated, they are found frequently (Bazzini *et al.*, 2014; Ji *et al.*, 2015; Mackowiak *et al.*, 2015; Chen *et al.*, 2020; Wu *et al.*, 2020b; Renz, Valdivia Francia and Sendoel, 2020; Wright *et al.*, 2022). Despite this there is little discussion of dORF function. Some may encode functional proteins (Bazzini *et al.*, 2014; Chong *et al.*, 2020; Ruiz Cuevas *et al.*, 2021). Alternatively, dORFs are suggested to be translated, acting as regulators of translation by increasing coding sequence (CDS) translation (Wu *et al.*, 2020b). The mechanism behind this function remains unknown, but is hypothesised to involve a looped model of mRNA where 3' UTR dORF translation passes initiation factors and ribosomal subunits to the 5' UTR and CDS start site, allowing more efficient CDS translation (Wu *et al.*, 2020b).

Ribosomes are found in 3' UTRs, however there is debate as to whether they are translating or not (Ingolia *et al.*, 2009; Guydosh and Green, 2014; Ji *et al.*, 2015; Miettinen and Björklund, 2015; Young *et al.*, 2015; Hsu *et al.*, 2016). 3' UTR ribosomal association is influenced by the cellular conditions, with stressful conditions increasing relative 3' UTR association (Ingolia *et al.*, 2009). This could mean these conditions influence dORF translation and function. Other 3' UTR translational regulators, including micro-RNAs (miRNAs), AU-rich elements (AREs) and cytoplasmic polyadenylation elements (CPEs), can have varied effects on translation. These regulators often involve binding of proteins or nucleic acids to 3' UTRs, associations that could be impacted by dORFs and translating ribosomes. Translation of 5' UTR uORFs is typically associated with repression of CDS translational efficiency, the opposite effect to that proposed for dORFs (Mueller and Hinnebusch, 1986; Krishna M Vattem and Wek, 2004; Brar *et al.*, 2012; Von Arnim, Jia and Vaughn, 2014; Wethmar *et al.*, 2014; Chew, Pauli and Schier, 2016; Johnstone, Bazzini and Giraldez, 2016; Renz, Valdivia Francia and Sendoel, 2020). Although uORF and dORF functions are suggested to be opposed, this relates to their location: translation of uORFs decreases CDS translation, whereas translation of dORFs does not, and instead may support CDS translation (Wu *et al.*, 2020b).

Regulating translation is essential to healthy cell function, and as research into translational regulation continues to expand, a growing number of diseases are associated with dysregulation of translation. The importance of translational regulation highlights the need to understand dORFs and their function, especially considering so much remains unknown. The worldwide impact of cancers, links to dysregulation of translation, and influence of cancers on uORF regulation, suggest that dORF function could be affected by cancers and dORFs could be involved in cancer pathology. To this end, ribosomal association with, and translation of, dORFs and their function was explored in healthy and cancerous tissues. Analysis of dORF composition sought to gain a better understanding of dORFs and whether this could be used identify further dORFs. Wu dORFs, the dORFs published by Wu *et al.* (2020b), were presented alongside the proposal of translational regulation by dORFs and the hypothesis mechanism of action (Wu *et al.*, 2020b). Investigating other dORFs and their function allows comparison with Wu dORFs and may determine whether this regulatory function applies to all dORF sequences. Conservation of dORFs would demonstrate their biological importance, especially by expanding on the evidence that some Wu dORFs are conserved in mice and zebrafish (Wu *et al.*, 2020b). This understanding of dORFs would be furthered by knowing how abundant dORFs and conserved dORFs are. uORF function and 3' UTR ribosomal association are affected by local conditions. Exploring 3' UTR and dORF ribosomal association across various cell types, cellular conditions, disease states, and treatments will expand the understanding of 3' UTR ribosomal association and establish whether these changes affect dORF ribosomal association. This could highlight cells or tissues where dORFs are most active and whether treatments, conditions or disease states impact dORF function.

dORF activity was found when investigating Wu dORFs in healthy kidney tissue. Translation of transcripts with Wu dORFs was upregulated compared to transcripts without these dORFs (Figure 3.2). In kidney tumour tissue (Figure 3.3) and A549 pulmonary adenocarcinomic human alveolar basal epithelial cells (Figure 3.13) this effect in transcripts with Wu dORFs is reduced but still present, whereas in RKO human colorectal carcinoma cells dORF activity is lost (Figure 3.16). Generally, Wu dORF activity could reduce in cancer, and may vary across different cancer types; although, the reason for this is unknown. However, some Wu dORF-containing transcripts had increased potential dORF activity in cancer datasets. Future research could investigate Wu dORF activity in oncogenes and tumour suppressors in cancer datasets, exploring whether dORFs found in oncogenes may maintain, or even increase, their activity, potentially driving cancer progression. The reason for reduced dORF activity in cancer was not explained by changes in 3' UTR processing which can influence translational regulation and could disrupt dORFs (Kuersten and Goodwin, 2003; Sandberg *et al.*, 2008; Mayr and Bartel, 2009; Singh *et al.*, 2009).

As 3' UTR ribosomal association increased the same was seen in CDSs (Figures 3.6, 3.7, 3.15, 3.18). Interestingly there was little evidence of any relationship between 3' UTR ribosomal presence and the translational regulation of the transcript (Figures 3.4, 3.5, 3.14, 3.17). If dORF translation is responsible for their regulatory function (Wu *et al.*, 2020b), it was anticipated that greater 3' UTR ribosomal association would be seen alongside increased CDS translation. Wu dORFs have ribosomal association, although there was little conclusive evidence of translation. Increased ribosome profiling (RP) read density in Wu dORFs compared to 3' UTRs (Figure 3.10), and distribution of RP reads across different reading frames, showing a preference different to 3' UTRs and somewhat similar to CDSs (Figure 3.12). This provides some signs that dORFs could be translated. Although still debated, the reading frame analysis suggested that generally 3' UTR ribosomes were not translating. Some dORFs are translated, their encoded proteins have been found using mass spectrometry (MS), and some of these dORFs are Wu dORFs.

Wu dORFs were not like 'mini-CDSs', their composition was more similar to 3' UTRs than CDSs, considering nucleotide (Table 4.4), dinucleotide (Figure 4.4) and trinucleotide compositions (Figure 4.6). Similarity between dORFs and CDSs seemed likely due to potential translation of dORFs compared to the remaining 3' UTR. Generally, the composition of Wu dORFs was unremarkable when compared to 3' UTRs, meaning the composition is not a reasonable method of identifying other active dORFs.

Ribosomal association with Wu dORFs, and the dORF-containing 3' UTRs, could be driven by CDS stop codon readthrough (Doronina and Brown, 2006; Namy and Rousset, 2010), and although not proposed in the hypothesised dORF mechanism of action (Wu *et al.*, 2020b), increased stop codon readthrough could increase dORF translation. In this case, 3' UTR ribosomes would be found in the same reading frame as the CDS, and the dORF peptide would be part of the extended CDS protein. If translation of dORFs was explained by readthrough, they would be less likely to have regulatory function, unless it was part of the readthrough mechanism, or through activity of the peptide which would need to be cleaved from the extended readthrough protein. Although readthrough cannot be ruled out, Wu dORF transcripts did not have a different CDS stop codon preference when compared with the rest of the genome (Table 4.5). Based on their CDS stop codon, transcripts with Wu dORFs are not more predisposed to readthrough, instead it remains more likely that dORF translation is driven by reinitiation and ribosomal recruitment (Wu *et al.*, 2020b).

When moving beyond Wu dORFs to investigate other dORF sequences, they were found to be widespread throughout 3' UTRs (Table 4.1). When compared to the number of potential dORF sequences that were found in randomised 3' UTR sequences, dORFs appeared more frequently. Although many of these potential dORF sequences had ribosomal association (Table 4.2), most lacked any translational regulation function (Figure 4.2) and have low ribosomal association. The lack of function for many dORF sequences, even when associated with ribosomes, raises important future enquiries into why some dORFs are functional and others may not be. What is it about Wu dORFs that may give them function? There could be tissue specificity, cellular conditions, undiscovered motifs, or criteria related to the translation or termination in the rest of the transcript, that make some dORFs active.

dORFs are also abundant in many other species, from chimpanzees to rice. Across a range of species, dORFs using an AUG start codon (AUG dORFs) were generally more conserved than their 3' UTRs (Figure 5.4 and Table 5.2) and the regions of 3' UTR surrounding the AUG dORFs (Table 5.5). Across species AUG dORFs appear around once (0.75-1.32) per 100 nucleotides in the 3' UTR (Table 5.1), highlighting their abundance. Where AUG dORFs were not found, the 3' UTRs were usually shorter, around 100 nucleotides long (Figure 5.2). Finding that AUG dORFs were more conserved than their 3' UTRs indicates their biological importance and implies function across species. Whether they are all potential translational regulators, or they encode functional proteins, dORFs are being conserved. The most highly conserved AUG dORFs often did not show the potential translational regulation seen with Wu dORFs (Table 5.9). However, these highly conserved dORFs seemed to be found more often in genes associated with transcriptional or translational regulation. Short ORFs (sORFs) have diverse functions, ranging from translational regulation through sORF translation (Iacono, Mignone and Pesole, 2005; Neafsey and Galagan, 2007; Johnstone, Bazzini and Giraldez, 2016; Ruiz-Orera and Albà, 2019), to encoding proteins active in cellular communication and signalling (Hashimoto *et al.*, 2001; Pauli *et al.*, 2014; Matsumoto *et al.*, 2017; Polycarpou-Schwarz *et al.*, 2018; Guo *et al.*, 2020; Wang *et al.*, 2020). It is possible that dORF functions differ between genes.

Cellular stress can influence 3' UTR ribosomal association (Ingolia *et al.*, 2009), and sORF translation is suggested to be tissue-specific (Couso and Patraquim, 2017). Considering this alongside the varied dORF activity and ribosomal association with 3' UTRs and dORFs in healthy and cancerous kidney tissue, led to exploration into the influence of tissue type, cellular conditions, disease states, and treatments on dORF and 3' UTR ribosomal association. When considering the shortlisted MSVW dORFs, which are Wu dORFs with MS validation of translation, and HC dORFs, the highly conserved dORFs from the conservation analysis, ribosomal association with 3' UTRs and dORFs changes depending on cell types, cellular conditions, and disease states (Figures 5.7-5.18). Under stressful cellular conditions relative 3' UTR and dORF ribosomal association often increased compared to the whole gene. There should be consideration of varying cell types, disease states, and cellular conditions when discussing dORF translation and 3' UTR ribosomal association. HC dORFs results showed inconsistency when compared to MSVW dORFs. Alongside the lack of evidence of translational regulation function for HC dORFs, they had low ribosomal association, especially compared to the MSVW dORF shortlist. This could be further evidence that dORF functions may be divergent.

Prior to this research there little was known about dORFs despite their abundance in the 3' UTR (Bazzini *et al.*, 2014; Ji *et al.*, 2015; Mackowiak *et al.*, 2015; Chen *et al.*, 2020; Wu *et al.*, 2020b; Renz, Valdivia Francia and Sendoel, 2020; Wright *et al.*, 2022). There were only a few reports of functional dORFs, either through encoded proteins (Chong *et al.*, 2020) or Wu dORFs with function as translational regulators (Wu *et al.*, 2020b). There was suggestion that parts of 3' UTRs could be conserved across species (Fournier *et al.*, 1994; Knee, Pitcher and Murphy, 1994; Silverman, 1994; Lipman, 1997), and that some Wu dORFs were conserved in zebrafish and mice, but the sequences generally were not (Wu *et al.*, 2020b). Wu dORF function was expected be consistent across different cell types and conditions (Wu *et al.*, 2020b).

dORFs have been at the centre of this research, increasing understanding of dORFs and their role in translational regulation. This research has raised new questions about dORFs and much still remains to be explored in relation to their function and mechanisms of action. dORFs are biologically important, highlighted by their potential function as translational regulators and their conservation across species. In this case the conservation also applies to dORF sequences, with the dORF sequence being more conserved than the surrounding 3' UTR. The potential translational upregulation in transcripts with Wu dORFs in healthy tissue supports the previous findings about dORFs in cells (Wu *et al.*, 2020b). This analysis has started to build a picture of dORF translation, some dORFs associate with ribosomes, whereas others do not. A novel finding is that disease states, such as cancer, and different cell types or cellular conditions also impact ribosomal association with dORFs, and potentially their translation. This is also seen with 3' UTR ribosomal association, which has previously been shown to change under stressful conditions (Ingolia *et al.*, 2009). This is particularly apparent in cancer for both dORFs and 3' UTRs. dORFs may play a crucial role in cancers, the regulatory activity of Wu dORFs appears to be reduced or lost in cancers when compared to healthy tissue. Wu dORFs composition does not highlight anything to distinguish these from other dORFs or 3' UTRs that they are found within. Investigations expanded beyond the Wu dORFs to include more dORFs in functional and conservation analysis. Wu dORFs and other dORFs could have different functions, with some acting as translational regulators, but others lacking that function whilst being conserved across species. Much like uORFs and other sORFs, the function of dORFs is not consistent and there may be different functions acting through different mechanisms. The lack of ribosomal association, especially with highly conserved dORFs, and difficulty identifying dORF translation could indicate that translation may not be central to the function of every dORF.

Although there is evidence for dORFs as potential translational regulators, there is some evidence that disputes this. Instead of having an alternative function, some dORFs could be randomly generated in the genome (Couso and Patraquim, 2017) and potentially associate with ribosomes that are not translating (Guydosh and Green, 2014; Miettinen and Björklund, 2015), or have undergone stop codon readthrough (Doronina and Brown, 2006; Namy and Rousset, 2010). This could be supported by the large number of dORFs that have no evidence of ribosomal association and the lack of activity for some dORFs that are associated with ribosomes. Across different samples and datasets, dORF activity can also vary and often dORFs do not have consistent ribosomal association. When dORF ribosomal association increases this is often not accompanied by greater dORF activity. The lack of publications demonstrating dORF activity also means there is only a relatively small amount of evidence behind them. Future investigation will be needed to find further evidence of dORFs and to understand why some dORFs could have no activity.

This research found two groups of dORFs which look particularly interesting for future investigation. One group are a subset of Wu dORFs which have MS validation of their translation. These dORFs have good evidence of translation and appear to have function as translational regulators, with some showing interesting changes in dORF activity in cancers. The other group are dORFs with highest conservation across multiple species, although these lack the ribosomal association and evidence of translational regulation of the other group, they are implied to be biologically important by their conservation, and the genes they appear in are often associated with transcriptional or translational regulation.

dORFs, due to their potential biological importance, should be an important consideration in mRNA exploration, translational regulation, and 3’ UTRs. dORFs expand the understanding of the complexity of regulating gene expression and provide another example of 3’ UTR regulators. This research shows how important dORFs could be and the importance of considering sORFs moving forward. dORFs are relatively unexplored, meaning many systems could be influenced by dORFs. The importance of translation regulation in a range of diseases from metabolic (Morita *et al.*, 2013) to neurological disorders (Buffington, Huang and Costa-

Mattioli, 2014) highlights the potential influence of dORFs with translational regulation function. There are many diseases where dORF function could be explored to understand whether they play a part in the pathology. The importance of dORFs may depend on the mRNA they exist in and the gene's function.

dORFs could form a future therapeutic option for a variety of diseases, targeting dORFs, or dORF start sites, for mutation in oncogenes could reduce translation of oncogenic mRNAs. Alternatively, dORFs could be introduced into target mRNAs, such as tumour suppressors to increase translation of these mRNAs. Aside from the translational regulation function, if dORFs encode functional proteins, these proteins could become future drug targets, or dORFs could be disrupted to prevent their production. Understanding that dORFs could be translated in 3' UTRs, brings greater complexity to consideration of mRNA structure, especially 3' UTRs. Additionally, dORF translation could have implications for other 3' UTR regulators which depend on 3' UTR binding, such as miRNAs. Alternative polyadenylation and 3' UTR truncation have the potential to also disrupt dORF sequences, potentially through removal of part, or the whole, of the dORF from the 3' UTR. This could have consequences depending on the disrupted, or removed, dORF function. The conservation of dORFs across species and difficulties accessing accurately annotated 3' UTR sequences in some species, shows how important developing comprehensive libraries of accurately annotated 3' UTRs for as many species as possible is. This is not just important for dORFs in other species, but also for other 3' UTR translational regulators.

There is a large amount of data from analysis in the final chapter, particularly around ribosomal association with 3' UTRs and dORFs relative to the gene (Section 5.5). In the future, investigating specific dORFs would be beneficial to see how the different cell types, treatments, and cellular condition influence ribosomal association with a particular dORF. This could identify candidate dORFs to study under different conditions to understand the mechanism of dORF translation. The conservation analysis presents an opportunity to explore dORF function in other species and allow comparison to humans. Generating additional paired RNAseq and RP datasets, especially with interesting treatments or cellular conditions could allow further study of Wu dORF function and highly conserved dORFs functions. To demonstrate dORF translation exploration of proteomics datasets to search for dORF peptides would also be useful.

This research shows how much still remains to be understood about dORFs and raises some interesting questions for future study. A key piece of research would explore why some dORFs appear to be functional and others are not, and what is driving this difference. Whether this is linked to tissue, or cell, specificity or something else. dORF translation does not seem to be related to stop codon readthrough, meaning it will be important to understand how dORFs recruit ribosomes. Where dORFs act as translational regulators there remains questions about how this happens, is it driven by the translation of the dORF? And does it match the published hypothesis (Wu *et al.*, 2020b)? There is also scope to explore the reasons behind the increased 3' UTR ribosomal association seen in cancer, and the potential reduction in dORF activity, and whether these occur in a range of different cancers. The reasons could involve stop codon readthrough or the influence of local conditions caused by cancer. It has been mentioned previously, but understanding how dORFs interact with, and are influenced by, other 3' UTR regulators would be very interesting.

Future Work

To explore dORF function and mechanism of action, a luciferase dual reporter system could be employed (Branchini *et al.*, 2018), similar to that used in original investigations of dORFs (Wu *et al.*, 2020b). A dual reporter system would report the CDS and dORF translation in a construct, allowing investigation of the effect of changes in, or loss of, dORF translation on the CDS (Branchini *et al.*, 2018). The CRISPR/Cas9 nuclease system could be used to modify the dORF start codon, sequence, length, and the region upstream of the dORF (Canver *et al.*, 2014; Mitschka, Fansler and Mayr, 2021). This would allow investigation of how these changes affect dORF regulatory activity and may establish which dORF regions are essential to their function. Secondary structure modelling of dORF-containing mRNAs, using a tool like Wfold (Yuan, Yang and Zhang, 2024), would also be interesting to explore dORF translation. With particular focus on the region between the CDS stop codon and the dORF start codon, searching for structures, or motifs, within this region to identify how ribosomes are recruited to dORFs. Additionally resolving mRNA structures, using methods such as cryo-electron microscopy, to show that dORF-containing transcripts form a looped structure would support the hypothesised dORF mechanism of action (Gabashvili *et al.*, 2000; Wu *et al.*, 2020b; Cao *et al.*, 2024).

Development of more paired RNAseq and RP datasets would also support further dORF investigation. Further comparisons in paired RNAseq and RP datasets between healthy and tumour tissue could explore the differences in cancers. Paired datasets for a wider range of cellular conditions, tissue types, treatments, and disease states would help to explore the variability of dORF activity. Generating RP datasets with greater sequencing depth would also be useful to increase the number of dORF aligned reads, providing a better opportunity for investigation of dORF translation (Tomuro *et al.*, 2024). Using proteomics datasets and MS validation of short proteins (Hartman *et al.*, 2024) is another way to demonstrate dORF translation. Methods such as Rp3, a ribosome profiling-assisted proteogenomic approach, could identify dORF encoded proteins (Vieira de Souza *et al.*, 2024). The potential for dORFs to have alternative functions to the proposed translational regulation also warrants investigation of the function of dORF peptides. The highly conserved dORFs would be a good place to start when investigating functional proteins.

In the future, the ideal way to investigate dORF function would involve real-time visualisation of ribosomes and translational factors on mRNA. Tagging ribosomal subunits, would allow visualisation of ribosomal recruitment to, and translation of, dORFs, and whether the ribosomal subunits are then passed to 5’ UTRs or CDS start sites. This could help to understand whether the hypothesised dORF mechanism of action, involving dORF translation driving increased CDS translation, is accurate. Otherwise, this would be a useful method to show dORF translation and how this occurs. There have been developments in methodologies focussed on tracking translation in real-time (Qureshi and Duss, 2024). There is still much about dORFs to be understood, meaning future investigations could go in many directions. This research has shown that dORFs are conserved, biologically important, and some could function as regulators of translation.

Conclusion

In conclusion, dORFs could be biologically important. dORFs are associated with ribosomes, however, ribosomal association with dORFs and 3’ UTRs varies depending on cellular conditions, tissue types and disease states. In addition to ribosomal association, dORF translation is confirmed by identification of their encoded proteins in protein databases. Some dORFs can act as translational regulators, increasing CDS translation in the transcript they appear in. This dORF regulatory activity appears to reduce in cancer through an unknown mechanism. Despite ribosomal association, some dORFs show no evidence of regulatory function and others lack consistent activity across samples and datasets. dORFs can still be disputed and need future investigation into dORF activity and why some dORFs could have no function. The nucleotide, dinucleotide, and trinucleotide composition of dORFs is unremarkable when compared to 3’ UTRs, sharing more similarity with this region than CDSs. dORFs are conserved across species, dORF sequences are more conserved than their 3’ UTRs and 3’ UTR regions surrounding the dORFs, highlighting their biological importance. dORFs may have divergent functions, some may act as translational regulators, whereas others may have different activity. dORFs are a novel, interesting, and important feature of mRNAs that require further study to understand more about their function and mechanism of action.

References

Afgan, E. *et al.* (2018) ‘The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update’, *Nucleic Acids Research*, 46(W1), pp. W537–W544. doi: 10.1093/nar/gky379.

Ahmed, N. *et al.* (2019) ‘Identifying A- and P-site locations on ribosome-protected mRNA fragments using Integer Programming’, *Scientific Reports*, 9(1). doi: 10.1038/S41598-019-42348-X.

Alekhina, O. M. *et al.* (2020) ‘Functional cyclization of eukaryotic mRNAs’, *International Journal of Molecular Sciences*, 21(5). doi: 10.3390/ijms21051677.

Alkalaeva, E. Z. *et al.* (2006) ‘In vitro reconstitution of eukaryotic translation reveals cooperativity between release factors eRF1 and eRF3’, *Cell*, 125(6), pp. 1125–1136. doi: 10.1016/J.CELL.2006.04.035.

