Characterisation of translated downstream Open Reading Frames in human cells

A Thesis

submitted to

Indian Institute of Science Education and Research Pune
in partial fulfilment of the requirements for the BS-MS Dual Degree Programme

by

Pritam Pathak

20191017

Indian Institute of Science Education and Research Pune

Dr. Homi Bhabha Road

Pashan, Pune 411008, INDIA

Date: April 2024

Under the guidance of

Supervisor: Dr. Ariel Bazzini

Associate Investigator

Stowers Institute for Medical Research, Kansas City, USA

From June 2023 to April 2024

Certificate

This is to certify that this dissertation entitled "**Characterisation of translated downstream Open Reading Frames in Human Cells"** towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune, represents the work carried out by **Pritam Pathak** at the Stowers Institute for Medical Research, USA under the supervision of **Dr Ariel Bazzini**, Associate Investigator, during the academic year 2023-2024.

Ariel Bazzini

Dr Ariel Bazzini

Associate Investigator

Stowers Institute for Medical Research

Thesis Advisory Committee (TAC)

Supervisor : Dr Ariel Bazzini

Thesis Advisor: Dr Mayurika Lahiri

This thesis is dedicated to

My 'Gurus'

Declaration

I hereby declare that the matter embodied in the report entitled "**Characterisation of translated downstream Open Reading Frames in Human Cells"** are the results of the work carried out by me at the Stowers Institute for Medical Research, under the supervision of Dr Ariel Bazzini and the same has not been submitted elsewhere for any other degree. Wherever others contribute, every effort is made to indicate this clearly, with due reference to the literature and acknowledgement of collaborative research and discussions.

Pritam Pathak

Pritam Pathak

20191017

Date : 15 March 2024

Table of Contents

Certificate ---- *3*

Declaration ---- *6*

List of Figures ---- *9*

Abstract ---- *10*

Abbreviations ---- *11*

Acknowledgements ---- *12*

Contributions ---- *14*

Chapter 1 Introduction ---- *15*

1.1 Ribosomes and their functions ---- 16

1.2 Open Reading Frames (ORF) ---- 17

1.3 Untranslated Regions (UTRs) ---- 18

1.4 Small Open Reading Frames ---- 20

1.5 Ribosome Profiling ---- 21

1.6 Upstream Open Reading Frames ---- 23

1.7 Downstream Open Reading Frames ---- 24

Chapter 2 ---- *27*

2.1 Materials ---- 27

2.2 Methods ---- 28

2.2.1 Tissue culture ---- 28

2.2.2 Cloning ---- 28

2.2.3 Transfection ---- 29

2.2.4 Cytometric Analysis ---- 29

2.2.5 Data analysis and plotting ---- 30

Chapter 3 Results ---- *31*

3.1 The effect of the length of the downstream Open Reading Frames ---- 31

3.1.1 Enhancement of the main Open Reading Frame translation is dependent on the length of the dORF ---- 31
3.1.2 The length dependence of the main ORF enhancement is independent of the iUTR. ---- 33
3.1.3 dORFs less than 6AA show possible translation enhancement ---- 35
3.1.4 Identification of short downstream Open Reading Frames ---- 35
3.2 The effect of the length of the 3'UTR on the enhancement effect of dORFs on main ORFs ---- 37
3.2.1 Enhancement of the main Open Reading Frame translation is independent of the 3'UTR length ---- 37
3.3 The effect of the iUTR length on the enhancement effect of dORFs on main ORFs ---- 39
3.3.1 Enhancement of the main Open Reading Frame is dependent on the internal UTR length. ---- 39
3.4 Validation of experimental results using *in silico* data ---- 41
Chapter 4 Discussion ---- 43
References ---- 48
Appendix ---- 53
Supplementary Figures ---- 53
Primers ---- 55