Altschul, S. F. *et al.* (1997) ‘Gapped BLAST and PSI-BLAST: a new generation of protein database search programs’, *Nucleic Acids Research*, 25(17), pp. 3389–3402. doi: 10.1093/NAR/25.17.3389.

Amrani, N. *et al.* (2008) ‘Translation factors promote the formation of two states of the closed-loop mRNP’, *Nature*, 453(7199), pp. 1276–1280. doi: 10.1038/nature06974.

Amrani, N., Sachs, M. S. and Jacobson, A. (2006) ‘Early nonsense: mRNA decay solves a translational problem’, *Nature Reviews Molecular Cell Biology*. Nat Rev Mol Cell Biol, pp. 415–425. doi: 10.1038/nrm1942.

Anders, S., Pyl, P. T. and Huber, W. (2015) ‘HTSeq—a Python framework to work with high-throughput sequencing data’, *Bioinformatics*, 31(2), pp. 166–169. doi: 10.1093/BIOINFORMATICS/BTU638.

Anderson, D. M. *et al.* (2015) ‘A micropeptide encoded by a putative long noncoding RNA regulates muscle performance’, *Cell*, 160(4), pp. 595–606. doi: 10.1016/j.cell.2015.01.009.

Andreev, D. E. *et al.* (2015) ‘Translation of 5’ leaders is pervasive in genes resistant to eIF2 repression’, *eLife*, 2015(4). doi: 10.7554/ELIFE.03971.

Andrews, S. (2010) *Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data*. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (Accessed: 22 February 2021).

Andrews, S. J. and Rothnagel, J. A. (2014) ‘Emerging evidence for functional peptides encoded by short open reading frames’, *Nature reviews. Genetics*, 15(3), pp. 193–204. doi: 10.1038/NRG3520.

Antequera, F. (2003) ‘Structure, function and evolution of CpG island promoters’, *Cellular and molecular life sciences : CMLS*, 60(8), pp. 1647–1658. doi: 10.1007/S00018-003-3088-6.

Antequera, F. and Bird, A. (1999) ‘CpG islands as genomic footprints of promoters that are associated with replication origins', *Current biology : CB*, 9(17). doi: 10.1016/S0960-9822(99)80418-7.

Von Arnim, A. G., Jia, Q. and Vaughn, J. N. (2014) 'Regulation of plant translation by upstream open reading frames', *Plant Science*. Plant Sci, pp. 1–12. doi: 10.1016/j.plantsci.2013.09.006.

Arribere, J. A. and Gilbert, W. V. (2013) 'Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing', *Genome Research*, 23(6), pp. 977–987. doi: 10.1101/gr.150342.112.

Aspden, J. L. *et al.* (2014) 'Extensive translation of small open reading frames revealed by poly-ribo-seq', *eLife*, 3(August2014), pp. 1–19. doi: 10.7554/ELIFE.03528.

Attwood, T. K. *et al.* (2019) 'A global perspective on evolving bioinformatics and data science training needs', *Briefings in Bioinformatics*, 20(2), p. 398. doi: 10.1093/BIB/BBX100.

Bakhti, S. Z. and Latifi-Navid, S. (2021) 'Non-coding RNA-Encoded Peptides/Proteins in Human Cancer: The Future for Cancer Therapy', *Current Medicinal Chemistry*, 29(22), pp. 3819–3835. doi: 10.2174/0929867328666211111163701.

Barbosa, C., Peixeiro, I. and Romão, L. (2013) 'Gene Expression Regulation by Upstream Open Reading Frames and Human Disease', *PLoS Genetics*. Edited by E. M. C. Fisher, 9(8), p. e1003529. doi: 10.1371/journal.pgen.1003529.

Barkoff, A., Ballantyne, S. and Wickens, M. (1998) 'Meiotic maturation in Xenopus requires polyadenylation of multiple mRNAs', *EMBO Journal*, 17(11), pp. 3168–3175. doi: 10.1093/emboj/17.11.3168.

Bartel, D. P. (2018) 'Metazoan MicroRNAs', *Cell*. Cell Press, pp. 20–51. doi: 10.1016/j.cell.2018.03.006.

Barth-Baus, D. *et al.* (2013) 'Influence of translation factor activities on start site selection in six different mRNAs', *Translation*, 1(1), p. e24419. doi: 10.4161/trla.24419.

Barthelme, D. *et al.* (2007) 'Structural organization of essential iron-sulfur clusters in the evolutionarily highly conserved ATP-binding cassette protein ABCE1', *The Journal of biological chemistry*, 282(19), pp. 14598–14607. doi: 10.1074/JBC.M700825200.

Barthelme, D. *et al.* (2011) 'Ribosome recycling depends on a mechanistic link between the FeS cluster domain and a conformational switch of the twin-ATPase ABCE1', *Proceedings of the National Academy of Sciences of the United States of America*, 108(8), pp. 3228–3233. doi: 10.1073/PNAS.1015953108/-/DCSUPPLEMENTAL.

Bartish, M. *et al.* (2023) 'The role of eIF4F-driven mRNA translation in regulating the tumour microenvironment', *Nature Reviews Cancer 2023 23:6*, 23(6), pp. 408–425. doi: 10.1038/s41568-023-00567-5.

Bartlett, A., Penders, B. and Lewis, J. (2017) 'Bioinformatics: indispensable, yet hidden in plain sight?', *BMC Bioinformatics*, 18(1). doi: 10.1186/S12859-017-1730-9.

Bashirullah, A., Cooperstock, R. L. and Lipshitz, H. D. (1998) 'RNA localization in development', *Annual review of biochemistry*, 67, pp. 335–394. doi: 10.1146/ANNUREV.BIOCHEM.67.1.335.

Basrai, M. A., Hieter, P. and Boeke, J. D. (1997) 'Small Open Reading Frames: Beautiful Needles in the Haystack', *Genome Research*, 7(8), pp. 768–771. doi: 10.1101/GR.7.8.768.

Bazzini, A. A. *et al.* (2014) 'Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation', *EMBO Journal*, 33(9), pp. 981–993. doi: 10.1002/embj.201488411.

Behm-Ansmant, I. *et al.* (2006) 'mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes', *Genes and Development*, 20(14), pp. 1885–1898. doi: 10.1101/gad.1424106.

Beier, H. and Grimm, M. (2001) 'Misreading of termination codons in eukaryotes by natural nonsense suppressor tRNAs', *Nucleic Acids Research*, 29(23), p. 4767. doi: 10.1093/NAR/29.23.4767.

Bellelli, R. *et al.* (2018) 'Polε Instability Drives Replication Stress, Abnormal Development, and Tumorigenesis', *Molecular Cell*, 70(4), p. 707. doi: 10.1016/J.MOLCEL.2018.04.008.

Bergman, S., Diament, A. and Tuller, T. (2021) 'New computational model for miRNA-mediated repression reveals novel regulatory roles of miRNA bindings inside the coding region', *Bioinformatics*, 36(22–23), pp. 5398–5404. doi: 10.1093/bioinformatics/btaa1021.

Bertram, G. *et al.* (2000) 'Terminating eukaryote translation: domain 1 of release factor eRF1 functions in stop codon recognition.', *RNA*, 6(9), p. 1236. doi: 10.1017/S1355838200000777.

Bhatta, A. *et al.* (2020) 'A Mitochondrial Micropeptide Is Required for Activation of the Nlrp3 Inflammasome', *Journal of immunology (Baltimore, Md. : 1950)*, 204(2), pp. 428–437. doi: 10.4049/JIMMUNOL.1900791.

Bidou, L. *et al.* (2004) 'Premature stop codons involved in muscular dystrophies show a broad spectrum of readthrough efficiencies in response to gentamicin treatment', *Gene therapy*, 11(7), pp. 619–627. doi: 10.1038/SJ.GT.3302211.

Biswas, J. *et al.* (2019) 'Fluorescence imaging methods to investigate translation in single cells', *Cold Spring Harbor Perspectives in Biology*, 11(4), p. a032722. doi: 10.1101/cshperspect.a032722.

Blankenberg, D. *et al.* (2007) 'A framework for collaborative analysis of ENCODE data: Making large-scale analyses biologist-friendly', *Genome Research*, 17(6), pp. 960–964. doi: 10.1101/gr.5578007.

Blankenberg, D. *et al.* (2010) 'Manipulation of FASTQ data with Galaxy', *Bioinformatics*, 26(14), pp. 1783–1785. doi: 10.1093/BIOINFORMATICS/BTQ281.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) ‘Trimmomatic: a flexible trimmer for Illumina sequence data’, *Bioinformatics*, 30(15), pp. 2114–2120. doi: 10.1093/BIOINFORMATICS/BTU170.

Bönisch, C. *et al.* (2007) ‘Degradation of hsp70 and other mRNAs in Drosophila via the 5′-3′ pathway and its regulation by heat shock’, *Journal of Biological Chemistry*, 282(30), pp. 21818–21828. doi: 10.1074/jbc.M702998200.

Branchini, B. R. *et al.* (2018) ‘A Firefly Luciferase Dual Color Bioluminescence Reporter Assay Using Two Substrates To Simultaneously Monitor Two Gene Expression Events’, *Scientific Reports 2018 8:1*, 8(1), pp. 1–7. doi: 10.1038/s41598-018-24278-2.

Brar, G. A. *et al.* (2012) ‘High-resolution view of the yeast meiotic program revealed by ribosome profiling’, *Science*. American Association for the Advancement of Science, pp. 552–557. doi: 10.1126/science.1215110.

Breaker, R. R. (2018) ‘Riboswitches and translation control’, *Cold Spring Harbor Perspectives in Biology*, 10(11), p. a032797. doi: 10.1101/cshperspect.a032797.

Brenner, S., Stretton, A. O. W. and Kaplan, S. (1965) ‘Genetic code: The “nonsense” triplets for chain termination and their suppression’, *Nature*, 206(4988), pp. 994–998. doi: 10.1038/206994a0.

Brogna, S. and Wen, J. (2009) ‘Nonsense-mediated mRNA decay (NMD) mechanisms’, *Nature Structural and Molecular Biology*. Nat Struct Mol Biol, pp. 107–113. doi: 10.1038/nsmb.1550.

Brown, A. *et al.* (2015) ‘Structural basis for stop codon recognition in eukaryotes’, *Nature*, 524(7566), p. 493. doi: 10.1038/NATURE14896.

Brůna, T. *et al.* (2021) ‘BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database’, *NAR Genomics and Bioinformatics*, 3(1), pp. 1–11. doi: 10.1093/NARGAB/LQAA108.

Brůna, T., Lomsadze, A. and Borodovsky, M. (2020) ‘GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins’, *NAR Genomics and Bioinformatics*, 2(2). doi: 10.1093/NARGAB/LQAA026.

Brunet, M. A. *et al.* (2018) ‘Recognition of the polycistronic nature of human genes is critical to understanding the genotype-phenotype relationship’, *Genome Research*. Cold Spring Harbor Laboratory Press, pp. 609–624. doi: 10.1101/gr.230938.117.

Brunet, M. A. *et al.* (2021) ‘OpenProt 2021: deeper functional annotation of the coding potential of eukaryotic genomes’, *Nucleic Acids Research*, 49(D1), pp. D380–D388. doi: 10.1093/NAR/GKAA1036.

Buffington, S. A., Huang, W. and Costa-Mattioli, M. (2014) ‘Translational control in synaptic plasticity and cognitive dysfunction’, *Annual Review of Neuroscience*. Annual Reviews Inc., pp. 17–38. doi: 10.1146/annurev-neuro-071013-014100.

Calviello, L. *et al.* (2016) ‘Detecting actively translated open reading frames in ribosome profiling data’, *Nature Methods*, 13(2), pp. 165–170. doi: 10.1038/nmeth.3688.

Camacho, C. *et al.* (2009) ‘BLAST+: architecture and applications’, *BMC Bioinformatics*, 10(1), p. 421. doi: 10.1186/1471-2105-10-421.

Canver, M. C. *et al.* (2014) ‘Characterization of Genomic Deletion Efficiency Mediated by Clustered Regularly Interspaced Palindromic Repeats (CRISPR)/Cas9 Nuclease System in Mammalian Cells’, *Journal of Biological Chemistry*, 289(31), pp. 21312–21324. doi: 10.1074/JBC.M114.564625.

Cao, X. *et al.* (2021) ‘Alt-RPL36 downregulates the PI3K-AKT-mTOR signaling pathway by interacting with TMEM24’, *Nature Communications 2021 12:1*, 12(1), pp. 1–15. doi: 10.1038/s41467-020-20841-6.

Cao, X. *et al.* (2024) ‘Identification of RNA structures and their roles in RNA functions’, *Nature Reviews Molecular Cell Biology 2024 25:10*, 25(10), pp. 784–801. doi: 10.1038/s41580-024-00748-6.

Carninci, P. *et al.* (2005) ‘Molecular biology: The transcriptional landscape of the mammalian genome’, *Science*, 309(5740), pp. 1559–1563. doi: 10.1126/SCIENCE.1112014/SUPPL_FILE/P1-2CAGE_MOUSE_ACCESSION.LIST.ZIP.

Carninci, P. *et al.* (2006) ‘Genome-wide analysis of mammalian promoter architecture and evolution’, *Nature genetics*, 38(6), pp. 626–635. doi: 10.1038/NG1789.

Carvunis, A. R. *et al.* (2012) ‘Proto-genes and de novo gene birth’, *Nature 2012 487:7407*, 487(7407), pp. 370–374. doi: 10.1038/nature11184.

Cazzola, M. and Skoda, R. C. (2000) ‘Translational pathophysiology: A novel molecular mechanism of human disease’, *Blood*. W.B. Saunders, pp. 3280–3288. doi: 10.1182/blood.v95.11.3280.011k41_3280_3288.

Chang, Y. F., Imam, J. S. and Wilkinson, M. F. (2007) ‘The Nonsense-mediated decay RNA surveillance pathway’, *Annual Review of Biochemistry*, 76, pp. 51–74. doi: 10.1146/ANNUREV.BIOCHEM.76.050106.093909.

Chen, C. H. *et al.* (2017) ‘Upregulation of MARCKS in kidney cancer and its potential as a therapeutic target’, *Oncogene*, 36(25), p. 3588. doi: 10.1038/ONC.2016.510.

Chen, C. Y. A. and Shyu, A. Bin (2011) ‘Mechanisms of deadenylation-dependent decay’, *Wiley Interdisciplinary Reviews: RNA*. Wiley Interdiscip Rev RNA, pp. 167–183. doi: 10.1002/wrna.40.

Chen, J. *et al.* (2020) ‘Pervasive functional translation of noncanonical human open reading frames’, *Science*, 367(6482), pp. 1140-1146. doi: 10.1126/science.aay0262.

Chew, G. L., Pauli, A. and Schier, A. F. (2016) ‘Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish’, *Nature Communications*, 7. doi: 10.1038/ncomms11663.

Cho, S. J. *et al.* (2013) ‘Ninjurin1, a target of p53, regulates p53 expression and p53-dependent cell survival, senescence, and radiation-induced mortality’, *Proceedings of the National Academy of Sciences of the United States of America*, 110(23), pp. 9362–9367. doi: 10.1073/PNAS.1221242110/-/DCSUPPLEMENTAL.

Chodavarapu, R. K. *et al.* (2010) 'Relationship between nucleosome positioning and DNA methylation', *Nature*, 466(7304), pp. 388–392. doi: 10.1038/NATURE09147.

Choe, J. *et al.* (2018) 'mRNA circularization by METTL3–eIF3h enhances translation and promotes oncogenesis', *Nature*, 561(7724), pp. 556–560. doi: 10.1038/s41586-018-0538-8.

Chong, C. *et al.* (2020) 'Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes', *Nature communications*, 11(1), 1293. doi: 10.1038/s41467-020-14968-9.

Choteau, S. A. *et al.* (2021) 'MetamORF: a repository of unique short open reading frames identified by both experimental and computational approaches for gene and metagene analyses', *Database : the journal of biological databases and curation*, 2021. doi: 10.1093/DATABASE/BAAB032.

Christie, M. *et al.* (2013) 'Structure of the PAN3 pseudokinase reveals the basis for interactions with the PAN2 deadenylase and the GW182 proteins', *Molecular Cell*, 51(3), pp. 360–373. doi: 10.1016/j.molcel.2013.07.011.

Chugunova, A. *et al.* (2019) 'LINC00116 codes for a mitochondrial peptide linking respiration and lipid metabolism', *Proceedings of the National Academy of Sciences of the United States of America*, 116(11), pp. 4940–4945. doi: 10.1073/PNAS.1809105116/-/DCSUPPLEMENTAL.

Cock, P. J. A. *et al.* (2009) 'Biopython: freely available Python tools for computational molecular biology and bioinformatics', *Bioinformatics*, 25(11), p. 1422. doi: 10.1093/BIOINFORMATICS/BTP163.

Cock, P. J. A. *et al.* (2015) 'NCBI BLAST+ integrated into Galaxy', *GigaScience*, 4(1), p. 39. doi: 10.1186/s13742-015-0080-7.

Conne, B., Stutz, A. and Vassalli, J. D. (2000) 'The 3' untranslated region of messenger RNA: A molecular "hotspot" for pathology?', *Nature medicine*, 6(6), pp. 637–641. doi: 10.1038/76211.

Couso, J. P. and Patraquim, P. (2017) 'Classification and function of small open reading frames', *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, pp. 575–589. doi: 10.1038/nrm.2017.58.

Couttet, P. *et al.* (1997) 'Messenger RNA deadenylylation precedes decapping in mammalian cells', *Proceedings of the National Academy of Sciences of the United States of America*, 94(11), pp. 5628–5633. doi: 10.1073/pnas.94.11.5628.

D'Lima, N. G. *et al.* (2017) 'A human microprotein that interacts with the mRNA decapping complex', *Nature chemical biology*, 13(2), pp. 174–180. doi: 10.1038/NCHEMBIO.2249.

Dabrowski, M., Bukowy-Bieryllo, Z. and Zietkiewicz, E. (2015) 'Translational readthrough potential of natural termination codons in eucaryotes – The impact of RNA sequence', *RNA Biology*, 12(9), pp. 950–958. doi: 10.1080/15476286.2015.1068497.

Decker, C. J. and Parker, R. (1993) 'A turnover pathway for both stable and unstable mRNAs in yeast: Evidence for a requirement for deadenylation', *Genes and*

Development, 7(8), pp. 1632–1643. doi: 10.1101/gad.7.8.1632.

Delcourt, V. *et al.* (2018) ‘Small Proteins Encoded by Unannotated ORFs are Rising Stars of the Proteome, Confirming Shortcomings in Genome Annotations and Current Vision of an mRNA’, *PROTEOMICS*, 18(10), p. 1700058. doi: 10.1002/PMIC.201700058.

Dever, T. E., Dinman, J. D. and Green, R. (2018) ‘Translation elongation and recoding in eukaryotes’, *Cold Spring Harbor Perspectives in Biology*, 10(8), p. a032649. doi: 10.1101/cshperspect.a032649.

Dever, T. E. and Green, R. (2012) ‘The elongation, termination, and recycling phases of translation in eukaryotes’, *Cold Spring Harbor Perspectives in Biology*, 4(7), pp. 1–16. doi: 10.1101/cshperspect.a013706.

Diao, M. Q. *et al.* (2019) ‘RPS27, a sORF-Encoded Polypeptide, Functions Antivirally by Activating the NF-κB Pathway and Interacting With Viral Envelope Proteins in Shrimp’, *Frontiers in Immunology*, 10. doi: 10.3389/FIMMU.2019.02763.

Doherty, J. K. *et al.* (1999) ‘An alternative HER-2/neu transcript of 8 kb has an extended 3’UTR and displays increased stability in SKOV-3 ovarian carcinoma cells’, *Gynecologic oncology*, 74(3), pp. 408–415. doi: 10.1006/GYNO.1999.5467.

Doronina, V. A. and Brown, J. D. (2006) ‘Non-canonical decoding events at stop codons in eukaryotes’, *Molekuliarnaia biologiia.*, 40(4), pp. 731–741. doi: 10.1134/S0026893306040182.

Duchaine, T. F. and Fabian, M. R. (2019) ‘Mechanistic insights into microrna-mediated gene silencing’, *Cold Spring Harbor Perspectives in Biology*, 11(3), p. a032771. doi: 10.1101/cshperspect.a032771.

Dumesic, P. A. *et al.* (2019) ‘An Evolutionarily Conserved uORF Regulates PGC1α and Oxidative Metabolism in Mice, Flies, and Bluefin Tuna’, *Cell Metabolism*, 30(1), pp. 190-200.e6. doi: 10.1016/j.cmet.2019.04.013.

Dunn, J. G. *et al.* (2013) ‘Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster’, *eLife*, 2013(2), p. 1179. doi: 10.7554/eLife.01179.

Duret, L., Dorkeld, F. and Gautier, C. (1993) ‘Strong conservation of non-coding sequences during vertebrates evolution: potential involvement in post-transcriptional regulation of gene expression’, *Nucleic Acids Research*, 21(10), pp. 2315–2322. doi: 10.1093/NAR/21.10.2315.

Eberhardt, W. *et al.* (2007) ‘Modulation of mRNA stability as a novel therapeutic approach’, *Pharmacology & therapeutics*, 114(1), pp. 56–73. doi: 10.1016/J.PHARMTHERA.2007.01.002.

Eden, E. *et al.* (2009) ‘GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists’, *BMC Bioinformatics*, 10, p. 48. doi: 10.1186/1471-2105-10-48.

Ehrlich, M. and Wang, R. Y. H. (1981) ‘5-Methylcytosine in eukaryotic DNA’, *Science (New York, N.Y.)*, 212(4501), pp. 1350–1357. doi:

10.1126/SCIENCE.6262918.

Eldad, N., Yosefzon, Y. and Arava, Y. (2008) ‘Identification and characterization of extensive intra-molecular associations between 3′-UTRs and their ORFs’, *Nucleic Acids Research*, 36(21), pp. 6728–6738. doi: 10.1093/nar/gkn754.

Erhard, F. *et al.* (2018) ‘Improved Ribo-seq enables identification of cryptic translation events’, *Nat Methods*, 15(5), pp. 363-366. doi: 10.1038/nmeth.4631.

Esteller, M. (2008) ‘Epigenetics in evolution and disease’, *The Lancet*, 372, pp. S90–S96. doi: 10.1016/s0140-6736(08)61887-5.

Eswarappa, S. M. *et al.* (2014) ‘Programmed translational readthrough generates antiangiogenic VEGF-Ax’, *Cell*, 157(7), pp. 1605–1618. doi: 10.1016/j.cell.2014.04.033.

Eswarappa, S. M. and Fox, P. L. (2015) ‘Antiangiogenic VEGF-Ax: A new participant in tumor angiogenesis’, *Cancer Research*. American Association for Cancer Research Inc., pp. 2765–2769. doi: 10.1158/0008-5472.CAN-14-3805.

Ewels, P. *et al.* (2016) ‘MultiQC: summarize analysis results for multiple tools and samples in a single report’, *Bioinformatics*, 32(19), pp. 3047–3048. doi: 10.1093/BIOINFORMATICS/BTW354.

Fakim, H. and Fabian, M. R. (2019) ‘Communication Is Key: 5′–3′ Interactions that Regulate mRNA Translation and Turnover’, in *Advances in Experimental Medicine and Biology*. Springer, pp. 149–164. doi: 10.1007/978-3-030-31434-7_6.

Fan-Minogue, H. *et al.* (2008) ‘Distinct eRF3 Requirements Suggest Alternate eRF1 Conformations Mediate Peptide Release During Eukaryotic Translation Termination’, *Molecular cell*, 30(5), p. 599. doi: 10.1016/J.MOLCEL.2008.03.020.

Fearon, K. *et al.* (1994) *THE JOURNAI. OF BIOI.OGICAI. CHEMISTRY Premature Translation Termination Mutations Are Efficiently Suppressed in a Highly Conserved Region of Yeast SteGp, a Member of the ATP-binding Cassette (ABC) Transporter Family**. doi: 10.1016/S0021-9258(17)32379-7.

Fesenko, I. *et al.* (2019) ‘Distinct types of short open reading frames are translated in plant cells’, *Genome Research*, 29(9), pp. 1464–1477. doi: 10.1101/GR.253302.119.

Flavell, S. W. *et al.* (2008) ‘Genome-wide analysis of MEF2 transcriptional program reveals synaptic target genes and neuronal activity-dependent polyadenylation site selection’, *Neuron*, 60(6), p. 1022. doi: 10.1016/J.NEURON.2008.11.029.

Floquet, C. *et al.* (2012) ‘Statistical Analysis of Readthrough Levels for Nonsense Mutations in Mammalian Cells Reveals a Major Determinant of Response to Gentamicin’, *PLoS Genetics*, 8(3), p. 1002608. doi: 10.1371/JOURNAL.PGEN.1002608.

Fournier, S. *et al.* (1994) ‘Role for Low-Affinity Receptor for IgE (CD23) in Normal and Leukemic B-Cell Proliferation’, *Blood*, 84(6), pp. 1881–1886. doi: 10.1182/BLOOD.V84.6.1881.1881.

Fox, C. A., Sheets, M. D. and Wickens, M. P. (1989) ‘Poly(A) addition during maturation of frog oocytes: distinct nuclear and cytoplasmic activities and regulation by the sequence UUUUUAU', *Genes & development*, 3(12B), pp. 2151–2162. doi: 10.1101/GAD.3.12B.2151.

Franckenberg, S., Becker, T. and Beckmann, R. (2012) 'Structural view on recycling of archaeal and eukaryotic ribosomes after canonical termination and ribosome rescue', *Current opinion in structural biology*, 22(6), pp. 786–796. doi: 10.1016/J.SBI.2012.08.002.

François, P. *et al.* (2021) 'RiboDoc: A Docker-based package for ribosome profiling analysis', *Computational and Structural Biotechnology Journal*, 19, pp. 2851–2860. doi: 10.1016/J.CSBJ.2021.05.014.

Frankish, A. *et al.* (2019) 'GENCODE reference annotation for the human and mouse genomes', *Nucleic Acids Research*, 47(D1), pp. D766–D773. doi: 10.1093/nar/gky955.

Freitag, J., Ast, J. and Bölker, M. (2012) 'Cryptic peroxisomal targeting via alternative splicing and stop codon read-through in fungi', *Nature*, 485(7399), pp. 522–525. doi: 10.1038/nature11051.

Fresno, M., Jiménez, A. and Vázquez, D. (1977) 'Inhibition of translation in eukaryotic systems by harringtonine.', *European journal of biochemistry*, 72(2), pp. 323–30. doi: 10.1111/j.1432-1033.1977.tb11256.x.

Frischmeyer, P. A. and Dietz, H. C. (1999) 'Nonsense-mediated mRNA decay in health and disease', *Human Molecular Genetics*, 8(10), pp. 1893–1900. doi: 10.1093/HMG/8.10.1893.

Frolova, L. Y. *et al.* (1999) 'Mutations in the highly conserved GGQ motif of class 1 polypeptide release factors abolish ability of human eRF1 to trigger peptidyl-tRNA hydrolysis.', *RNA*, 5(8), p. 1014. doi: 10.1017/S135583829999043X.

Fu, G. *et al.* (2013) 'MicroRNAs in human placental development and pregnancy complications', *International Journal of Molecular Sciences*. Int J Mol Sci, pp. 5519–5544. doi: 10.3390/ijms14035519.

Fujii, K. *et al.* (2017) 'Pervasive translational regulation of the cell signalling circuitry underlies mammalian development', *Nature Communications 2017 8:1*, 8(1), pp. 1–13. doi: 10.1038/ncomms14443.

Gabashvili, I. S. *et al.* (2000) 'Solution structure of the E. coli 70S ribosome at 11.5 A resolution', *Cell*, 100(5), pp. 537–549. doi: 10.1016/S0092-8674(00)80690-X.

Gagnon, M. *et al.* (2021) 'Potentiation of B2 receptor signaling by AltB2R, a newly identified alternative protein encoded in the human bradykinin B2 receptor gene', *Journal of Biological Chemistry*, 296, p. 100329. doi: 10.1016/J.JBC.2021.100329.

Gallie, D. R. (1991) 'The cap and poly(A) tail function synergistically to regulate mRNA translational efficiency', *Genes and Development*, 5(11), pp. 2108–2116. doi: 10.1101/gad.5.11.2108.

Garcia-Garijo, A., Fajardo, C. A. and Gros, A. (2019) 'Determinants for neoantigen identification', *Frontiers in Immunology*, 10(JUN), p. 1392. doi: 10.3389/FIMMU.2019.01392/BIBTEX.

Ge, Q. *et al.* (2021) 'Micropeptide ASAP encoded by LINC00467 promotes colorectal cancer progression by directly modulating ATP synthase activity', *The Journal of clinical investigation*, 131(22). doi: 10.1172/JCI152911.

Ge, Y. and Porse, B. T. (2014) 'The functional consequences of intron retention: Alternative splicing coupled to NMD as a regulator of gene expression', *BioEssays*, 36(3), pp. 236–243. doi: 10.1002/BIES.201300156.

Gebauer, F. *et al.* (1994) 'Translational control by cytoplasmic polyadenylation of c-mos mRNA is necessary for oocyte maturation in the mouse', *EMBO Journal*, 13(23), pp. 5712–5720. doi: 10.1002/j.1460-2075.1994.tb06909.x.

Grabherr, M. G. *et al.* (2011) 'Full-length transcriptome assembly from RNA-Seq data without a reference genome', *Nature Biotechnology 2011 29:7*, 29(7), pp. 644–652. doi: 10.1038/nbt.1883.

Grüning, B. *et al.* (2016) 'galaxytools: July 2016 release'. doi: 10.5281/ZENODO.58846.

Grzybowska, E. A., Wilczynska, A. and Siedlecki, J. A. (2001) 'Regulatory functions of 3'UTRs', *Biochemical and biophysical research communications*, 288(2), pp. 291–295. doi: 10.1006/BBRC.2001.5738.

Guhaniyogi, J. and Brewer, G. (2001) 'Regulation of mRNA stability in mammalian cells', *Gene*, 265(1–2), pp. 11–23. doi: 10.1016/S0378-1119(01)00350-X.

Guo, B. *et al.* (2003) 'Humanin peptide suppresses apoptosis by interfering with Bax activation', *Nature*, 423(6938), pp. 456–461. doi: 10.1038/NATURE01627.

Guo, B. *et al.* (2020) 'Micropeptide CIP2A-BP encoded by LINC00665 inhibits triple-negative breast cancer progression', *The EMBO journal*, 39(1). doi: 10.15252/EMBJ.2019102190.

Guo, J. U. and Bartel, D. P. (2016) 'RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria', *Science*, 353(6306). doi: 10.1126/science.aaf5371.

Guttman, M. *et al.* (2013) 'Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins', *Cell*, 154(1), pp. 240–251. doi: 10.1016/j.cell.2013.06.009.

Guydosh, N. R. and Green, R. (2014) 'Dom34 rescues ribosomes in 3′ untranslated regions', *Cell*, 156(5), pp. 950–962. doi: 10.1016/j.cell.2014.02.006.

Ha, M. and Kim, V. N. (2014) 'Regulation of microRNA biogenesis', *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, pp. 509–524. doi: 10.1038/nrm3838.

Haghighat, A. *et al.* (1995) 'Repression of cap-dependent translation by 4E-binding protein 1: competition with p220 for binding to eukaryotic initiation factor-4E', *EMBO J.*, 14(22), pp. 5701–5709. doi: 10.1002/j.1460-2075.1995.tb00257.x.

Hake, L. E. and Richter, J. D. (1994) 'CPEB is a specificity factor that mediates cytoplasmic polyadenylation during Xenopus oocyte maturation', *Cell*, 79(4), pp. 617–627. doi: 10.1016/0092-8674(94)90547-9.

Hann, S. R. *et al.* (1988) 'A non-AUG translational initiation in c-myc exon 1 generates an N-terminally distinct protein whose synthesis is disrupted in Burkitt's lymphomas', *Cell*, 52(2), pp. 185–195. doi: 10.1016/0092-8674(88)90507-7.

Hao, Y. *et al.* (2018) 'SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci', *Briefings in Bioinformatics*, 19(4), pp. 636–643. doi: 10.1093/BIB/BBX005.

Harding, H. P. *et al.* (2000) 'Regulated translation initiation controls stress-induced gene expression in mammalian cells', *Molecular Cell*, 6(5), pp. 1099–1108. doi: 10.1016/S1097-2765(00)00108-8.

Harrison, P. W. *et al.* (2024) 'Ensembl 2024', *Nucleic Acids Research*, 52(D1), pp. D891–D899. doi: 10.1093/NAR/GKAD1049.

Hartman, E. *et al.* (2024) 'Peptide clustering enhances large-scale analyses and reveals proteolytic signatures in mass spectrometry data', *Nature Communications 2024 15:1*, 15(1), pp. 1–15. doi: 10.1038/s41467-024-51589-y.

Harvey, R. F. *et al.* (2018) 'Trans-acting translational regulatory RNA binding proteins', *Wiley Interdisciplinary Reviews. RNA*, 9(3). doi: 10.1002/WRNA.1465.

Hashimoto, Y. *et al.* (2001) 'A rescue factor abolishing neuronal cell death by a wide spectrum of familial Alzheimer's disease genes and Abeta', *Proceedings of the National Academy of Sciences of the United States of America*, 98(11), pp. 6336–6341. doi: 10.1073/PNAS.101133498.

Hayes, J., Peruzzi, P. P. and Lawler, S. (2014) 'MicroRNAs in cancer: Biomarkers, functions and therapy', *Trends in Molecular Medicine*. Elsevier Ltd, pp. 460–469. doi: 10.1016/j.molmed.2014.06.005.

He, F. and Jacobson, A. (2015) 'Nonsense-Mediated mRNA Decay: Degradation of Defective Transcripts Is Only Part of the Story', *Annual review of genetics*, 49, p. 339. doi: 10.1146/ANNUREV-GENET-112414-054639.

He, L. *et al.* (2021) 'Circular RNAs' cap-independent translation protein and its roles in carcinomas', *Molecular cancer*, 20(1). doi: 10.1186/S12943-021-01417-4.

van Heesch, S. *et al.* (2019) 'The Translational Landscape of the Human Heart', *Cell*, 178(1), pp. 242-260.e29. doi: 10.1016/J.CELL.2019.05.010.

Hellen, C. U. T. (2018) 'Translation termination and ribosome recycling in eukaryotes', *Cold Spring Harbor Perspectives in Biology*, 10(10). doi: 10.1101/cshperspect.a032656.

Hershey, J. W. B., Sonenberg, N. and Mathews, M. B. (2019) 'Principles of translational control', *Cold Spring Harbor Perspectives in Biology*, 11(9), p. a032607. doi: 10.1101/cshperspect.a032607.

Heuer, A. *et al.* (2017) 'Structure of the 40S-ABCE1 post-splitting complex in ribosome recycling and translation initiation', *Nature structural & molecular biology*, 24(5), pp. 453–460. doi: 10.1038/NSMB.3396.

Hinnebusch, A. G. (1988) 'Mechanisms of gene regulation in the general control of amino acid biosynthesis in Saccharomyces cerevisiae', *Microbiological Reviews*.

American Society for Microbiology (ASM), pp. 248–273. doi: 10.1128/mmbr.52.2.248-273.1988.

Hinnebusch, A. G. (2005) ‘Translational regulation of GCN4 and the general amino acid control of yeast’, *Annual Review of Microbiology*. Annu Rev Microbiol, pp. 407–450. doi: 10.1146/annurev.micro.59.031805.133833.

Hinnebusch, A. G. and Lorsch, J. R. (2012) ‘The mechanism of eukaryotic translation initiation: New insights and challenges’, *Cold Spring Harbor Perspectives in Biology*, 4(10). doi: 10.1101/cshperspect.a011544.

Ho, J. *et al.* (2020) ‘Detection of CSF1 rearrangements deleting the 3’ UTR in tenosynovial giant cell tumors’, *Genes, chromosomes & cancer*, 59(2), pp. 96–105. doi: 10.1002/GCC.22807.

Ho, V. *et al.* (1995) ‘Use of a marked erythropoietin gene for investigation of its cis-acting elements’, *The Journal of biological chemistry*, 270(17), pp. 10084–10090. doi: 10.1074/JBC.270.17.10084.

Holbrook, J. A. *et al.* (2004) ‘Nonsense-mediated decay approaches the clinic’, *Nature Genetics*, 36(8), pp. 801–808. doi: 10.1038/NG1403.

Hong, D. and Jeong, S. (2023) ‘3’UTR Diversity: Expanding Repertoire of RNA Alterations in Human mRNAs’, *Molecules and Cells*, 46(1), pp. 48–56. doi: 10.14348/MOLCELLS.2023.0003.

Horie, C. *et al.* (2022) ‘Motile sperm domain containing 1 is upregulated by the Wnt/β-catenin signaling pathway in colorectal cancer’, *Oncology Letters*, 24(2). doi: 10.3892/OL.2022.13402.

Hoshino, S. I. *et al.* (1999) ‘The eukaryotic polypeptide chain releasing factor (eRF3/GSPT) carrying the translation termination signal to the 3’-Poly(A) tail of mRNA. Direct association of erf3/GSPT with polyadenylate-binding protein’, *The Journal of biological chemistry*, 274(24), pp. 16677–16680. doi: 10.1074/JBC.274.24.16677.

Howard, M. T. *et al.* (2000) ‘Sequence Specificity of Aminoglycoside-Induced Stop Codon Readthrough: Potential Implications for Treatment of Duchenne Muscular Dystrophy’, *Ann Neurol*, 48, pp. 164–169. doi: 10.1002/1531-8249(200008)48:2.

Hsu, P. Y. *et al.* (2016) ‘Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis’, *Proceedings of the National Academy of Sciences of the United States of America*, 113(45), pp. E7126–E7135. doi: 10.1073/pnas.1614788113.

Huang, J. Z. *et al.* (2017) ‘A Peptide Encoded by a Putative lncRNA HOXB-AS3 Suppresses Colon Cancer Growth’, *Molecular cell*, 68(1), pp. 171-184.e6. doi: 10.1016/J.MOLCEL.2017.09.015.

Huang, W. (2017) ‘MicroRNAs: Biomarkers, diagnostics, and therapeutics’, *Methods in Molecular Biology*, 1617, pp. 57–67. doi: 10.1007/978-1-4939-7046-9_4.

Huntzinger, E. and Izaurralde, E. (2011) ‘Gene silencing by microRNAs: Contributions of translational repression and mRNA decay’, *Nature Reviews*

Genetics. Nat Rev Genet, pp. 99–110. doi: 10.1038/nrg2936.

Iacono, M., Mignone, F. and Pesole, G. (2005) ‘uAUG and uORFs in human and rodent 5′untranslated mRNAs’, *Gene*, 349, pp. 97–105. doi: 10.1016/j.gene.2004.11.041.

Iadevaia, V. *et al.* (2012) ‘Roles of the mammalian target of rapamycin, mTOR, in controlling ribosome biogenesis and protein synthesis’, *Biochemical Society Transactions*. Biochem Soc Trans, pp. 168–172. doi: 10.1042/BST20110682.

Illingworth, R. S. and Bird, A. P. (2009) ‘CpG islands – “A rough guide”’, *FEBS Letters*, 583(11), pp. 1713–1720. doi: 10.1016/J.FEBSLET.2009.04.012.

Ingolia, N. T. *et al.* (2009) ‘Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling’, *Science*, 324(5924), pp. 218–223. doi: 10.1126/science.1168978.

Ingolia, N. T. *et al.* (2014) ‘Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes’, *Cell reports*, 8(5), pp. 1365–1379. doi: 10.1016/J.CELREP.2014.07.045.

Ingolia, N. T., Hussmann, J. A. and Weissman, J. S. (2019) ‘Ribosome profiling: Global views of translation’, *Cold Spring Harbor Perspectives in Biology*, 11(5), p. a032698. doi: 10.1101/cshperspect.a032698.

Ingolia, N. T., Lareau, L. F. and Weissman, J. S. (2011) ‘Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity of Mammalian Proteomes’, *Cell*, 147(4), p. 789. doi: 10.1016/J.CELL.2011.10.002.

Ipsaro, J. J. and Joshua-Tor, L. (2015) ‘From guide to target: Molecular insights into eukaryotic RNA-interference machinery’, *Nature Structural and Molecular Biology*. Nature Publishing Group, pp. 20–28. doi: 10.1038/nsmb.2931.

Ishikawa, Y. *et al.* (2008) ‘The rough endoplasmic reticulum-resident FK506-binding protein FKBP65 is a molecular chaperone that interacts with collagens’, *The Journal of biological chemistry*, 283(46), pp. 31584–31590. doi: 10.1074/JBC.M802535200.

Ivanov, A. *et al.* (2016) ‘PABP enhances release factor recruitment and stop codon recognition during translation termination’, *Nucleic Acids Research*, 44(16), pp. 7766–7776. doi: 10.1093/nar/gkw635.

Ivanov, P., Kedersha, N. and Anderson, P. (2019) ‘Stress granules and processing bodies in translational control’, *Cold Spring Harbor Perspectives in Biology*, 11(5), p. a032813. doi: 10.1101/cshperspect.a032813.

Ivanov, P. V. *et al.* (2008) ‘Interactions between UPF1, eRFs, PABP and the exon junction complex suggest an integrated model for mammalian NMD pathways’, *The EMBO Journal*, 27(5), p. 736. doi: 10.1038/EMBOJ.2008.17.

Ivshina, M., Lasko, P. and Richter, J. D. (2014) ‘Cytoplasmic polyadenylation element binding proteins in development, health, and disease’, *Annual review of cell and developmental biology*, 30, pp. 393–415. doi: 10.1146/ANNUREV-CELLBIO-101011-155831.

Jackson, R. J. (2013) ‘The Current Status of Vertebrate Cellular mRNA IRESs’, *Cold Spring Harbor Perspectives in Biology*, 5(2). doi: 10.1101/CSHPERSPECT.A011569.

Jackson, R. J., Hellen, C. U. T. and Pestova, T. V. (2010) ‘The mechanism of eukaryotic translation initiation and principles of its regulation’, *Nature Reviews Molecular Cell Biology*. Nat Rev Mol Cell Biol, pp. 113–127. doi: 10.1038/nrm2838.

Jackson, R. J., Hellen, C. U. T. and Pestova, T. V. (2012) ‘Termination and post-termination events in eukaryotic translation’, in *Advances in Protein Chemistry and Structural Biology*. Academic Press Inc., pp. 45–93. doi: 10.1016/B978-0-12-386497-0.00002-5.

Jacobs, G. H. *et al.* (2002) ‘Transterm: a database of mRNAs and translational control elements’, *Nucleic Acids Research*, 30(1), p. 310. doi: 10.1093/NAR/30.1.310.

Jaenisch, R. and Bird, A. (2003) ‘Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals’, *Nature Genetics*, 33(3S), pp. 245–254. doi: 10.1038/NG1089.

Jang, H. S. *et al.* (2017) ‘CpG and Non-CpG Methylation in Epigenetic Gene Regulation and Brain Function’, *Genes*, 8(6), pp. 2–20. doi: 10.3390/GENES8060148.

Ji, Z. *et al.* (2009) ‘Progressive lengthening of 3′ untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development’, *Proceedings of the National Academy of Sciences of the United States of America*, 106(17), p. 7028. doi: 10.1073/PNAS.0900028106.

Ji, Z. *et al.* (2015) ‘Many lncRNAs, 5′UTRs, and pseudogenes are translated and some are likely to express functional proteins’, *eLife*, 4(DECEMBER2015). doi: 10.7554/eLife.08890.

Jia, L. *et al.* (2020) ‘Decoding mRNA translatability and stability from the 5′ UTR’, *Nature Structural & Molecular Biology 2020 27:9*, 27(9), pp. 814–821. doi: 10.1038/s41594-020-0465-x.

Jiang, G. *et al.* (2024) ‘A comprehensive workflow for optimizing RNA-seq data analysis’, *BMC Genomics*, 25(1), pp. 1–21. doi: 10.1186/S12864-024-10414-Y/FIGURES/9.

Jin, X. *et al.* (2003) ‘The two upstream open reading frames of oncogene mdm2 have different translational regulatory properties’, *The Journal of biological chemistry*, 278(28), pp. 25716–25721. doi: 10.1074/JBC.M300316200.

Johnstone, T. G., Bazzini, A. A. and Giraldez, A. J. (2016) ‘ Upstream ORF s are prevalent translational repressors in vertebrates ’, *The EMBO Journal*, 35(7), pp. 706–723. doi: 10.15252/embj.201592759.

Jonas, S. and Izaurralde, E. (2015) ‘Towards a molecular understanding of microRNA-mediated gene silencing’, *Nature Reviews Genetics*. Nature Publishing Group, pp. 421–433. doi: 10.1038/nrg3965.