List of Figures

Figure No	Title	Page No
Figure 1	A graphical representation of central dogma in eukaryotes.	15
Figure 2	A basic structure of an eukaryotic mRNA.	17
Figure 3	Graphical representation of ribosome profiling.	22
Figure 4	uORFs repress translation of the canonical ORF.	23
Figure 5	dORFs enhance the translation of the canonical ORF.	25
Figure 6	Enhancement of the main Open Reading Frame is dependent on the length of the dORF.	32
Figure 7	Enhancement of translation of main ORF by dORF is iUTR-independent	34
Figure 8	Identification of short downstream Open Reading Frames.	36
Figure 9	Enhancement of the main Open Reading Frame translation is independent of the 3'UTR length.	38
Figure 10	Enhancement of the main Open Reading Frame is dependent on the internal UTR length.	40
Figure 11	*In silico* data supports experimental findings of 3'UTR and iUTR lengths	42
Figure 12	Schematic highlighting translation of dORFs.	43
Figure 13	Schematic depicting proposed hypothesis of main ORF translation enhancement by dORF.	45
Figure 14	Schematic depicting multi-faceted approaches and future perspectives based on this thesis or in relation to dORF in general.	46
Supp Fig 1	Schematic explaining the experimental validation of whether long iUTRs get translated or not.	53
Supp Fig 2	Schematic explaining the experimental validation of checking whether short dORFs get translated.	54

Abstract

This study delves into the characterisation of translated downstream Open Reading Frames (dORFs) in human cells, exploring their role in post-transcriptional gene regulation. dORFs are small ORFs found in the 3'UTR region of an mRNA and have been shown to enhance the expression of its associated main ORF when it is translated in human cells and zebrafish. dORFs as a novel post-transcriptional regulator that works contrary to upstream ORFs (uORF) was established fairly recently. Moreover, we also know that although the presence of dORFs is conserved, the amino acid sequence is not in Orthologous genes in humans and zebrafish, suggesting evolutionary conservation of dORFs. However, how the various structural aspects like dORF length, the 3'UTR length and the iUTR length affect its enhancement activity has not been known until now.

Employing a fluorescent reporter expression system, we characterised dORFs to be able to propose how its functional components affect the enhancement ability of the associated canonical open reading frame. Our results demonstrate that the length of dORFs plays a crucial role in modulating translation efficiency, with the presence of an optimal length that exhibits greater enhancement of main ORF expression. We also found that dORFs shorter than 6AA cannot enhance the translation of the main ORF. Additionally, we found the influence of 3'UTR and internal UTR (iUTR) length on dORF-mediated translation enhancement, revealing a dependency on iUTR length but not 3'UTR length for optimal dORF function. With elongated iUTRs, the dORFs lose their capability to enhance the expression of the canonical ORF. These findings provide novel insights into the molecular mechanisms governing post-transcriptional gene regulation, highlighting the diverse roles of dORFs in fine-tuning protein expression levels. Our study lays the groundwork for further investigations into the functional significance of dORFs and their implications for cellular physiology and disease pathogenesis.

Abbreviations

- **ORF** : Open Reading Frame
- **dORF** : Downstream Open Reading Frame
- **uORF** : Upstream Open Reading Frame
- **sORF** : Small Open Reading Frames
- i**UTR** : Internal Untranslated Region
- **tRNA** : transfer RNA
- **mRNA** : messenger RNA
- **rRNA** : ribosomal RNA
- **CCDC 167** : Coiled-Coil domain-containing Protein 167
- **CENP-A** : Centromere Protein A
- **IRES** : Internal Ribosome Entry Site

Acknowledgements

I believe this thesis work is not only the product of my work carried out in the past one year at the Stowers Institute in Kansas City. It rather started very early from my childhood and later, when I joined Ramakrishna Mission Vidyapith Purulia in class V. I have been fortunate enough to have great teachers throughout my life until now who have mentored me and shaped me into the person I am today. They have come in different roles that were not necessarily of traditional teachers. It will be a herculean job to mention them all. But I will try.

My maternal grandfather instilled within me the passion for Biology that I have carried on till now. I have been greatly inspired by him since childhood. My school was a residential one. I have spent six years, a significant part of my childhood, in that place that became my second home. That place was full of exemplary teachers and monks. I have been fortunate to learn from them, and it would be disrespectful to others if I named some. They have taught me life. However, I must say that I owe a lot to Arijit Maharaj. He has taught me leading by example. The friends I got there are a treasure. It is like a brotherhood, and I need not name any specific one among the 89.

I will always be indebted to Mr Anup Samanta. He agreed to take me in and contact all the necessary teachers for my higher secondary education and convinced them to teach me for free or at a minimum cost. I would not be where I am without him and the support from those teachers. Saptarshi da, Biswajit Sir, Biswanath Sir, Bagchi Sir, Rudrajit Sir, everyone had gone beyond their capacities to help in every ways possible.