Joshi, B., Yan, R. and Rhoads, R. E. (1994) ‘In vitro synthesis of human protein synthesis initiation factor 4 gamma and its localization on 43 and 48S initiation complexes’, *J. Biol. Chem.*, 269(3), pp. 2048–2055. doi: 10.1016/s0021-9258(17)42133-8.

Jungreis, I. *et al.* (2011) ‘Evidence of abundant stop codon readthrough in Drosophila and other metazoa’, *Genome Research*, 21(12), p. 2096. doi: 10.1101/GR.119974.110.

Jungreis, I. *et al.* (2016) ‘Evolutionary dynamics of abundant stop codon readthrough’, *Molecular Biology and Evolution*, 33(12), pp. 3108–3132. doi: 10.1093/molbev/msw189.

Kahvejian, A. *et al.* (2005) ‘Mammalian poly(A)-binding protein is a eukaryotic translation initiation factor, which acts via multiple mechanisms’, *Genes and Development*, 19(1), pp. 104–113. doi: 10.1101/gad.1262905.

Karcher, A., Schele, A. and Hopfner, K. P. (2008) ‘X-ray structure of the complete ABC enzyme ABCE1 from Pyrococcus abyssi’, *The Journal of biological chemistry*, 283(12), pp. 7962–7971. doi: 10.1074/JBC.M707347200.

Karlin, S. and Mrázek, J. (1997) ‘Compositional differences within and between eukaryotic genomes’, *Proceedings of the National Academy of Sciences of the United States of America*, 94(19), p. 10227. doi: 10.1073/PNAS.94.19.10227.

Karousis, E. D. and Mühlemann, O. (2019) ‘Nonsense-mediated mRNA decay begins where translation ends’, *Cold Spring Harbor Perspectives in Biology*, 11(2), p. a032862. doi: 10.1101/cshperspect.a032862.

Kataoka, K. *et al.* (2016) ‘Aberrant PD-L1 expression through 3’-UTR disruption in multiple cancers’, *Nature*, 534(7607), pp. 402–406. doi: 10.1038/NATURE18294.

Kawamata, T. and Tomari, Y. (2010) ‘Making RISC’, *Trends in Biochemical Sciences*. Trends Biochem Sci, pp. 368–376. doi: 10.1016/j.tibs.2010.03.009.

Kearse, M. G. and Wilusz, J. E. (2017) ‘Non-AUG translation: a new start for protein synthesis in eukaryotes’, *Genes & Development*, 31(17), pp. 1717–1731. doi: 10.1101/GAD.305250.117.

Keeling, K. M. *et al.* (2004) ‘Leaky termination at premature stop codons antagonizes nonsense-mediated mRNA decay in S. cerevisiae’, *RNA*, 10(4), pp. 691–703. doi: 10.1261/rna.5147804.

Kent, W. J. (2002) ‘BLAT—The BLAST-Like Alignment Tool’, *Genome Research*, 12(4), p. 656. doi: 10.1101/GR.229202.

Kervestin, S. and Jacobson, A. (2012) ‘NMD: A multifaceted response to premature translational termination’, *Nature Reviews Molecular Cell Biology*, 13(11), pp. 700–712. doi: 10.1038/NRM3454.

Kessler, S. H. and Sachs, A. B. (1998) ‘RNA Recognition Motif 2 of Yeast Pab1p Is Required for Its Functional Interaction with Eukaryotic Translation Initiation Factor 4G’, *Molecular and Cellular Biology*, 18(1), pp. 51–57. doi: 10.1128/mcb.18.1.51.

Khajavi, M., Inoue, K. and Lupski, J. R. (2006) ‘Nonsense-mediated mRNA decay modulates clinical outcome of genetic disease', *European Journal of Human Genetics*, 14(10), pp. 1074–1081. doi: 10.1038/SJ.EJHG.5201649.

Kim, D. *et al.* (2019) 'Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype', *Nature Biotechnology*, 37(8), pp. 907–915. doi: 10.1038/s41587-019-0201-4.

Kluge, F., Götze, M. and Wahle, E. (2020) 'Establishment of 5′-3′ interactions in mRNA independent of a continuous ribose-phosphate backbone', *RNA*, 26(5), pp. 613–628. doi: 10.1261/rna.073759.119.

Knee, R. S., Pitcher, S. E. and Murphy, P. R. (1994) 'Basic fibroblast growth factor sense (FGF) and antisense (gfg) RNA transcripts are expressed in unfertilized human oocytes and in differentiated adult tissues', *Biochemical and biophysical research communications*, 205(1), pp. 577–583. doi: 10.1006/BBRC.1994.2704.

Kopczynski, J. B., Raff, A. C. and Bonner, J. J. (1992) 'Translational readthrough at nonsense mutations in the HSF1 gene of Saccharomyces cerevisme', *MGG Molecular & General Genetics*, 234(3), pp. 369–378. doi: 10.1007/BF00538696.

Kozak, M. (1984) 'Selection of initiation sites by eucaryotic ribosomes: effect of inserting AUG triplets upstream from the coding sequence for preproinsulin.', *Nucleic Acids Research*, 12(9), p. 3873. doi: 10.1093/NAR/12.9.3873.

Kozak, M. (1987) 'An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs.', *Nucleic acids research*, 15(20), pp. 8125–48. doi: 10.1093/nar/15.20.8125.

Kozak, M. (2001) 'Constraints on reinitiation of translation in mammals', *Nucleic Acids Research*, 29(24), p. 5226. doi: 10.1093/NAR/29.24.5226.

Kozak, M. (2002) 'Pushing the limits of the scanning mechanism for initiation of translation', *Gene*, 299(1), p. 1. doi: 10.1016/S0378-1119(02)01056-9.

Krueger, F. (2021) *Trim Galore!: A Wrapper Tool Around Cutadapt and FastQC to Consistently Apply Quality and Adapter Trimming to FastQ Files*. Available at: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ (Accessed: 27 June 2024).

Kryuchkova, P. *et al.* (2013) 'Two-step model of stop codon recognition by eukaryotic release factor eRF1', *Nucleic Acids Research*, 41(8), p. 4573. doi: 10.1093/NAR/GKT113.

Kuersten, S. and Goodwin, E. B. (2003) 'The power of the 3' UTR: translational control and development', *Nature reviews. Genetics*, 4(8), pp. 626–637. doi: 10.1038/NRG1125.

Kumar, S. and Sharawat, S. K. (2018) 'Epigenetic regulators of programmed death-ligand 1 expression in human cancers', *Translational research : the journal of laboratory and clinical medicine*, 202, pp. 129–145. doi: 10.1016/J.TRSL.2018.05.011.

Kuroda, A. (2009) 'Insulin gene expression is regulated by DNA methylation', *PLoS ONE*, 4(9), p. e6953. doi: 10.1371/journal.pone.00069534.

Lai, W. J. C. *et al.* (2018) 'mRNAs and lncRNAs intrinsically form secondary structures with short end-to-end distances', *Nature Communications*, 9(1), pp. 1–11. doi: 10.1038/s41467-018-06792-z.

Larizza, A. *et al.* (2002) 'Evolutionary dynamics of mammalian mRNA untranslated regions by comparative analysis of orthologous human, artiodactyl and rodent gene pairs', *Computers & Chemistry*, 26(5), pp. 479–490. doi: 10.1016/S0097-8485(02)00009-8.

Laumont, C. M. *et al.* (2016) 'Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames', *Nature Communications 2016 7:1*, 7(1), pp. 1–12. doi: 10.1038/ncomms10238.

Law, J. A. and Jacobsen, S. E. (2010) 'Establishing, maintaining and modifying DNA methylation patterns in plants and animals', *Nature Reviews Genetics*, 11(3), pp. 204–220. doi: 10.1038/NRG2719.

Leblanc, S. *et al.* (2024) 'OpenProt 2.0 builds a path to the functional characterization of alternative proteins', *Nucleic Acids Research*, 52(D1), p. D522. doi: 10.1093/NAR/GKAD1050.

Lee, C. *et al.* (2015) 'The Mitochondrial-Derived Peptide MOTS-c Promotes Metabolic Homeostasis and Reduces Obesity and Insulin Resistance', *Cell Metabolism*, 21(3), pp. 443–454. doi: 10.1016/J.CMET.2015.02.009.

Lee, C. Q. E. *et al.* (2021) 'Coding and non-coding roles of MOCCI (C15ORF48) coordinate to regulate host inflammation and immunity', *Nature Communications 2021 12:1*, 12(1), pp. 1–22. doi: 10.1038/s41467-021-22397-5.

Lee, J. *et al.* (2015) 'A novel troponin T peptide in humans: Assay, biochemistry and preliminary findings in acute coronary syndromes', *International Journal of Cardiology*, 190(1), pp. 68–74. doi: 10.1016/j.ijcard.2015.04.145.

Lee, Sooncheol *et al.* (2012) 'Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution', *Proceedings of the National Academy of Sciences of the United States of America*, 109(37), pp. E2424–E2432. doi: 10.1073/PNAS.1207846109/-/DCSUPPLEMENTAL.

Lei, M. *et al.* (2020) 'Translation and functional roles of circular RNAs in human cancer', *Molecular cancer*, 19(1). doi: 10.1186/S12943-020-1135-7.

Leinonen, R., Sugawara, H. and Shumway, M. (2011) 'The sequence read archive', *Nucleic Acids Research*, 39(SUPPL. 1), pp. D19–D21. doi: 10.1093/nar/gkq1019.

Leong, A. Z. X. *et al.* (2022) 'Short open reading frames (sORFs) and microproteins: an update on their identification and validation measures', *Journal of Biomedical Science 2022 29:1*, 29(1), pp. 1–15. doi: 10.1186/S12929-022-00802-5.

Lewis, B. P., Green, R. E. and Brenner, S. E. (2003) 'Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans', *Proceedings of the National Academy of Sciences of the United States of America*, 100(1), pp. 189–192. doi: 10.1073/PNAS.0136770100.

Li, C. *et al.* (2021) 'PTPN18 promotes colorectal cancer progression by regulating the c-MYC-CDK4 axis', *Genes & Diseases*, 8(6), p. 838. doi:

10.1016/J.GENDIS.2020.08.001.

Li, H. *et al.* (2009) ‘The Sequence Alignment/Map format and SAMtools’, *Bioinformatics*, 25(16), p. 2078. doi: 10.1093/BIOINFORMATICS/BTP352.

Li, Y. *et al.* (2020) ‘Transcription levels and prognostic significance of the NFI family members in human cancers’, *PeerJ*, 2020(3). doi: 10.7717/PEERJ.8816/SUPP-25.

Li, Y. *et al.* (2021) ‘SmProt: A Reliable Repository with Comprehensive Annotation of Small Proteins Identified from Ribosome Profiling’, *Genomics, Proteomics & Bioinformatics*, 19(4), p. 602. doi: 10.1016/J.GPB.2021.09.002.

Liang, W. C. *et al.* (2019) ‘Translation of the circular RNA circβ-catenin promotes liver cancer cell growth through activation of the Wnt pathway’, *Genome biology*, 20(1). doi: 10.1186/S13059-019-1685-4.

Liao, C. *et al.* (2022) ‘SPINKs in Tumors: Potential Therapeutic Targets’, *Frontiers in Oncology*, 12. doi: 10.3389/FONC.2022.833741.

Lin, M. si *et al.* (2022) ‘Pan-cancer analysis of oncogenic TNFAIP2 identifying its prognostic value and immunological function in acute myeloid leukemia’, *BMC Cancer*, 22(1), p. 1068. doi: 10.1186/S12885-022-10155-9.

Lingelbach, K. and Dobberstein, B. (1988) ‘An extended RNA/RNA duplex structure within the coding region of mRNA does not block translational elongation’, *Nucleic Acids Research*, 16(8), pp. 3405–3414. doi: 10.1093/nar/16.8.3405.

Lipman, D. J. (1997) ‘Making (anti)sense of non-coding sequence conservation’, *Nucleic acids research*, 25(18), pp. 3580–3583. doi: 10.1093/NAR/25.18.3580.

Lister, R. *et al.* (2009) ‘Human DNA methylomes at base resolution show widespread epigenomic differences’, *Nature*, 462(7271), pp. 315–322. doi: 10.1038/NATURE08514.

Litterman, A. J. *et al.* (2019) ‘A massively parallel 3′ UTR reporter assay reveals relationships between nucleotide content, sequence conservation, and mRNA destabilization’, *Genome Research*, 29(6), pp. 896–906. doi: 10.1101/GR.242552.118/-/DC1.

Liu, Q. *et al.* (2020) ‘RiboToolkit: an integrated platform for analysis and annotation of ribosome profiling data to decode mRNA translation at codon resolution’, *Nucleic Acids Research*, 48(W1), pp. W218–W229. doi: 10.1093/NAR/GKAA395.

López De Silanes, I., Paz Quesada, M. and Esteller, M. (2007) ‘Aberrant Regulation of Messenger RNA 3′-Untranslated Region in Human Cancer’, *Cellular Oncology : the Official Journal of the International Society for Cellular Oncology*, 29(1), p. 1. doi: 10.1155/2007/586139.

Loughran, G. *et al.* (2014) ‘Evidence of efficient stop codon readthrough in four mammalian genes’, *Nucleic Acids Research*, 42(14), p. 8928. doi: 10.1093/NAR/GKU608.

Lu, J. *et al.* (2005) ‘MicroRNA expression profiles classify human cancers’, *Nature*, 435(7043), pp. 834–838. doi: 10.1038/NATURE03702.

Lu, S. *et al.* (2019) ‘A hidden human proteome encoded by “non-coding” genes’, *Nucleic Acids Research*, 47(15), pp. 8111–8125. doi: 10.1093/NAR/GKZ646.

Lucas, C. L. *et al.* (2016) ‘PI3Kδ and primary immunodeficiencies’, *Nature Reviews Immunology*. Nature Publishing Group, pp. 702–714. doi: 10.1038/nri.2016.93.

Luukkonen, B. G. M., Tan, W. and Schwartz, S. (1995) ‘Efficiency of reinitiation of translation on human immunodeficiency virus type 1 mRNAs is determined by the length of the upstream open reading frame and by intercistronic distance.’, *Journal of Virology*, 69(7), p. 4086. doi: 10.1128/jvi.69.7.4086-4094.1995.

Ma, J. *et al.* (2016) ‘Improved Identification and Analysis of Small Open Reading Frame Encoded Polypeptides’, *Analytical Chemistry*, 88(7), pp. 3967–3975. doi: 10.1021/acs.analchem.6b00191.

Mackowiak, S. D. *et al.* (2015) ‘Extensive identification and analysis of conserved small ORFs in animals’, *Genome Biology*, 16(1). doi: 10.1186/s13059-015-0742-x.

Mader, S. *et al.* (1995) ‘The translation initiation factor eIF-4E binds to a common motif shared by the translation factor eIF-4 gamma and the translational repressors 4E-binding proteins’, *Mol. Cell Biol.*, 15(9), pp. 4990–4997. doi: 10.1128/mcb.15.9.4990.

Makarewich, C. A. *et al.* (2018) ‘MOXI Is a Mitochondrial Micropeptide That Enhances Fatty Acid β-Oxidation’, *Cell reports*, 23(13), pp. 3701–3709. doi: 10.1016/J.CELREP.2018.05.058.

Makarewich, C. A. and Olson, E. N. (2017) ‘Mining for Micropeptides’, *Trends in Cell Biology*. Elsevier Ltd, pp. 685–696. doi: 10.1016/j.tcb.2017.04.006.

Mangkalaphiban, K. *et al.* (2021) ‘Transcriptome-wide investigation of stop codon readthrough in Saccharomyces cerevisiae’, *PLOS Genetics*. Edited by A. K. Hopper, 17(4), p. e1009538. doi: 10.1371/journal.pgen.1009538.

Mantsyzov, A. B. *et al.* (2010) ‘NMR solution structure and function of the C-terminal domain of eukaryotic class 1 polypeptide chain release factor’, *The Febs Journal*, 277(12), p. 2611. doi: 10.1111/J.1742-4658.2010.07672.X.

Manuvakhova, M., Keeling, K. and Bedwell, D. M. (2000) ‘Aminoglycoside antibiotics mediate context-dependent suppression of termination codons in a mammalian translation system.’, *RNA*, 6(7), p. 1044. doi: 10.1017/S1355838200000716.

Martinez, T. F. *et al.* (2019) ‘Accurate annotation of human protein-coding small open reading frames’, *Nature Chemical Biology 2019 16:4*, 16(4), pp. 458–468. doi: 10.1038/s41589-019-0425-0.

Matsumoto, A. *et al.* (2017) ‘mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide’, *Nature*, 541(7636), pp. 228–232. doi: 10.1038/NATURE21034.

Mayr, C. and Bartel, D. P. (2009) ‘Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells’, *Cell*, 138(4), p. 673. doi: 10.1016/J.CELL.2009.06.016.

Mayr, C. (2016) ‘Evolution and Biological Roles of Alternative 3’UTRs’, *Trends in Cell Biology*, 26(3), pp. 227–237. doi: 10.1016/j.tcb.2015.10.012.

Mccaughan, K. K. *et al.* (1995) ‘Translational termination efficiency in mammals is influenced by the base following the stop codon.’, *Proceedings of the National Academy of Sciences of the United States of America*, 92(12), p. 5431. doi: 10.1073/PNAS.92.12.5431.

McGlincy, N. J. and Smith, C. W. J. (2008) ‘Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense?’, *Trends in Biochemical Sciences*, 33(8), pp. 385–393. doi: 10.1016/J.TIBS.2008.06.001.

McKinney, W. (2011) ‘pandas: a Foundational Python Library for Data Analysis and Statistics’. Available at: https://docslib.org/doc/4231522/a-foundational-python-library-for-data-analysis-and-statistics (Accessed: 27 January 2025)

Merino-Valverde, I., Greco, E. and Abad, M. (2020) ‘The microproteome of cancer: From invisibility to relevance’, *Experimental Cell Research*, 392(1), p. 111997. doi: 10.1016/J.YEXCR.2020.111997.

Merrick, W. C. and Pavitt, G. D. (2018) ‘Protein synthesis initiation in eukaryotic cells’, *Cold Spring Harbor Perspectives in Biology*, 10(12), p. a033092. doi: 10.1101/cshperspect.a033092.

Meyer, K. D. *et al.* (2015) ‘5′ UTR m6A Promotes Cap-Independent Translation’, *Cell*, 163(4), pp. 999–1010. doi: 10.1016/j.cell.2015.10.012.

Michelangeli, F. and East, J. M. (2011) ‘A diversity of SERCA Ca2+ pump inhibitors’, *Biochemical Society transactions*, 39(3), pp. 789–797. doi: 10.1042/BST0390789.

Miettinen, T. P. and Björklund, M. (2015) ‘Modified ribosome profiling reveals high abundance of ribosome protected mRNA fragments derived from 3′ untranslated regions’, *Nucleic Acids Research*, 43(2), pp. 1019–1034. doi: 10.1093/nar/gku1310.

Mignone, F. *et al.* (2002) ‘Untranslated regions of mRNAs’, *Genome Biology*, 3(3), p. reviews0004.1. doi: 10.1186/GB-2002-3-3-REVIEWS0004.

Mills, E. W. *et al.* (2016) ‘Dynamic regulation of a ribosome rescue pathway in erythroid cells and platelets’, *Cell reports*, 17(1), p. 1. doi: 10.1016/J.CELREP.2016.08.088.

Mitchell, S. F. and Parker, R. (2014) ‘Principles and Properties of Eukaryotic mRNPs’, *Molecular Cell*. Cell Press, pp. 547–558. doi: 10.1016/j.molcel.2014.04.033.

Mitschka, S., Fansler, M. M. and Mayr, C. (2021) ‘Generation of 3′UTR knockout cell lines by CRISPR/Cas9-mediated genome editing’, *Methods in Enzymology*, 655, pp. 427–457. doi: 10.1016/BS.MIE.2021.03.014.

Morisaki, T. *et al.* (2016) ‘Real-time quantification of single RNA translation dynamics in living cells’, *Science*, 352(6292), pp. 1425–1429. doi: 10.1126/science.aaf0899.

Morita, M. *et al.* (2013) ‘MTORC1 controls mitochondrial activity and biogenesis through 4E-BP-dependent translational regulation', *Cell Metabolism*, 18(5), pp. 698–711. doi: 10.1016/j.cmet.2013.10.001.

Mueller, P. P. and Hinnebusch, A. G. (1986) 'Multiple upstream AUG codons mediate translational control of GCN4', *Cell*, 45(2), pp. 201–207. doi: 10.1016/0092-8674(86)90384-3.

Muhlrad, D., Decker, C. J. and Parker, R. (1994) 'Deadenylation of the unstable mRNA encoded by the yeast MFA2 gene leads to decapping followed by 5' → 3' digestion of the transcript', *Genes and Development*, 8(7), pp. 855–866. doi: 10.1101/gad.8.7.855.

Muhs, M. *et al.* (2015) 'Cryo-EM of ribosomal 80S complexes with termination factors reveal the translocated cricket paralysis virus IRES', *Molecular cell*, 57(3), p. 422. doi: 10.1016/J.MOLCEL.2014.12.016.

Na, C. H. *et al.* (2018) 'Discovery of noncanonical translation initiation sites through mass spectrometric analysis of protein N termini', *Genome Research*, 28(1), pp. 25–36. doi: 10.1101/gr.226050.117.

Namy, O., Duchateau-Nguyen, G. and Rousset, J. P. (2002) 'Translational readthrough of the PDE2 stop codon modulates cAMP levels in Saccharomyces cerevisiae', *Molecular Microbiology*, 43(3), pp. 641–652. doi: 10.1046/j.1365-2958.2002.02770.x.

Namy, O. and Rousset, J.-P. (2010) 'Specification of Standard Amino Acids by Stop Codons', in. Springer, New York, NY, pp. 79–100. doi: 10.1007/978-0-387-89382-2_4.

Neafsey, D. E. and Galagan, J. E. (2007) 'Dual modes of natural selection on upstream open reading frames', *Molecular Biology and Evolution*, 24(8), pp. 1744–1751. doi: 10.1093/molbev/msm093.

Nicholson, B. L. and White, K. A. (2011) '3′ Cap-independent translation enhancers of positive-strand RNA plant viruses', *Current Opinion in Virology*. Elsevier B.V., pp. 373–380. doi: 10.1016/j.coviro.2011.10.002.

Nie, K. *et al.* (2020) 'COX6B2 drives metabolic reprogramming toward oxidative phosphorylation to promote metastasis in pancreatic ductal cancer cells', *Oncogenesis*, 9(5). doi: 10.1038/S41389-020-0231-2.

Niu, L. *et al.* (2020) 'A micropeptide encoded by lncRNA MIR155HG suppresses autoimmune inflammation via modulating antigen presentation', *Science Advances*, 6(21). doi: 10.1126/SCIADV.AAZ2059/SUPPL_FILE/AAZ2059_SM.PDF.

Nobuta, R. *et al.* (2020) 'eIF4G-driven translation initiation of downstream ORFs in mammalian cells', *Nucleic Acids Research*, 48(18), p. 10441. doi: 10.1093/NAR/GKAA728.

O'Brien, J. *et al.* (2018) 'Overview of microRNA biogenesis, mechanisms of actions, and circulation', *Frontiers in Endocrinology*. Frontiers Media S.A., p. 402. doi: 10.3389/fendo.2018.00402.

O'Leary, N. A. *et al.* (2016) 'Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation', *Nucleic Acids*

Research, 44(Database issue), p. D733. doi: 10.1093/NAR/GKV1189.

Orr, M. W. *et al.* (2021) ‘Alternative ORFs and small ORFs: Shedding light on the dark proteome’, *Nucleic Acids Research*, 48(3), pp. 1029–1042. doi: 10.1093/NAR/GKZ734.

Ouspenskaia, T. *et al.* (2020) ‘Thousands of novel unannotated proteins expand the MHC I immunopeptidome in cancer’, *bioRxiv*, p. 2020.02.12.945840. doi: 10.1101/2020.02.12.945840.

Paek, K. Y. *et al.* (2015) ‘Translation initiation mediated by RNA looping’, *Proceedings of the National Academy of Sciences of the United States of America*, 112(4), pp. 1041–1046. doi: 10.1073/pnas.1416883112.

Palmiter, R. D. (1975) ‘Quantitation of parameters that determine the rate of ovalbumin synthesis’, *Cell*. Elsevier, pp. 189–197. doi: 10.1016/0092-8674(75)90167-1.

Park, E. H. *et al.* (2011) ‘Multiple elements in the eIF4G1 N-terminus promote assembly of eIF4G1•PABP mRNPs in vivo’, *EMBO Journal*, 30(2), pp. 302–316. doi: 10.1038/emboj.2010.312.

Paul, P. *et al.* (2018) ‘Interplay between miRNAs and human diseases’, *Journal of Cellular Physiology*. Wiley-Liss Inc., pp. 2007–2018. doi: 10.1002/jcp.25854.

Pauli, A. *et al.* (2014) ‘Toddler: an embryonic signal that promotes cell movement via Apelin receptors’, *Science (New York, N.Y.)*, 343(6172). doi: 10.1126/SCIENCE.1248636.

Peer, E. *et al.* (2019) ‘The epitranscriptome in translation regulation’, *Cold Spring Harbor Perspectives in Biology*, 11(8), p. a032623. doi: 10.1101/cshperspect.a032623.

Pelechano, V., Wei, W. and Steinmetz, L. M. (2013) ‘Extensive transcriptional heterogeneity revealed by isoform profiling’, *Nature*, 497(7447), pp. 127–131. doi: 10.1038/nature12121.

Pelletier, J. and Sonenberg, N. (2019) ‘The Organizing Principles of Eukaryotic Ribosome Recruitment’, *Annual Review of Biochemistry*. Annual Reviews Inc., pp. 307–335. doi: 10.1146/annurev-biochem-013118-111042.

Pesole, G. *et al.* (1997) ‘Structural and compositional features of untranslated regions of eukaryotic mRNAs’, *Gene*, 205(1–2), pp. 95–102. doi: 10.1016/S0378-1119(97)00407-1.

Pesole, G. *et al.* (2001) ‘Structural and functional features of eukaryotic mRNA untranslated regions’, *Gene*, 276(1–2), pp. 73–81. doi: 10.1016/S0378-1119(01)00674-6.

Pesole, G. *et al.* (2002) ‘UTRdb and UTRsite: Specialized databases of sequences and functional elements of 5′ and 3′ untranslated regions of eukaryotic mRNAs. Update 2002’, *Nucleic Acids Research*, 30(1), pp. 335–340. doi: 10.1093/NAR/30.1.335,.

Pesole, G., Bernardi, G. and Saccone, C. (1999) ‘Isochore specificity of AUG initiator context of human genes', *FEBS letters*, 464(1–2), pp. 60–62. doi: 10.1016/S0014-5793(99)01675-0.

Peterková, L. *et al.* (2020) 'Sarco/Endoplasmic Reticulum Calcium ATPase Inhibitors: Beyond Anticancer Perspective', *Journal of medicinal chemistry*, 63(5), pp. 1937–1963. doi: 10.1021/ACS.JMEDCHEM.9B01509.

Piccirillo, C. A. *et al.* (2014) 'Translational control of immune responses: From transcripts to translatomes', *Nature Immunology*. Nature Publishing Group, pp. 503–511. doi: 10.1038/ni.2891.

Pisarev, A. V. *et al.* (2010) 'The role of ABCE1 in eukaryotic post-termination ribosomal recycling', *Molecular cell*, 37(2), p. 196. doi: 10.1016/J.MOLCEL.2009.12.034.

Plass, M., Rasmussen, S. H. and Krogh, A. (2017) 'Highly accessible AU-rich regions in 3' untranslated regions are hotspots for binding of regulatory factors', *PLoS Computational Biology*, 13(4). doi: 10.1371/JOURNAL.PCBI.1005460.

Polycarpou-Schwarz, M. *et al.* (2018) 'The cancer-associated microprotein CASIMO1 controls cell proliferation and interacts with squalene epoxidase modulating lipid droplet formation', *Oncogene*, 37(34), pp. 4750–4768. doi: 10.1038/S41388-018-0281-5.

Portela, A. and Esteller, M. (2010) 'Epigenetic modifications and human disease', *Nature Biotechnology 2010 28:10*, 28(10), pp. 1057–1068. doi: 10.1038/nbt.1685.

Pöyry, T. A. A., Kaminski, A. and Jackson, R. J. (2004) 'What determines whether mammalian ribosomes resume scanning after translation of a short upstream open reading frame?', *Genes & Development*, 18(1), p. 62. doi: 10.1101/GAD.276504.

Preis, A. *et al.* (2014) 'Cryoelectron Microscopic Structures of Eukaryotic Translation Termination Complexes Containing eRF1-eRF3 or eRF1-ABCE1', *Cell reports*, 8(1), p. 59. doi: 10.1016/J.CELREP.2014.04.058.

Prensner, J. R. *et al.* (2021) 'Noncanonical open reading frames encode functional proteins essential for cancer cell survival', *Nature Biotechnology 2021 39:6*, 39(6), pp. 697–704. doi: 10.1038/s41587-020-00806-2.

Proud, C. G. (2019) 'Phosphorylation and signal transduction pathways in translational control', *Cold Spring Harbor Perspectives in Biology*, 11(7), p. a033050. doi: 10.1101/cshperspect.a033050.

Proudfoot, N. J. and Brownlee, G. G. (1976) '3′ Non-coding region sequences in eukaryotic messenger RNA', *Nature*, 263(5574), pp. 211–214. doi: 10.1038/263211A0,.

Pueyo, J. I. *et al.* (2016) 'Hemotin, a Regulator of Phagocytosis Encoded by a Small ORF and Conserved across Metazoans', *PLoS biology*, 14(3). doi: 10.1371/JOURNAL.PBIO.1002395.

Qu, X. *et al.* (2011) 'The ribosome uses two active mechanisms to unwind messenger RNA during translation', *Nature*, 475(7354), pp. 118–121. doi: 10.1038/nature10126.

Qureshi, N. S. and Duss, O. (2024) ‘Tracking transcription–translation coupling in real time’, *Nature 2024*, pp. 1–9. doi: 10.1038/s41586-024-08308-w.

Rabani, M. *et al.* (2017) ‘A Massively Parallel Reporter Assay of 3′ UTR Sequences Identifies In Vivo Rules for mRNA Degradation’, *Molecular Cell*, 68(6), pp. 1083-1094.e5. doi: 10.1016/J.MOLCEL.2017.11.014.

Ramadori, G. *et al.* (2020) ‘FKBP10 Regulates Protein Translation to Sustain Lung Cancer Growth’, *Cell Reports*, 30(11), pp. 3851-3863.e6. doi: 10.1016/j.celrep.2020.02.082.

Raudvere, U. *et al.* (2019) ‘g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)’, *Nucleic Acids Research*, 47(W1), p. W191. doi: 10.1093/NAR/GKZ369.

Rees, D. C., Johnson, E. and Lewinson, O. (2009) ‘ABC transporters: The power to change’, *Nature reviews. Molecular cell biology*, 10(3), p. 218. doi: 10.1038/NRM2646.

Renz, P. F., Valdivia Francia, F. and Sendoel, A. (2020) ‘Some like it translated: small ORFs in the 5′UTR’, *Experimental Cell Research*. Elsevier Inc., p. 112229. doi: 10.1016/j.yexcr.2020.112229.

Rice, P, Longden, I. and Bleasby, A. (2000) ‘EMBOSS: the European Molecular Biology Open Software Suite.’, *Trends in genetics : TIG*, 16(6), pp. 276–7. doi: 10.1016/s0168-9525(00)02024-2.

Robichaud, N. *et al.* (2019) ‘Translational control in cancer’, *Cold Spring Harbor Perspectives in Biology*, 11(7), p. a032896. doi: 10.1101/cshperspect.a032896.

Robinson, J. T. *et al.* (2011) ‘Integrative Genomics Viewer’, *Nature biotechnology*, 29(1), p. 24. doi: 10.1038/NBT.1754.

Rodnina, M. V. *et al.* (2020) ‘Survey and summary: Translational recoding: Canonical translation mechanisms reinterpreted’, *Nucleic Acids Research*, 48(3), pp. 1056–1067. doi: 10.1093/NAR/GKZ783.

Rodriguez, C. M. *et al.* (2019) ‘Translation of upstream open reading frames in a model of neuronal differentiation’, *BMC Genomics*, 20(1), pp. 1–18. doi: 10.1186/S12864-019-5775-1/FIGURES/7.

Rucci, E. *et al.* (2018) ‘SWIFOLD: Smith-Waterman implementation on FPGA with OpenCL for long DNA sequences’, *BMC Systems Biology*, 12(5), pp. 43–53. doi: 10.1186/S12918-018-0614-6/TABLES/7.

Ruiz-Orera, J. *et al.* (2018) ‘Translation of neutrally evolving peptides provides a basis for de novo gene evolution’, *Nature Ecology and Evolution*, 2(5), pp. 890–896. doi: 10.1038/s41559-018-0506-6.

Ruiz-Orera, J. and Albà, M. M. (2019) ‘Translation of Small Open Reading Frames: Roles in Regulation and Evolutionary Innovation’, *Trends in Genetics*. Elsevier Ltd, pp. 186–198. doi: 10.1016/j.tig.2018.12.003.

Ruiz Cuevas, M. V. *et al.* (2021) ‘Most non-canonical proteins uniquely populate the proteome or immunopeptidome’, *Cell Reports*, 34(10), p. 108815. doi:

10.1016/J.CELREP.2021.108815.

Samandi, S. *et al.* (2017) ‘Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins’, *eLife*, 6. doi: 10.7554/ELIFE.27860.

Sandberg, R. *et al.* (2008) ‘Proliferating cells express mRNAs with shortened 3′ UTRs and fewer microRNA target sites’, *Science (New York, N.Y.)*, 320(5883), p. 1643. doi: 10.1126/SCIENCE.1155390.

Sayers, E. W. *et al.* (2022) ‘Database resources of the National Center for BiotechnologyInformation’, *Nucleic Acids Research*, 50(D1), p. D20. doi: 10.1093/NAR/GKAB1112.

Schleich, S. *et al.* (2014) ‘DENR-MCT-1 promotes translation re-initiation downstream of uORFs to control tissue growth’, *Nature*, 512(7513), pp. 208–212. doi: 10.1038/nature13401.

Schleich, S. *et al.* (2017) ‘Identification of transcripts with short stuORFs as targets for DENR•MCTS1-dependent translation in human cells’, *Scientific Reports*, 7(1). doi: 10.1038/S41598-017-03949-6.

Schlesinger, D. and Elsässer, S. J. (2022) ‘Revisiting sORFs: overcoming challenges to identify and characterize functional microproteins’, *The FEBS Journal*, 289(1), pp. 53–74. doi: 10.1111/FEBS.15769.

Schneider-Poetsch, T. *et al.* (2010) ‘Inhibition of Eukaryotic Translation Elongation by Cycloheximide and Lactimidomycin’, *Nature chemical biology*, 6(3), p. 209. doi: 10.1038/NCHEMBIO.304.

Schott, J. *et al.* (2021) ‘Nascent Ribo-Seq measures ribosomal loading time and reveals kinetic impact on ribosome density’, *Nature Methods 2021 18:9*, 18(9), pp. 1068–1074. doi: 10.1038/s41592-021-01250-z.

Schwab, S. R. *et al.* (2003) ‘Constitutive display of cryptic translation products by MHC class I molecules’, *Science (New York, N.Y.)*, 301(5638), pp. 1367–1371. doi: 10.1126/SCIENCE.1085650.

Schwartz, D. C. and Parker, R. (2000) ‘mRNA Decapping in Yeast Requires Dissociation of the Cap Binding Protein, Eukaryotic Translation Initiation Factor 4E’, *Molecular and Cellular Biology*, 20(21), pp. 7933–7942. doi: 10.1128/mcb.20.21.7933-7942.2000.

Schweingruber, C. *et al.* (2013) ‘Nonsense-mediated mRNA decay - Mechanisms of substrate mRNA recognition and degradation in mammalian cells’, *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*. Biochim Biophys Acta, pp. 612–623. doi: 10.1016/j.bbagrm.2013.02.005.

Sendoel, A. *et al.* (2017) ‘Translation from unconventional 5′ start sites drives tumour initiation’, *Nature*, 541(7638), pp. 494–499. doi: 10.1038/nature21036.

Shabalina, S. A. *et al.* (2003) ‘Patterns in interspecies similarity correlate with nucleotide composition in mammalian 3′UTRs’, *Nucleic Acids Research*, 31(18), pp. 5433–5439. doi: 10.1093/NAR/GKG751.

Shabalina, S. A. *et al.* (2004) ‘Comparative analysis of orthologous eukaryotic mRNAs: potential hidden functional signals’, *Nucleic Acids Research*, 32(5), p. 1774. doi: 10.1093/NAR/GKH313.

Shabalina, S. A., Ogurtsov, A. Y. and Spiridonov, N. A. (2006) ‘A periodic pattern of mRNA secondary structure created by the genetic code’, *Nucleic Acids Research*, 34(8), p. 2428. doi: 10.1093/NAR/GKL287.

Sherman, B. T. *et al.* (2022) ‘DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update)’, *Nucleic Acids Research*, 50(W1), p. W216. doi: 10.1093/NAR/GKAC194.

Shi, Y. and Manley, J. L. (2015) ‘The end of the message: Multiple protein–RNA interactions define the mRNA polyadenylation site’, *Genes and Development*, 29(9), pp. 889–897. doi: 10.1101/GAD.261974.115,.

Shoemaker, C. J. and Green, R. (2011) ‘Kinetic analysis reveals the ordered coupling of translation termination and ribosome recycling in yeast’, *Proceedings of the National Academy of Sciences of the United States of America*, 108(51). doi: 10.1073/pnas.1113956108.

Shukla, S. *et al.* (2011) ‘CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing’, *Nature*, 479(7371), pp. 74–79. doi: 10.1038/NATURE10442.

Siepel, A. *et al.* (2005) ‘Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes’, *Genome Research*, 15(8), pp. 1034–1050. doi: 10.1101/GR.3715005.

Silva, A. L. and Romão, L. (2009) ‘The mammalian nonsense-mediated mRNA decay pathway: To decay or not to decay! Which players make the decision?’, *FEBS Letters*, 583(3), pp. 499–505. doi: 10.1016/J.FEBSLET.2008.12.058.

Silva, J., Fernandes, R. and Romão, L. (2017) 'Gene expression regulation by upstream open reading frames in rare diseases', *Dis Res Treat*, 2(4), pp. 33-38. doi: 10.29245/2572-9411/2017/4.1121.

Silverman, R. H. (1994) ‘Fascination with 2-5A-dependent RNase: a unique enzyme that functions in interferon action’, *Journal of interferon research*, 14(3), pp. 101–104. doi: 10.1089/JIR.1994.14.101.

Simon, A. E. and Miller, W. A. (2013) ‘3′ cap-independent translation enhancers of plant viruses’, *Annual Review of Microbiology*, 67, pp. 21–42. doi: 10.1146/annurev-micro-092412-155609.

Singh, P. *et al.* (2009) ‘Global changes in processing of mRNA 3’ untranslated regions characterize clinically distinct cancer subtypes’, *Cancer research*, 69(24), pp. 9422–9430. doi: 10.1158/0008-5472.CAN-09-2236.

Skabkin, M. A. *et al.* (2013) ‘Reinitiation and other unconventional post-termination events during eukaryotic translation’, *Molecular cell*, 51(2), p. 249. doi: 10.1016/J.MOLCEL.2013.05.026.

Slavoff, S. A. *et al.* (2013) ‘Peptidomic discovery of short open reading frame-encoded peptides in human cells’, *Nature Chemical Biology*, 9(1), pp. 59–64. doi:

10.1038/nchembio.1120.

Smith, T. F. and Waterman, M. S. (1981) ‘Identification of common molecular subsequences.’, *Journal of molecular biology*, 147(1), pp. 195–7. doi: 10.1016/0022-2836(81)90087-5.

Smith, Z. D. and Meissner, A. (2013) ‘DNA methylation: Roles in mammalian development’, *Nature Reviews Genetics*, 14(3), pp. 204–220. doi: 10.1038/NRG3354.

Solassol, J., Mange, A. and Maudelonde, T. (2011) ‘FKBP family proteins as promising new biomarkers for cancer’, *Current opinion in pharmacology*, 11(4), pp. 320–325. doi: 10.1016/J.COPH.2011.03.012.

Somers, J., Pöyry, T. and Willis, A. E. (2013) ‘A perspective on mammalian upstream open reading frame function’, *International Journal of Biochemistry and Cell Biology*. Elsevier Ltd, pp. 1690–1700. doi: 10.1016/j.biocel.2013.04.020.

Sonenberg, N. *et al.* (1978) ‘A polypeptide in eukaryotic initiation factors that crosslinks specifically to the 5′-terminal cap in mRNA’, *Proc. Natl Acad. Sci. USA*, 75(10), pp. 4843–4847. doi: 10.1073/pnas.75.10.4843.

Song, H. *et al.* (2000) ‘The crystal structure of human eukaryotic release factor eRF1--mechanism of stop codon recognition and peptidyl-tRNA hydrolysis’, *Cell*, 100(3), pp. 311–321. doi: 10.1016/S0092-8674(00)80667-4.

Sood, P. *et al.* (2006) ‘Cell-type-specific signatures of microRNAs on target mRNA expression’, *Proceedings of the National Academy of Sciences of the United States of America*, 103(8), pp. 2746–2751. doi: 10.1073/PNAS.0511045103,.

Sriram, A., Bohlen, J. and Teleman, A. A. (2018) ‘Translation acrobatics: how cancer cells exploit alternate modes of translational initiation’, *EMBO reports*, 19(10). doi: 10.15252/EMBR.201845947.

Starck, S. *et al.* (2016) ‘Translation from the 5′ untranslated region shapes the integrated stress response’, *Science*, 351(6272). doi: 10.1126/science.aad3867.

Starck, S. R. *et al.* (2012) ‘Leucine-tRNA initiates at CUG start codons for protein synthesis and presentation by MHC class I’, *Science*, 336(6089), pp. 1719–1723. doi: 10.1126/science.1220270.

Stein, C. S. *et al.* (2018) ‘Mitoregulin: A lncRNA-Encoded Microprotein that Supports Mitochondrial Supercomplexes and Respiratory Efficiency’, *Cell reports*, 23(13), pp. 3710-3720.e8. doi: 10.1016/J.CELREP.2018.06.002.

Steneberg, P. and Samakovlis, C. (2001) ‘A novel stop codon readthrough mechanism produces functional Headcase protein in Drosophila trachea’, *EMBO Reports*, 2(7), p. 593. doi: 10.1093/EMBO-REPORTS/KVE128.

Stern-Ginossar, N. and Ingolia, N. T. (2015) ‘Ribosome Profiling as a Tool to Decipher Viral Complexity’, *Annual Review of Virology*. Annual Reviews Inc., pp. 335–349. doi: 10.1146/annurev-virology-100114-054854.

Stoiber, M. H. *et al.* (2015) ‘Extensive cross-regulation of post-transcriptional regulatory networks in Drosophila’, *Genome Research*, 25(11), p. 1692. doi:

10.1101/GR.182675.114.

Subramanian, K. *et al.* (2022) ‘The Codon Statistics Database: A Database of Codon Usage Bias’, *Molecular Biology and Evolution*, 39(8). doi: 10.1093/MOLBEV/MSAC157.

Sun, G. *et al.* (2015) ‘B cell translocation gene 1 reduces the biological outcome of kidney cancer through induction of cell proliferation, cell cycle arrest, cell apoptosis and cell metastasis’, *International journal of molecular medicine*, 35(3), pp. 777–783. doi: 10.3892/IJMM.2014.2058.

Tahmasebi, S. *et al.* (2018) ‘Translation deregulation in human disease’, *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, pp. 791–807. doi: 10.1038/s41580-018-0034-x.

Takahashi, K. *et al.* (2005) ‘Evolutionarily conserved non-AUG translation initiation in NAT1/p97/DAP5 (EIF4G2)’, *Genomics*, 85(3), pp. 360–371. doi: 10.1016/J.YGENO.2004.11.012.

Takata, M. A. *et al.* (2017) ‘CG-dinucleotide suppression enables antiviral defense targeting non-self RNA’, *Nature*, 550(7674), p. 124. doi: 10.1038/NATURE24039.

Takyar, S., Hickerson, R. P. and Noller, H. F. (2005) ‘mRNA helicase activity of the ribosome’, *Cell*, 120(1), pp. 49–58. doi: 10.1016/j.cell.2004.11.042.

Tarun, S. Z. and Sachs, A. B. (1996) ‘Association of the yeast poly(A) tail binding protein with translation initiation factor eIF-4G’, *EMBO Journal*, 15(24), pp. 7168–7177. doi: 10.1002/j.1460-2075.1996.tb01108.x.

Tazi, J., Bakkour, N. and Stamm, S. (2009) ‘Alternative splicing and disease’, *Biochimica et Biophysica Acta - Molecular Basis of Disease*. NIH Public Access, pp. 14–26. doi: 10.1016/j.bbadis.2008.09.017.

Thompson, M. K. and Gilbert, W. V. (2017) ‘mRNA length-sensing in eukaryotic translation: reconsidering the “closed loop” and its implications for translational control’, *Current Genetics*. Springer Verlag, pp. 613–620. doi: 10.1007/s00294-016-0674-3.

Tian, B. and Graber, J. H. (2012) ‘Signals for pre-mRNA cleavage and polyadenylation’, *Wiley Interdisciplinary Reviews: RNA*, 3(3), pp. 385–396. doi: 10.1002/WRNA.116,.

Tomuro, K. *et al.* (2024) ‘Calibrated ribosome profiling assesses the dynamics of ribosomal flux on transcripts’, *Nature Communications 2024 15:1*, 15(1), pp. 1–17. doi: 10.1038/s41467-024-51258-0.

Torabi, N. and Kruglyak, L. (2012) ‘Genetic basis of hidden phenotypic variation revealed by increased translational readthrough in yeast’, *PLoS Genetics*, 8(3). doi: 10.1371/journal.pgen.1002546.

Tüfekci, K. U. *et al.* (2014) ‘The role of microRNAs in human diseases’, *Methods in Molecular Biology*, 1107, pp. 33–50. doi: 10.1007/978-1-62703-748-8_3.

Vanderperre, B. *et al.* (2013) ‘Direct detection of alternative open reading frames translation products in human significantly expands the proteome’, *PloS one*, 8(8).

doi: 10.1371/JOURNAL.PONE.0070698.

Vattem, Krishna M. and Wek, R. C. (2004) ‘Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells’, *Proceedings of the National Academy of Sciences of the United States of America*, 101(31), pp. 11269–11274. doi: 10.1073/pnas.0400541101.

Vicens, Q., Kieft, J. S. and Rissland, O. S. (2018) ‘Revisiting the Closed-Loop Model and the Nature of mRNA 5′–3′ Communication’, *Molecular Cell*. Cell Press, pp. 805–812. doi: 10.1016/j.molcel.2018.10.047.

Vieira de Souza, E. *et al.* (2024) ‘Rp3: Ribosome profiling-assisted proteogenomics improves coverage and confidence during microprotein discovery’, *Nature Communications 2024 15:1*, 15(1), pp. 1–14. doi: 10.1038/s41467-024-50301-4.

Wang, J., Chen, J. and Sen, S. (2016) ‘MicroRNA as Biomarkers and Diagnostics’, *Journal of Cellular Physiology*. Wiley-Liss Inc., pp. 25–30. doi: 10.1002/jcp.25056.

Wang, L. *et al.* (2017) ‘Epigenetic inactivation of HOXA11, a novel functional tumor suppressor for renal cell carcinoma, is associated with RCC TNM classification’, *Oncotarget*, 8(13), p. 21861. doi: 10.18632/ONCOTARGET.15668.

Wang, L., Wang, S. and Li, W. (2012) ‘RSeQC: quality control of RNA-seq experiments’, *Bioinformatics*, 28(16), pp. 2184–2185. doi: 10.1093/BIOINFORMATICS/BTS356.

Wang, S. *et al.* (2023) ‘Targeting ACYP1-mediated glycolysis reverses lenvatinib resistance and restricts hepatocellular carcinoma progression’, *Drug resistance updates : reviews and commentaries in antimicrobial and anticancer chemotherapy*, 69. doi: 10.1016/J.DRUP.2023.100976.

Wang, Weiguang *et al.* (2009) ‘Atf4 regulates chondrocyte proliferation and differentiation during endochondral ossification by activating Ihh transcription’, *Development*, 136(24), pp. 4143–4153. doi: 10.1242/dev.043281.

Wang, W. *et al.* (2019) ‘Evolutionary and functional implications of 3′ untranslated region length of mRNAs by comprehensive investigation among four taxonomically diverse metazoan species’, *Genes and Genomics*, 41(7), pp. 747–755. doi: 10.1007/S13258-019-00808-8,.

Wang, Y. *et al.* (2020) ‘LncRNA-encoded polypeptide ASRPS inhibits triple-negative breast cancer angiogenesis’, *The Journal of experimental medicine*, 217(3). doi: 10.1084/JEM.20190950.

Wangen, J. R. and Green, R. (2020) ‘Stop codon context influences genome-wide stimulation of termination codon readthrough by aminoglycosides’, *eLife*, 9. doi: 10.7554/ELIFE.52611.

Weaver, J. *et al.* (2019) ‘Identifying Small Proteins by Ribosome Profiling with Stalled Initiation Complexes’, *mBio*, 10(2). doi: 10.1128/MBIO.02819-18.

Weeks, N. T. and Luecke, G. R. (2017) ‘Performance analysis and optimization of SAMtools sorting’, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10104 LNCS, pp. 409–420. doi: 10.1007/978-3-319-58943-5_33.

Wei-Lin Popp, M. and Maquat, L. E. (2013) ‘Organizing principles of Mammalian nonsense-mediated mRNA decay’, *Annual Review of Genetics*, 47, pp. 139–165. doi: 10.1146/ANNUREV-GENET-111212-133424.

Weingarten-Gabbay, S. *et al.* (2016) ‘Comparative genetics. Systematic discovery of cap-independent translation sequences in human and viral genomes’, *Science (New York, N.Y.)*, 351(6270). doi: 10.1126/SCIENCE.AAD4939.

Wek, R. C. (2018) ‘Role of eIF2α kinases in translational control and adaptation to cellular stress’, *Cold Spring Harbor Perspectives in Biology*, 10(7), p. a032870. doi: 10.1101/cshperspect.a032870.

Wells, S. E. *et al.* (1998) ‘Circularization of mRNA by eukaryotic translation initiation factors’, *Molecular Cell*, 2(1), pp. 135–140. doi: 10.1016/S1097-2765(00)80122-7.

Werner, M. *et al.* (1987) ‘The leader peptide of yeast gene CPA1 is essential for the translational repression of its expression’, *Cell*, 49(6), pp. 805–813. doi: 10.1016/0092-8674(87)90618-0.

Wethmar, K. *et al.* (2014) ‘UORFdb - A comprehensive literature database on eukaryotic uORF biology’, *Nucleic Acids Research*, 42(D1), p. D60. doi: 10.1093/nar/gkt952.

Wilson, B. A. and Masel, J. (2011) ‘Putatively noncoding transcripts show extensive association with ribosomes’, *Genome Biology and Evolution*, 3(1), pp. 1245–1252. doi: 10.1093/gbe/evr099.

Wise, M. J. (2003) ‘Alignment algorithms revisited: Alignment algorithms for low similarity protein sequence comparisons’, *European Control Conference, ECC 2003*, pp. 3386–3391. doi: 10.23919/ECC.2003.7086563.

Wollerton, M. C. *et al.* (2004) ‘Autoregulation of Polypyrimidine Tract Binding Protein by Alternative Splicing Leading to Nonsense-Mediated Decay’, *Molecular Cell*, 13(1), pp. 91–100. doi: 10.1016/S1097-2765(03)00502-1.

Wong, L. E. *et al.* (2012) ‘Selectivity of stop codon recognition in translation termination is modulated by multiple conformations of GTS loop in eRF1’, *Nucleic Acids Research*, 40(12), p. 5751. doi: 10.1093/NAR/GKS192.

Wright, B. W. *et al.* (2022) ‘The dark proteome: translation from noncanonical open reading frames’, *Trends in Cell Biology*, 32(3), pp. 243–258. doi: 10.1016/J.TCB.2021.10.010.

Wright, B. W., Molloy, M. P. and Jaschke, P. R. (2021) ‘Overlapping genes in natural and engineered genomes’, *Nature Reviews Genetics 2021 23:3*, 23(3), pp. 154–168. doi: 10.1038/s41576-021-00417-w.

Wu, B. *et al.* (2016) ‘Translation dynamics of single mRNAs in live cells and neurons’, *Science*, 352(6292), pp. 1430–1435. doi: 10.1126/science.aaf1084.

Wu, P. *et al.* (2020a) ‘Emerging role of tumor-related functional peptides encoded by lncRNA and circRNA’, *Molecular cancer*, 19(1). doi: 10.1186/S12943-020-1147-3.

Wu, Q. *et al.* (2020b) 'Translation of small downstream ORFs enhances translation of canonical main open reading frames', *The EMBO Journal*, 39(17). doi: 10.15252/embj.2020104763.

Wu, S. *et al.* (2020) 'A novel micropeptide encoded by y-linked LINC00278 links cigarette smoking and ar signaling in male esophageal squamous cell carcinoma', *Cancer Research*, 80(13), pp. 2790–2803. doi: 10.1158/0008-5472.CAN-19-3440/654126/AM/A-NOVEL-MICROPEPTIDE-ENCODED-BY-Y-LINKED-LINC00278.

Wyllie, K., Panagopoulos, V. and Cox, T. R. (2023) 'The role of peroxidasin in solid cancer progression', *Biochemical Society Transactions*, 51(5), p. 1881. doi: 10.1042/BST20230018.

Xia, X. *et al.* (2019) 'A novel tumor suppressor protein encoded by circular AKT3 RNA inhibits glioblastoma tumorigenicity by competing with active phosphoinositide-dependent Kinase-1', *Molecular cancer*, 18(1). doi: 10.1186/S12943-019-1056-5.

Xia, Z. *et al.* (2022) 'A Review of Parallel Implementations for the Smith–Waterman Algorithm', *Interdisciplinary Sciences – Computational Life Sciences*, 14(1). doi: 10.1007/S12539-021-00473-0.

Yamashita, A. *et al.* (2005) 'Concerted action of poly(A) nucleases and decapping enzyme in mammalian mRNA turnover', *Nature Structural and Molecular Biology*, 12(12), pp. 1054–1063. doi: 10.1038/nsmb1016.

Yan, X. *et al.* (2016) 'Dynamics of Translation of Single mRNA Molecules in Vivo', *Cell*, 165(4), pp. 976–989. doi: 10.1016/j.cell.2016.04.034.

Yao, Y. L. *et al.* (2011) 'FKBPs in chromatin modification and cancer', *Current opinion in pharmacology*, 11(4), pp. 301–307. doi: 10.1016/J.COPH.2011.03.005.

Young, D. J. *et al.* (2015) 'Rli1/ABCE1 Recycles Terminating Ribosomes and Controls Translation Reinitiation in 3′UTRs In Vivo', *Cell*, 162(4), pp. 872–884. doi: 10.1016/j.cell.2015.07.041.

Young, S. K. and Wek, R. C. (2016) 'Upstream open reading frames differentially regulate genespecific translation in the integrated stress response', *Journal of Biological Chemistry*. American Society for Biochemistry and Molecular Biology Inc., pp. 16927–16935. doi: 10.1074/jbc.R116.733899.

Yu, X. *et al.* (2017) 'Long non-coding RNA Linc-RAM enhances myogenic differentiation by interacting with MyoD', *Nature Communications 2017 8:1*, 8(1), pp. 1–12. doi: 10.1038/ncomms14016.

Yuan, Y., Yang, E. and Zhang, R. (2024) 'Wfold: A new method for predicting RNA secondary structure with deep learning', *Computers in Biology and Medicine*, 182, p. 109207. doi: 10.1016/J.COMPBIOMED.2024.109207.

Yuniati, L. *et al.* (2019) 'Tumor suppressors BTG1 and BTG2: Beyond growth control', *Journal of Cellular Physiology*, 234(5), p. 5379. doi: 10.1002/JCP.27407.

Zahdeh, F. and Carmel, L. (2016) 'The role of nucleotide composition in premature termination codon recognition', *BMC Bioinformatics*, 17(1). doi: 10.1186/S12859-

016-1384-Z.

Zhang, H. *et al.* (2021) ‘Determinants of genome-wide distribution and evolution of uORFs in eukaryotes’, *Nature Communications 2021 12:1*, 12(1), pp. 1–17. doi: 10.1038/s41467-021-21394-y.

Zhang, S. *et al.* (2020) ‘Mitochondrial peptide BRAWNIN is essential for vertebrate respiratory complex III assembly’, *Nature communications*, 11(1). doi: 10.1038/S41467-020-14999-2.

Zhang, X. *et al.* (2015) ‘Translational control of the cytosolic stress response by mitochondrial ribosomal protein L18’, *Nature Structural & Molecular Biology 2015 22:5*, 22(5), pp. 404–410. doi: 10.1038/nsmb.3010.

Zhao, N. *et al.* (2017) ‘SZRD1 is a Novel Protein that Functions as a Potential Tumor Suppressor in Cervical Cancer’, *Journal of Cancer*, 8(11), p. 2132. doi: 10.7150/JCA.18806.

Zhou, B. *et al.* (2021) ‘Translation of noncoding RNAs and cancer’, *Cancer letters*, 497, pp. 89–99. doi: 10.1016/J.CANLET.2020.10.002.

Zhou, J. *et al.* (2015) ‘Dynamic m6 A mRNA methylation directs translational control of heat shock response’, *Nature*, 526(7574), pp. 591–594. doi: 10.1038/nature15377.

Zhou, J. *et al.* (2018) ‘N6-Methyladenosine Guides mRNA Alternative Translation during Integrated Stress Response’, *Molecular Cell*, 69(4), pp. 636-647.e7. doi: 10.1016/j.molcel.2018.01.019.

Zhou, L. *et al.* (2022a) ‘ACYP1 Is a Pancancer Prognostic Indicator and Affects the Immune Microenvironment in LIHC’, *Frontiers in Oncology*, 12, p. 1. doi: 10.3389/FONC.2022.875097/FULL.

Zhou, X. *et al.* (2022b) ‘Discovery of the hidden coding information in cancers: Mechanisms and biological functions’, *International Journal of Cancer*. doi: 10.1002/IJC.34360.

Zinoviev, A., Hellen, C. U. T. and Pestova, T. V. (2015) ‘Multiple mechanisms of reinitiation on bicistronic calicivirus mRNAs’, *Molecular cell*, 57(6), p. 1059. doi: 10.1016/J.MOLCEL.2015.01.039.

Zinshteyn, B., Rojas-Duran, M. F. and Gilbert, W. V. (2017) ‘Translation initiation factor eIF4G1 preferentially binds yeast transcript leaders containing conserved oligo-uridine motifs’, *RNA*, 23(9), pp. 1365–1375. doi: 10.1261/RNA.062059.117/-/DC1.

Appendices

Appendix 1 – Additional quality control data from RNAseq and RP datasets to support Materials and Methods chapter

Table 8.1: Summary of the MultiQC data for RNAseq datasets before any processing in a Galaxy workflow. *Quality control data of RNAseq datasets gathered from the FastQC outputs, pulled together using MultiQC, for each dataset within the Galaxy workflow used on these datasets.*

Project	RNAseq Dataset	Total Reads	Sequences flagged as poor quality	Sequence length	GC base Percentage	Duplicates Percentage	Basic Statistics	Per Base Sequence Quality	Per Sequence Quality Scores	Adapter Content
PRJNA256316	SRR2064424	19254252	0	51	56	53.51682	pass	pass	pass	pass
	SRR2064425	15506913	0	51	59	59.58383	pass	pass	pass	pass
	SRR2064426	31392714	0	51	50	34.93515	pass	pass	pass	pass
	SRR2064427	27246329	0	51	50	34.74044	pass	pass	pass	pass
	SRR2064428	18512001	0	51	49	28.79826	pass	pass	pass	pass
	SRR2064429	25087358	0	51	52	33.64363	pass	pass	pass	pass
PRJNA880902	SRR21595026	34892969	0	151	54	73.70766	pass	warn	pass	fail
	SRR21595027	24019540	0	151	54	71.9712	pass	warn	pass	fail
	SRR21595028	38533467	0	151	53	74.2223	pass	warn	pass	fail
	SRR21595029	30706264	0	151	54	71.96845	pass	warn	pass	fail
	SRR21595030	27895948	0	151	54	72.36307	pass	warn	pass	fail
	SRR21595031	32295595	0	151	53	72.74266	pass	warn	pass	fail
PRJNA532400	SRR8883215	91609994	0	50	48	88.59359	pass	pass	pass	pass
	SRR8883216	79080659	0	50	48	91.9542	pass	pass	pass	pass
	SRR8883217	75782840	0	50	47	90.89907	pass	pass	pass	pass
	SRR8883218	96545767	0	50	48	93.69606	pass	pass	pass	pass

***Table 8.2: Summary of the MultiQC data for ribosome profiling (RP) datasets before any processing in a Galaxy workflow**. Quality control data of RP datasets gathered from the FastQC outputs, pulled together using MultiQC, for each dataset within the Galaxy workflow used on these datasets.*

Project	RP Dataset	Total Reads	Sequences flagged as poor quality	Sequence length	GC base Percentage	Duplicates Percentage	Basic Statistics	Per Base Sequence Quality	Per Sequence Quality Scores	Adapter Content
PRJNA256316	SRR1528686	19725261	0	51	59	91.68568	pass	pass	pass	fail
	SRR1528687	27215800	0	51	59	92.41518	pass	pass	pass	fail
	SRR1528688	17521963	0	51	58	88.98193	pass	pass	pass	fail
	SRR1528689	24145183	0	51	58	89.82847	pass	pass	pass	fail
	SRR1528690	22206265	0	51	60	77.71798	pass	pass	pass	fail
	SRR1528691	31147450	0	51	60	79.53531	pass	pass	pass	fail
	SRR1528692	25657299	0	51	57	78.84923	pass	pass	pass	fail
	SRR1528693	36181563	0	51	57	80.48734	pass	pass	pass	fail
	SRR1528694	14984509	0	51	59	95.8838	pass	pass	pass	fail
	SRR1528695	20492315	0	51	59	96.27529	pass	pass	pass	fail
	SRR1528696	20302267	0	51	59	85.03634	pass	pass	pass	fail
	SRR1528697	28365554	0	51	59	86.20666	pass	pass	pass	fail
PRJNA880902	SRR21595020	33507379	0	36-151	53	74.89432	pass	pass	pass	pass
	SRR21595021	22948925	0	36-151	53	73.27434	pass	pass	pass	pass
	SRR21595022	36989777	0	36-151	53	75.34412	pass	pass	pass	pass
	SRR21595023	29418207	0	36-151	53	73.26433	pass	pass	pass	pass
	SRR21595024	26810239	0	36-151	53	73.52129	pass	pass	pass	pass
	SRR21595025	31025645	0	36-151	52	73.81721	pass	pass	pass	pass
PRJNA532400	SRR8883211	48329936	0	50	65	79.58471	pass	pass	pass	pass
	SRR8883212	37835349	0	50	65	77.74159	pass	pass	pass	pass
	SRR8883213	43460625	0	50	63	90.31357	pass	pass	pass	pass
	SRR8883214	36322120	0	50	68	79.38972	pass	pass	pass	pass

Table 8.3: Summary of the MultiQC and alignment to human genome data for RNAseq datasets processed with Galaxy workflow. *Quality control data of RNAseq datasets gathered from the FastQC outputs, pulled together using MultiQC, for each dataset within the Galaxy workflow used on these datasets. Also included the overall alignment percentage for the RNAseq datasets against the human genome.*

Project	RNAseq Dataset	Total Reads	Sequences flagged as poor quality	Sequence length	GC base Percentage	Duplicates Percentage	Basic Statistics	Per Base Sequence Quality	Per Sequence Quality Scores	Adapter Content	Human genome overall alignment percentage
PRJNA256316	SRR2064424	18898979	0	25-51	56	52.63	pass	pass	pass	pass	93.73
	SRR2064425	14857518	0	25-51	58	57.82	pass	pass	pass	pass	94.86
	SRR2064426	31086128	0	25-51	50	34.28	pass	pass	pass	pass	97.29
	SRR2064427	26920232	0	25-51	50	33.89	pass	pass	pass	pass	97.49
	SRR2064428	18223371	0	25-51	49	27.89	pass	pass	pass	pass	92.06
	SRR2064429	24762149	0	25-51	52	32.89	pass	pass	pass	pass	96.01
PRJNA880902	SRR21595026	31689447	0	25-151	52	75.81	pass	pass	pass	pass	95.40
	SRR21595027	21448979	0	25-151	53	74.45	pass	pass	pass	pass	94.81
	SRR21595028	34953636	0	25-151	52	76.17	pass	pass	pass	pass	95.40
	SRR21595029	27612036	0	25-151	52	74.43	pass	pass	pass	pass	95.17
	SRR21595030	25384923	0	25-151	52	74.42	pass	pass	pass	pass	95.49
	SRR21595031	29346667	0	25-151	52	74.78	pass	pass	pass	pass	95.48
PRJNA532400	SRR8883215	91229140	0	25-50	48	87.41	pass	pass	pass	pass	65.43
	SRR8883216	78704593	0	25-50	48	90.78	pass	pass	pass	pass	63.99
	SRR8883217	75528464	0	25-50	47	90.00	pass	pass	pass	pass	46.43
	SRR8883218	96220164	0	25-50	48	92.86	pass	pass	pass	pass	47.78

Table 8.4: Summary of the MultiQC and alignment to human genome data for Ribosome Profiling (RP) datasets processed with Galaxy workflow. *Quality control data of RP datasets gathered from the FastQC outputs, pulled together using MultiQC, for each dataset within the Galaxy workflow used on these datasets. Also included the overall alignment percentage for the RP datasets against the human genome.*

Project	RP Dataset	Total Reads	Sequences flagged as poor quality	Sequence length	GC base Percentage	Duplicates Percentage	Basic Statistics	Per Base Sequence Quality	Per Sequence Quality Scores	Adapter Content	Human genome overall alignment percentage
PRJNA256316	SRR1528686	19254400	0	25-35	61	92.90017	pass	pass	pass	pass	91.44
	SRR1528687	26439789	0	25-35	61	93.40329	pass	pass	pass	pass	91.27
	SRR1528688	17140427	0	25-35	60	90.29463	pass	pass	pass	pass	92.73
	SRR1528689	23532265	0	25-35	60	90.82256	pass	pass	pass	pass	92.59
	SRR1528690	20942457	0	25-35	64	79.56826	pass	pass	pass	pass	87.82
	SRR1528691	29222920	0	25-35	64	80.96067	pass	pass	pass	pass	87.72
	SRR1528692	24705823	0	25-35	60	80.6093	pass	pass	pass	pass	89.12
	SRR1528693	34658967	0	25-35	60	81.9315	pass	pass	pass	pass	89.05
	SRR1528694	14720141	0	25-35	61	96.79558	pass	pass	pass	pass	94.85
	SRR1528695	20037361	0	25-35	61	97.01567	pass	pass	pass	pass	94.8
	SRR1528696	19009350	0	25-35	63	86.27238	pass	pass	pass	pass	87.55
	SRR1528697	26447143	0	25-35	63	87.14579	pass	pass	pass	pass	87.54
PRJNA880902	SRR21595020	33012759	0	25-35	53	78.08159	pass	pass	pass	pass	93.94
	SRR21595021	22549382	0	25-35	53	77.31465	pass	pass	pass	pass	93.17
	SRR21595022	36434744	0	25-35	53	78.62782	pass	pass	pass	pass	93.84
	SRR21595023	28937744	0	25-35	52	77.13692	pass	pass	pass	pass	93.42
	SRR21595024	26420304	0	25-35	53	77.00078	pass	pass	pass	pass	93.91
	SRR21595025	30568394	0	25-35	52	77.2468	pass	pass	pass	pass	93.84
PRJNA532400	SRR8883211	47858141	0	25-35	69	83.61498	pass	pass	pass	pass	84.41
	SRR8883212	37461208	0	25-35	70	82.15682	pass	pass	pass	pass	86.21
	SRR8883213	43052280	0	25-35	67	93.16985	pass	pass	pass	pass	86.37
	SRR8883214	35989091	0	25-35	72	84.16579	pass	pass	pass	pass	88.05

Appendix 2: Comparing the effect of altering the length criteria for alignment of dORFs in conservation analysis

Table 8.5: Filtering results from the Smith-Waterman alignments based on the alignment length has little effect on the mean similarity percentages of downstream open reading frame (dORF) sequences which start with an AUG start codon between humans and various species. *The mean AUG dORF similarity between other species and humans and the number of comparisons are reported where the alignment length is used as described in the methods section (100%) and also where the alignment length can be 90% or 80% of the shortest dORF being aligned.*

	Mean AUG dORF similarity to human with altered alignment length			**AUG dORF comparisons with altered alignment length**		
Species	**100%**	**90%**	**80%**	**100%**	**90%**	**80%**
P. troglodytes	92.46	90.95	89.74	350026	373559	405280
M. mulatta	83.47	82.11	80.90	241772	283883	320093
C. lupus	71.54	71.66	71.23	167113	226139	272998
B. taurus	68.36	68.88	68.68	160201	209846	254819
M. musculus	66.2	66.91	66.96	138168	194208	245393
R. norvegicus	63.83	64.78	65.06	150024	205911	256830
G. gallus	60.69	61.88	62.68	100081	144499	189621
X. tropicalis	56.14	57.63	58.80	135091	185331	236334
D. rerio	56.01	57.41	58.53	119397	164430	209375
A. gambiae	50.43	51.54	52.55	23178	28986	34582
D. melanogaster	53.51	54.83	55.91	35381	46068	56428
C. elegans	50.96	52.06	52.94	35079	43236	50456
S. pombe	52.28	53.38	54.33	19479	24169	28466
M. oryzae	51.41	52.59	53.50	19929	25202	29859
N. crassa	51.09	52.17	53.10	17998	22600	26884
A. thaliana	52.18	53.38	54.34	39584	49467	58786
O. sativa	53.14	54.13	54.98	34187	43511	51996

Appendix 3: dORF vs 3’ UTR similarity with length restriction <10000, <5000, <2500

Table 8.6: The mean similarity percentages of downstream open reading frame (dORF) sequences which start with an AUG start codon is increased compared to the 3'untranslated region (UTR) sequences containing the AUG dORFs between humans and various species when 3' UTR length is less than 10000 nucleotides. *One-Way ANOVA multiple comparisons test with Šídák adjusted P values was used to compare the means and the data generated by this test are included in the table.*

Species	Mean 3' UTR similarity to human	Mean AUG dORF similarity to human	Mean Difference	Standard Error of Difference	95% Confidence Intervals of difference	3' UTR comparisons	AUG dORF comparisons	Adjusted P Value
P. troglodytes	85.91	92.98	7.073	0.1294	6.689 to 7.457	322375	13404	<0.0001
M. mulatta	81.7	84.94	3.246	0.1543	2.789 to 3.704	221139	9447	<0.0001
C. lupus	62.15	72.47	10.32	0.1752	9.800 to 10.84	150913	7372	<0.0001
B. taurus	61.44	69.28	7.841	0.1774	7.315 to 8.368	143488	7191	<0.0001
M. musculus	55.79	67.09	11.31	0.1866	10.75 to 11.86	122460	6521	<0.0001
R. norvegicus	55.14	64.89	9.742	0.1762	9.219 to 10.26	130322	7332	<0.0001
G. gallus	44.4	50.92	6.516	0.2182	5.868 to 7.163	88669	4771	<0.0001
X. tropicalis	41.9	56.41	14.51	0.1908	13.94 to 15.07	111147	6255	<0.0001
D. rerio	40.77	55.88	15.11	0.1959	14.53 to 15.69	102034	5949	<0.0001
A. gambiae	40.62	50.85	10.22	0.4653	8.843 to 11.60	18726	1052	<0.0001
D. melanogaster	41.11	53.6	12.48	0.3963	11.31 to 13.66	28414	1443	<0.0001
C. elegans	41.45	51.07	9.624	0.3734	8.516 to 10.73	30181	1630	<0.0001
S. pombe	42.32	52.47	10.15	0.5507	8.520 to 11.79	13708	750	<0.0001
M. oryzae	40.83	51.62	10.79	0.5321	9.210 to 12.37	16283	799	<0.0001
N. crassa	40.21	52.01	11.79	0.6198	9.954 to 13.63	11973	589	<0.0001
A. thaliana	41.63	53.13	11.5	0.4148	10.27 to 12.73	26221	1316	<0.0001
O. sativa	41.24	52.96	11.72	0.4199	10.48 to 12.97	25237	1285	<0.0001

Table 8.7: The mean similarity percentages of downstream open reading frame (dORF) sequences which start with an AUG start codon is increased compared to the 3'untranslated region (UTR) sequences containing the AUG dORFs between humans and various species when 3' UTR length is less than 5000 nucleotides. *One-Way ANOVA multiple comparisons test with Šídák adjusted P values was used to compare the means and the data generated by this test are included in the table.*

Species	Mean 3' UTR similarity to human	Mean AUG dORF similarity to human	Mean Difference	Standard Error of Difference	95% Confidence Intervals of difference	3' UTR comparisons	AUG dORF comparisons	Adjusted P Value
P. troglodytes	85.97	93.13	7.159	0.1322	6.766 to 7.551	243239	13021	<0.0001
M. mulatta	81.86	85.94	4.079	0.1589	3.607 to 4.550	168654	9016	<0.0001
C. lupus	62.28	72.75	10.47	0.1804	9.936 to 11.01	116513	7043	<0.0001
B. taurus	61.39	69.62	8.228	0.1826	7.686 to 8.770	111325	6883	<0.0001
M. musculus	55.94	67.28	11.34	0.1921	10.77 to 11.91	96159	6239	<0.0001
R. norvegicus	55.23	65.28	10.05	0.1822	9.508 to 10.59	101499	6954	<0.0001
G. gallus	44.46	52.93	8.472	0.2272	7.798 to 9.146	68017	4459	<0.0001
X. tropicalis	42	56.48	14.49	0.202	13.89 to 15.08	79419	5675	<0.0001
D. rerio	40.88	55.91	15.02	0.2074	14.41 to 15.64	73358	5393	<0.0001
A. gambiae	40.56	50.88	10.31	0.4904	8.860 to 11.77	13867	961	<0.0001
D. melanogaster	41.19	53.32	12.13	0.4177	10.89 to 13.37	20106	1320	<0.0001
C. elegans	41.49	51.17	9.677	0.3978	8.497 to 10.86	20758	1462	<0.0001
S. pombe	42.36	52.45	10.08	0.5761	8.376 to 11.79	10327	695	<0.0001
M. oryzae	40.89	51.94	11.05	0.5587	9.389 to 12.70	11689	736	<0.0001
N. crassa	40.33	51.91	11.59	0.643	9.681 to 13.50	9266	554	<0.0001
A. thaliana	41.67	53.19	11.52	0.4374	10.22 to 12.82	18292	1204	<0.0001
O. sativa	41.32	53.07	11.75	0.4426	10.44 to 13.06	18277	1174	<0.0001

Table 8.8: The mean similarity percentages of downstream open reading frame (dORF) sequences which start with an AUG start codon is increased compared to the 3'untranslated region (UTR) sequences containing the AUG dORFs between humans and various species when 3' UTR length is less than 2500 nucleotides. *One-Way ANOVA multiple comparisons test with Šídák adjusted P values was used to compare the means and the data generated by this test are included in the table.*

Species	Mean 3' UTR similarity to human	Mean AUG dORF similarity to human	Mean Difference	Standard Error of Difference	95% Confidence Intervals of difference	3' UTR comparisons	AUG dORF comparisons	Adjusted P Value
P. troglodytes	86.32	93.17	6.843	0.1441	6.415 to 7.270	125456	11380	<0.0001
M. mulatta	81.92	86.3	4.383	0.1748	3.864 to 4.901	86637	7721	<0.0001
C. lupus	62.47	71.66	9.186	0.1995	8.594 to 9.777	62579	5961	<0.0001
B. taurus	61.5	69	7.503	0.2034	6.899 to 8.106	59532	5740	<0.0001
M. musculus	56.05	66.58	10.53	0.2124	9.902 to 11.16	54041	5266	<0.0001
R. norvegicus	55.46	64.96	9.5	0.2035	8.897 to 10.10	56133	5770	<0.0001
G. gallus	44.67	59.67	15	0.2598	14.23 to 15.77	36365	3520	<0.0001
X. tropicalis	42.23	56.36	14.14	0.2389	13.43 to 14.85	38765	4207	<0.0001
D. rerio	41.21	55.66	14.44	0.2459	13.71 to 15.17	36360	3974	<0.0001
A. gambiae	40.64	50.9	10.26	0.5795	8.542 to 11.98	6752	713	<0.0001
D. melanogaster	41.45	52.72	11.27	0.4927	9.804 to 12.73	9641	983	<0.0001
C. elegans	41.73	50.95	9.224	0.4768	7.809 to 10.64	9723	1056	<0.0001
S. pombe	42.63	52.57	9.934	0.6708	7.945 to 11.92	5240	530	<0.0001
M. oryzae	40.99	51.8	10.81	0.6429	8.906 to 12.72	5598	578	<0.0001
N. crassa	40.77	51.62	10.84	0.7592	8.592 to 13.10	4379	411	<0.0001
A. thaliana	41.81	53.16	11.34	0.5108	9.827 to 12.86	8929	915	<0.0001
O. sativa	41.5	52.98	11.48	0.5182	9.945 to 13.02	8615	890	<0.0001

Appendix 4: Violin plots of 3’ UTR, control sequences and AUG dORF similarity between human and homolog species for control sequences 500, 200 and 100 bases up and downstream of the AUG dORF

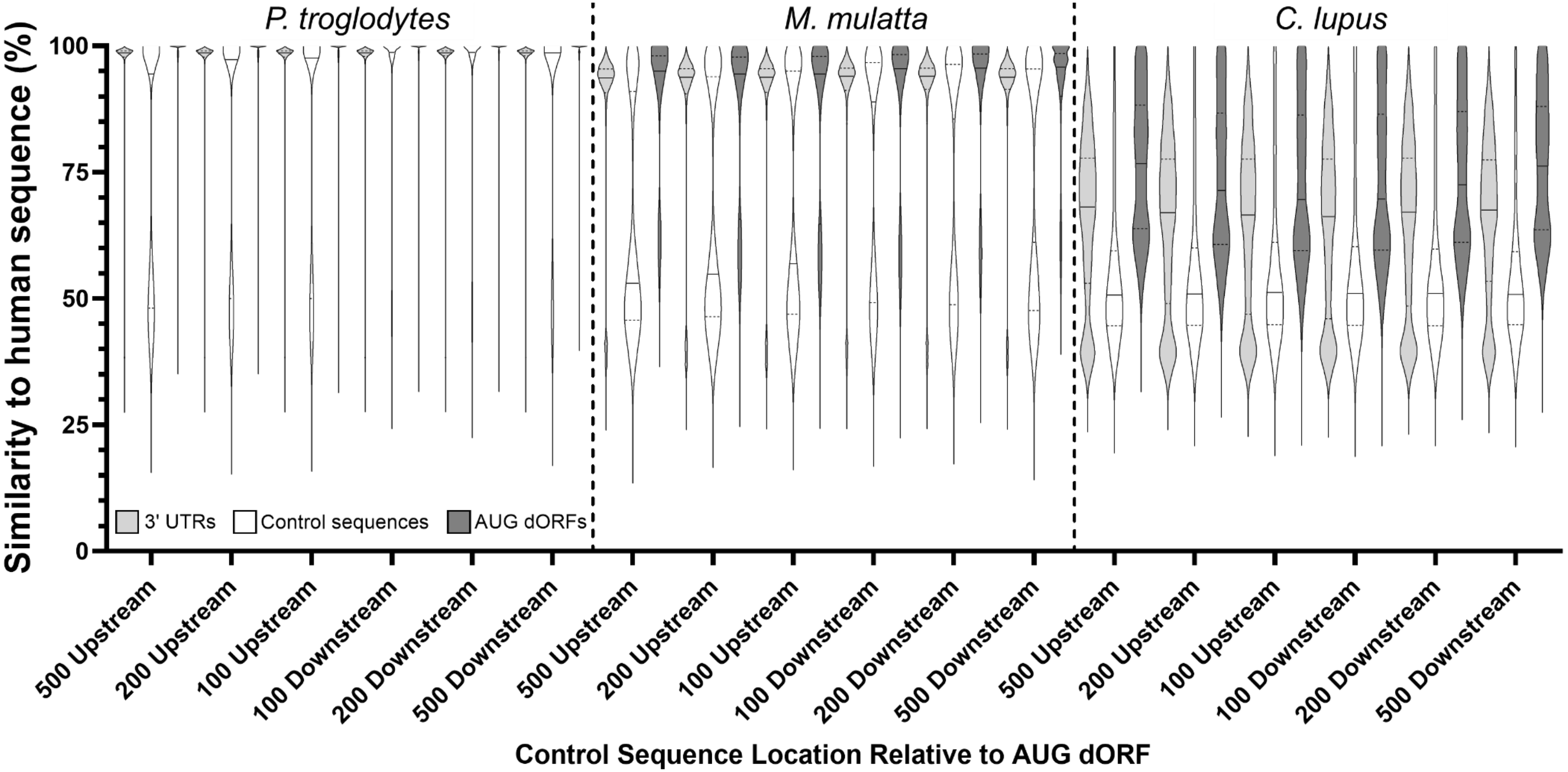

Figure 8.1: The similarity of downstream open reading frame (dORF) sequences which use an AUG start codon is greater than the 3' untranslated region (UTR) sequences containing these dORFs, and control sequences up and downstream of the AUG dORF between humans and P. troglodytes, M. mulatta, and C. lupus. *Control sequences are generated in the human and homolog 3' UTR sequences containing AUG dORFs and are sequences of the same length as the AUG dORF which end either, 500, 200, or 100 nucleotides upstream of the AUG dORF start codon in the 3' UTRs, or start 500, 200, or 100 nucleotides downstream of the AUG dORF stop codon in the 3' UTRs. The figure is split by dashed vertical lines into each species. The control sequence similarity percentage for each species is the similarity of the human control sequence compared against the homolog species control sequence. For comparison the mean similarity percentage between human and homolog species is included for the 3' UTR sequences containing the control sequence and the AUG dORF downstream of the control sequences. Violin plots, generated in GraphPad Prism (10.2.3), were used to show the distribution of the similarity percentages, the horizontal line within the violin marks the median and the dashed lines mark the 25th to the 75th percentile of the values distribution. Violin plots in light grey show the similarity percentages in the 3' UTRs, white for control sequences, and dark grey for the AUG dORFs.*

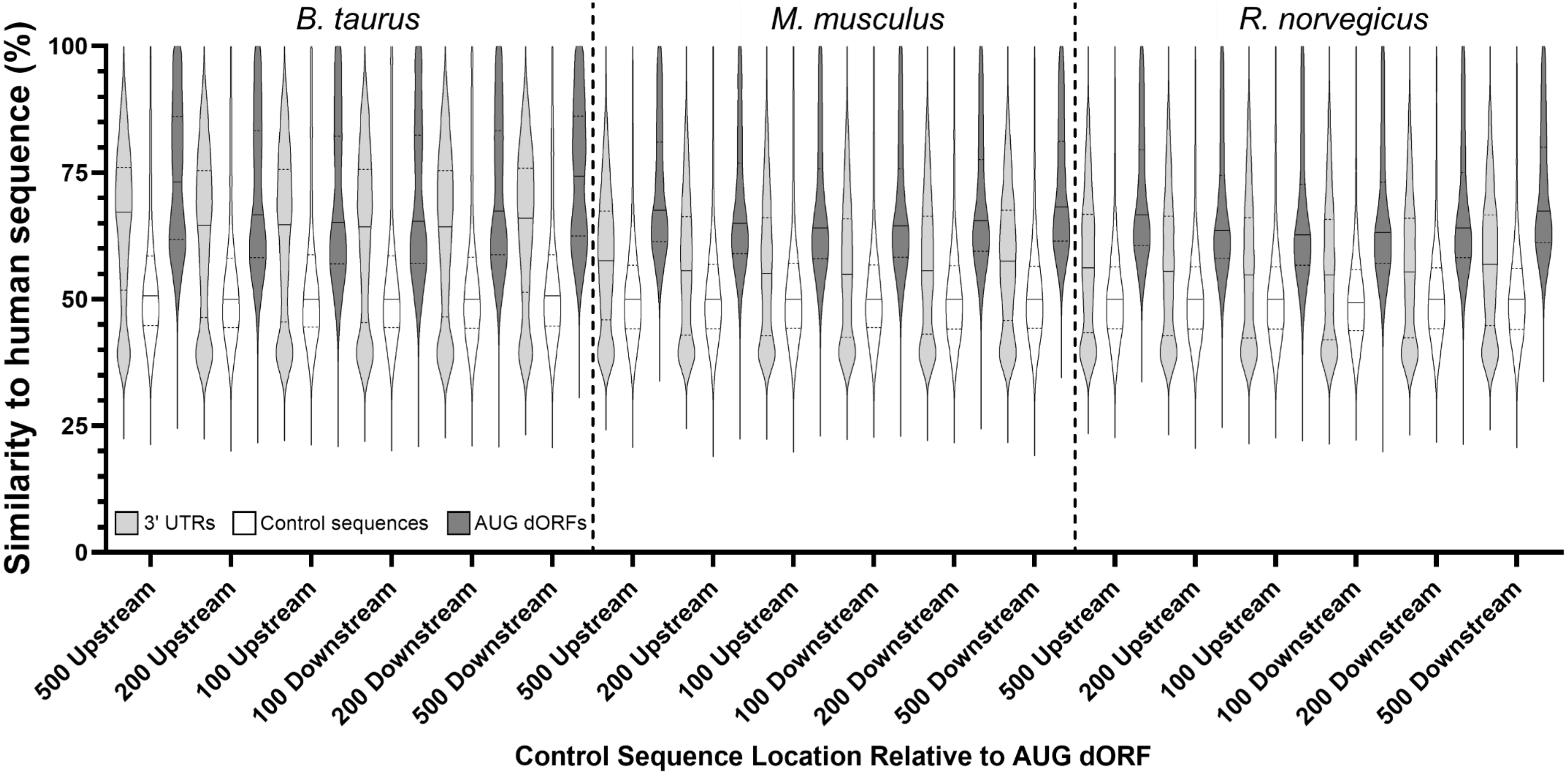

Figure 8.2: The similarity of downstream open reading frame (dORF) sequences which use an AUG start codon is greater than the 3' untranslated region (UTR) sequences containing these dORFs, and control sequences up and downstream of the AUG dORF between humans and B. taurus, M. musculus, and R. norvegicus. *Control sequences are generated in the human and homolog 3' UTR sequences containing AUG dORFs and are sequences of the same length as the AUG dORF which end either, 500, 200, or 100 nucleotides upstream of the AUG dORF start codon in the 3' UTRs, or start 500, 200, or 100 nucleotides downstream of the AUG dORF stop codon in the 3' UTRs. The figure is split by dashed vertical lines into each species. The control sequence similarity percentage for each species is the similarity of the human control sequence compared against the homolog species control sequence. For comparison the mean similarity percentage between human and homolog species is included for the 3' UTR sequences containing the control sequence and the AUG dORF downstream of the control sequences. Violin plots, generated in GraphPad Prism (10.2.3), were used to show the distribution of the similarity percentages, the horizontal line within the violin marks the median and the dashed lines mark the 25th to the 75th percentile of the values distribution. Violin plots in light grey show the similarity percentages in the 3' UTRs, white for control sequences, and dark grey for the AUG dORFs.*

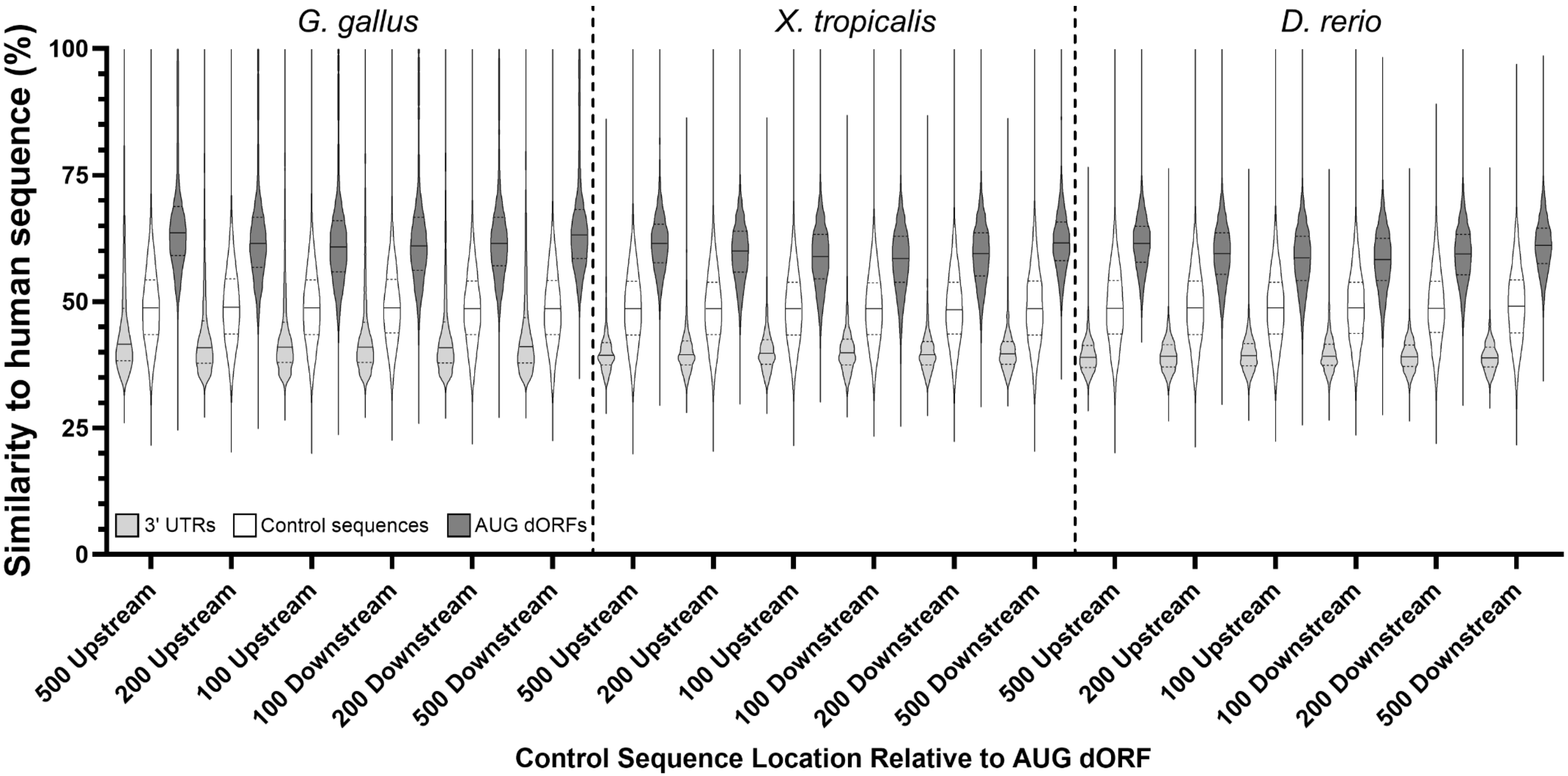

Figure 8.3: The similarity of downstream open reading frame (dORF) sequences which use an AUG start codon is greater than the 3' untranslated region (UTR) sequences containing these dORFs, and control sequences up and downstream of the AUG dORF between humans and G. gallus, X. tropicalis, and D. rerio. *Control sequences are generated in the human and homolog 3' UTR sequences containing AUG dORFs and are sequences of the same length as the AUG dORF which end either, 500, 200, or 100 nucleotides upstream of the AUG dORF start codon in the 3' UTRs, or start 500, 200, or 100 nucleotides downstream of the AUG dORF stop codon in the 3' UTRs. The figure is split by dashed vertical lines into each species. The control sequence similarity percentage for each species is the similarity of the human control sequence compared against the homolog species control sequence. For comparison the mean similarity percentage between human and homolog species is included for the 3' UTR sequences containing the control sequence and the AUG dORF downstream of the control sequences. Violin plots, generated in GraphPad Prism (10.2.3), were used to show the distribution of the similarity percentages, the horizontal line within the violin marks the median and the dashed lines mark the 25th to the 75th percentile of the values distribution. Violin plots in light grey show the similarity percentages in the 3' UTRs, white for control sequences, and dark grey for the AUG dORFs.*

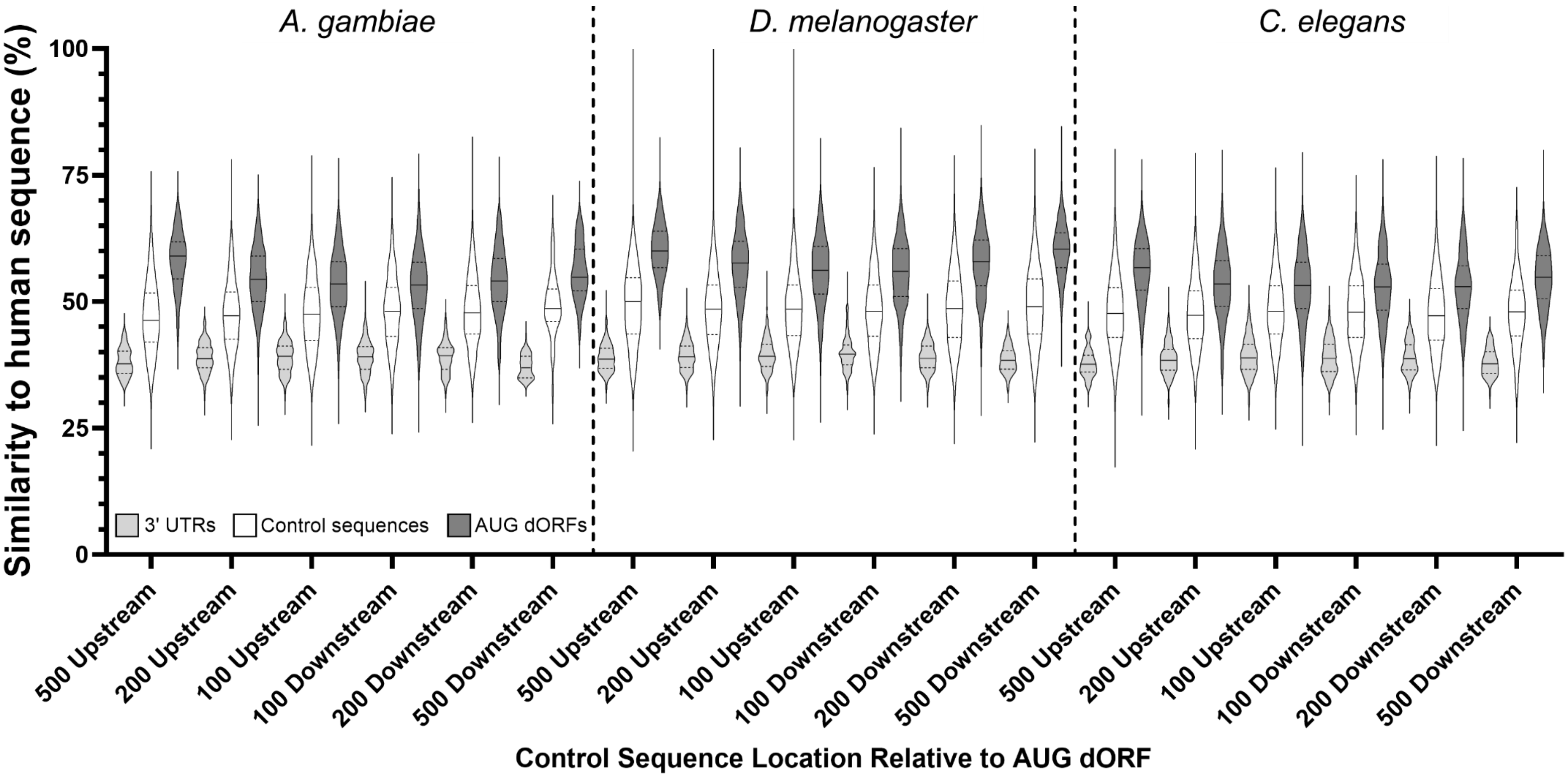

Figure 8.4: The similarity of downstream open reading frame (dORF) sequences which use an AUG start codon is greater than the 3' untranslated region (UTR) sequences containing these dORFs, and control sequences up and downstream of the AUG dORF between humans and A. gambiae, D. melanogaster, and C. elegans. *Control sequences are generated in the human and homolog 3' UTR sequences containing AUG dORFs and are sequences of the same length as the AUG dORF which end either, 500, 200, or 100 nucleotides upstream of the AUG dORF start codon in the 3' UTRs, or start 500, 200, or 100 nucleotides downstream of the AUG dORF stop codon in the 3' UTRs. The figure is split by dashed vertical lines into each species. The control sequence similarity percentage for each species is the similarity of the human control sequence compared against the homolog species control sequence. For comparison the mean similarity percentage between human and homolog species is included for the 3' UTR sequences containing the control sequence and the AUG dORF downstream of the control sequences. Violin plots, generated in GraphPad Prism (10.2.3), were used to show the distribution of the similarity percentages, the horizontal line within the violin marks the median and the dashed lines mark the 25th to the 75th percentile of the values distribution. Violin plots in light grey show the similarity percentages in the 3' UTRs, white for control sequences, and dark grey for the AUG dORFs.*

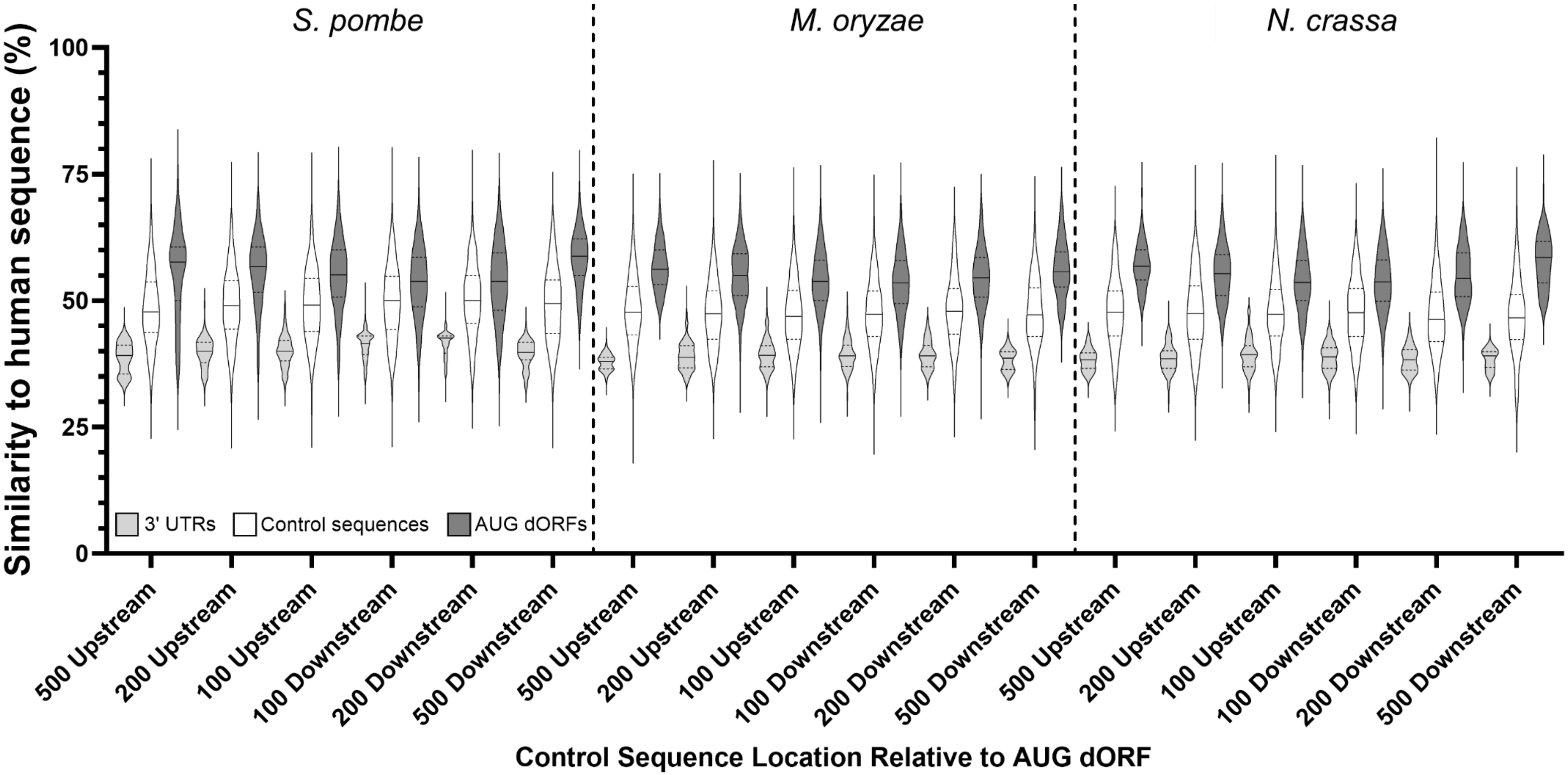

Figure 8.5: The similarity of downstream open reading frame (dORF) sequences which use an AUG start codon is greater than the 3' untranslated region (UTR) sequences containing these dORFs, and control sequences up and downstream of the AUG dORF between humans and S. pombe, M. oryzae, and N. crassa. *Control sequences are generated in the human and homolog 3' UTR sequences containing AUG dORFs and are sequences of the same length as the AUG dORF which end either, 500, 200, or 100 nucleotides upstream of the AUG dORF start codon in the 3' UTRs, or start 500, 200, or 100 nucleotides downstream of the AUG dORF stop codon in the 3' UTRs. The figure is split by dashed vertical lines into each species. The control sequence similarity percentage for each species is the similarity of the human control sequence compared against the homolog species control sequence. For comparison the mean similarity percentage between human and homolog species is included for the 3' UTR sequences containing the control sequence and the AUG dORF downstream of the control sequences. Violin plots, generated in GraphPad Prism (10.2.3), were used to show the distribution of the similarity percentages, the horizontal line within the violin marks the median and the dashed lines mark the 25th to the 75th percentile of the values distribution. Violin plots in light grey show the similarity percentages in the 3' UTRs, white for control sequences, and dark grey for the AUG dORFs.*

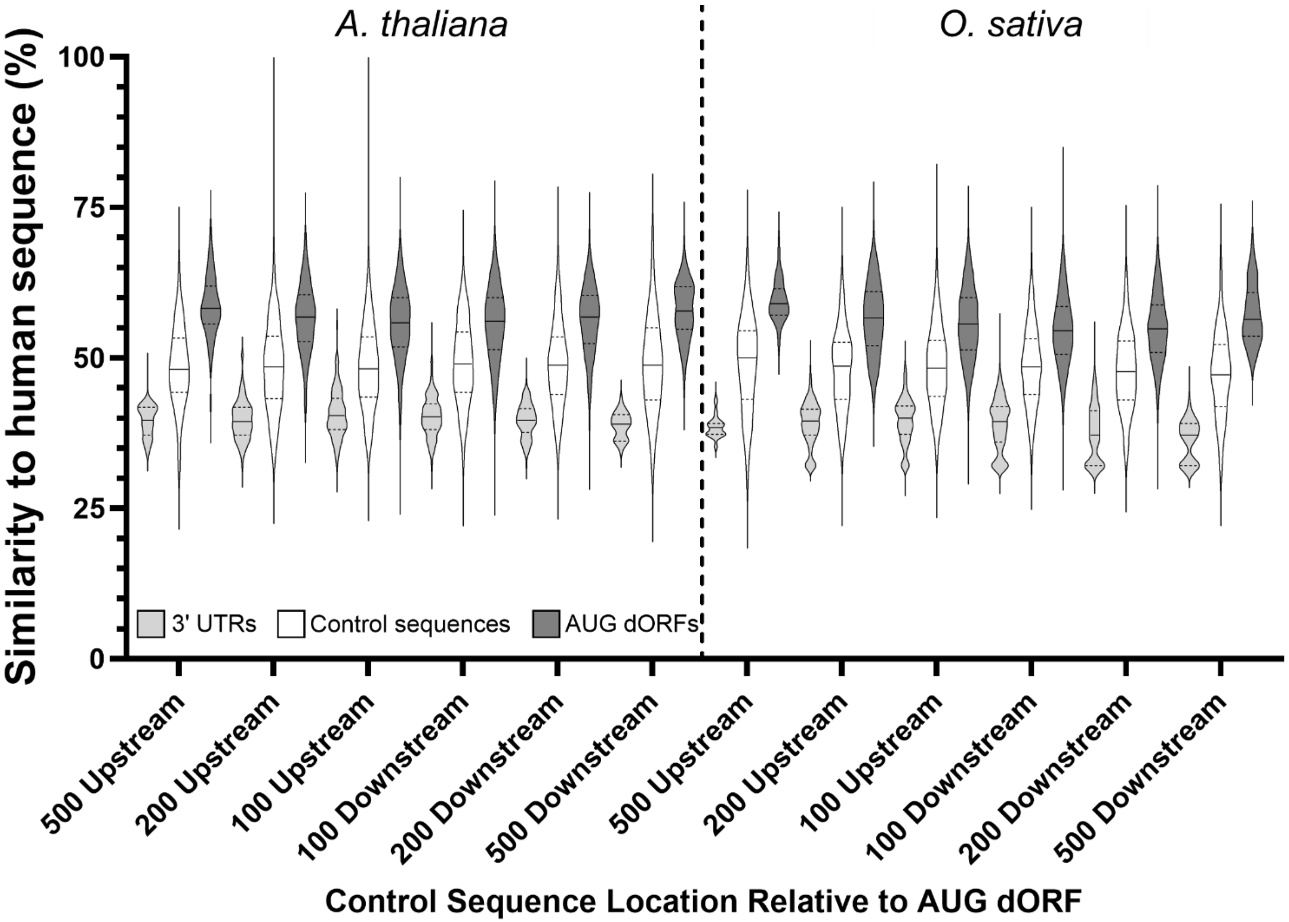

Figure 8.6: The similarity of downstream open reading frame (dORF) sequences which use an AUG start codon is greater than the 3' untranslated region (UTR) sequences containing these dORFs, and control sequences up and downstream of the AUG dORF between humans and A. thaliana and O. sativa. *Control sequences are generated in the human and homolog 3' UTR sequences containing AUG dORFs and are sequences of the same length as the AUG dORF which end either, 500, 200, or 100 nucleotides upstream of the AUG dORF start codon in the 3' UTRs, or start 500, 200, or 100 nucleotides downstream of the AUG dORF stop codon in the 3' UTRs. The figure is split by dashed vertical lines into each species. The control sequence similarity percentage for each species is the similarity of the human control sequence compared against the homolog species control sequence. For comparison the mean similarity percentage between human and homolog species is included for the 3' UTR sequences containing the control sequence and the AUG dORF downstream of the control sequences. Violin plots, generated in GraphPad Prism (10.2.3), were used to show the distribution of the similarity percentages, the horizontal line within the violin marks the median and the dashed lines mark the 25th to the 75th percentile of the values distribution. Violin plots in light grey show the similarity percentages in the 3' UTRs, white for control sequences, and dark grey for the AUG dORFs.*

Appendix 5: PIP Placement Reflection

The Nottingham Festival of Science and Curiosity (FOSAC) is an annual event which brings science, technology, engineering and maths into the community. The festival encourages sharing of knowledge and celebrates curiosity within Nottinghamshire. This project intended to evaluate the impact of FOSAC. The project aimed to provide a narrative of the festival and describe how the festival takes shape and would collate the data generated from previous festivals to allow evaluation of the growth of the festival over previous years. This growth would be assessed in a variety of ways, ranging from the audience attendance to the spread of the festival and the groups involved. The project then intended to consider the impact of the festival beyond the quantitative statistics around the growth of the festival. Thematic analysis through semi-structured interviews would help to understand the impact of the festival on individuals and companies involved with the festival. For comparison, interviews with individuals/companies that are not involved in the festival were planned too. This project was intended to culminate in multi-format impact reports to ensure that the results and discoveries were accessible to the public and partners alike. The goals of this project included, evaluating the quantitative data gathered from previous editions of FOSAC and producing a report on the impact of FOSAC involvement on those previously involved with the festival during the course of the placement. Other intended outcomes of the placement included interaction with the UK Science Festival Network Conference and possible presentation at their conference, contact and networking with a range of companies and institutes, and development of a range of skills relating to ethics application, evaluation, interviews, report writing, communication of qualitative data.

Before starting the placement, I worked alongside my placement supervisors to complete an ethics application for the proposed research involving interviews. This was a useful experience and something I hadn’t done previously. Production of a project proposal, data management plan, ethics application form, participant information form, and consent form, developed my project management and organisation, whilst improving my understanding of ethics processes. Although submitted prior to the placement, ethical approval took some time to come through. This alongside the ethics committee recommendations changed to project to no longer carry out interviews with those not previously involved with FOSAC. The qualitative research consisting of semi-structured interviews was my first time carrying out this type of research. I enjoyed developing the interview structure and carrying out the interviews. I completed some training relating to interview technique, took advice from my supervisors, and used feedback from a mock interview. This project really developed my interview and communication skills. The major challenge during this project was recruiting candidates for the interviews. Despite advice and support from supervisors the project failed to recruit as many participants as originally planned. Although this was disappointing it was an opportunity to develop problem solving skills and led to some adaption of the project to still make it worthwhile. To gather further data on the impacts of FOSAC I also ran a workshop with the FOSAC steering group. This was another opportunity to work on my communication skills and a chance to carry out a different type of event and presentation. Thematic analysis, and analysis of transcripts more generally, was another skill that I could develop in this project. Literature search and guidance from supervisors helped me to initially understand and then carry out thematic analysis. The smaller sample did mean that my initial there was a manageable amount of data to work with but did make drawing conclusions and carrying out thematic analysis more limited. Writing up a report was slightly different to the type of reports I usually complete and in addition to being a good opportunity to practice writing up, feedback from several people helped me to target my writing to my audience, in this case the FOSAC board.

Aside from the impact report and interviews I also had the opportunity to develop my data analysis and presentation skills, through working with previous FOSAC data. This was a good opportunity to work with a range of different types of data. I worked with survey data through to geographic and deprivation data. I thoroughly enjoyed the data analysis, from pulling things together to producing meaningful charts and graphics. I had support from the team I worked closely with at Ignite!, the organisation that produces FOSAC, in data discovery, making all their FOSAC data accessible to me. It was also interesting to report on data for a more public audience, rather than a scientific one, this helped to improve my ability to report for different audiences. I also worked with Ignite! to produce short data reports for them for all of the other programmes that they do, with a focus on their school programmes. I enjoyed producing these targeted reports which Ignite! could use to demonstrate the importance of their school programmes. My work on the data from previous festivals helped me to understand the evaluation process and highlight any issues that had arisen. I worked with the Ignite! team to develop a plan for FOSAC 2023 evaluation and planned the survey forms. This project gave me the chance to present to a variety of different people and be involved in some meetings to expand my network. These included the UK Science Festival Network, STEMCity, and the FOSAC board meeting. Ignite! also produce a magazine to go alongside the festival, I enjoyed working with a different media to produce a couple of activities alongside some short articles based on genetics. This was challenging but useful to work on something so different. I also had the opportunity to do some engagement work. I attended a local school to get feedback on magazine articles, including my own, which was very informative. I also attended a couple of Creative Sparks sessions, working with primary school children at a library in St. Anns on creative activities. This was an opportunity to develop my engagement skills, especially working with children.