IISER had been a place of reckoning. It was far from home and full of the best and the brightest from the country. Joining IISER was my choice and my dream, and when I was selected for Pune, I never had a second thought. Well, not everything goes as per plan and COVID hit. After the dust settled and we came back to campus, I joined Dr Mayurika Lahiri's lab. She has been a constant guide in every aspect of life. I learned all the basics of research in that lab and am grateful to all the past lab members, especially Ben and Abhijith.

I have made some friends that have become part of my life. Sayandeep, Ranojoy and Hritwik have been part of my journey and became integral to it. Vikas, Ritvee, Amisha, and Shruti all came to Stowers the same as me for their masters, and I shared my journey with them. I have met some fantastic people here as well. Gopal, Amruta, Kurne, Tolka, we seldom realise how important a role some people play in a very short stint of our lives. I knew Shruti before, but she has been a constant companion since coming here, providing motivation and insights when I did not realise I needed them. I cannot possibly express here all that I want to. Thanks !

Coming to the US for the thesis was a big step. I would like to thank the Stowers Institute and IISER, Pune for the opportunity and SHRM for the VISA sponsorship. The whole of Stowers is like a family. More so is my lab. The lab has made me one of their own in a very short span. I never felt isolated for a speck of time. They have helped me whenever I needed and in whatever ways possible.

As I stated earlier, I have had great teachers all throughout. Ariel is a person who made me believe that someone could possibly be this excited about science. He is always enthusiastic

with the tiniest of ideas and is ready to help. He is usually the first one to the lab and is always available if needed. I owe him a lot and to the whole lab. From the beginning, Majo, Eugenia, and Cameron helped me familiarise myself with the various techniques. Sergio and I started together on very related projects, often collaborating, starting from very minor things. I owe a lot to everyone, =Dani, Gabe, Luli, Michelle, Daniel, Josefina, Catalina, Gopal, AJ.

Sara, our AA, helped me with all the administrative work. I did not know official work could be this smooth, efficient and thorough. The core teams of the institute have been phenomenal. I believe the whole of the institute is what it is because of the technology centres. My special thanks to Cytometry, CTOC, Sequencing and Microscopy for all the help and guidance.

I wanted to save the end for my family. My mother has always been supportive of my decisions. She has believed me when no one else would and has taught me what no one else could with her actions and choices. My father has provided me with the best with what he could and often going beyond his capacities for me. I hope they know that I am ever grateful and thankful to them and that nothing would have ever been possible without them. My whole extended family has protected me whenever necessary and showered me with love, compassion and blessings.

I know that this acknowledgement still misses a lot of names and cannot possibly contain all of them. But this is my try with my limited capability. I thank everyone who has helped me directly or indirectly, whenever possible, in any possible ways.

Contributions

Contributor name	Contributor role
Bazzini lab	Conceptualisation Ideas
Pritam and Bazzini lab members	Methodology
Pritam Pathak	Software
Pritam Pathak	Validation
Pritam Pathak	Formal analysis
Pritam and Bazzini lab members	Investigation
Pritam and Bazzini lab members	Resources
-	Data Curation
Pritam Pathak	Writing - original draft preparation
Pritam and Ariel Bazzini	Writing - review and editing
Pritam Pathak	Visualisation
Ariel Bazzini	Supervision
Ariel Bazzini	Project administration
Ariel Bazzini	Funding acquisition

Chapter 1

Introduction

The central dogma of biology is a fundamental concept in the field of molecular biology that talks about the basic flow of genetic information within biological systems. The central dogma describes the unidirectional flow of genetic information from DNA to RNA to protein (Cobb, 2017) (Fig1). DNA serves as the repository of genetic instructions within cells, encoding the information necessary for synthesising proteins and regulating cellular processes. Through the process of DNA replication, genetic material is duplicated, which ensures that genetic information is transmitted to daughter cells during cell division(Cobb, 2017)(CRICK, 1970).

Transcription, which is the next step in the central dogma, involves the synthesis of ribonucleic acid (RNA) molecules from DNA templates. This process is catalysed by the enzyme RNA polymerase, which reads the DNA sequence and generates complementary RNA molecules through base-pairing interactions. The resulting RNA transcripts, including transfer RNA (tRNA), messenger RNA (mRNA), and ribosomal RNA (rRNA), serve as intermediaries which convey genetic instructions from the nucleus to the cytoplasm, where protein synthesis occurs(Chatterjee et al., 2021; Webster, 2021).

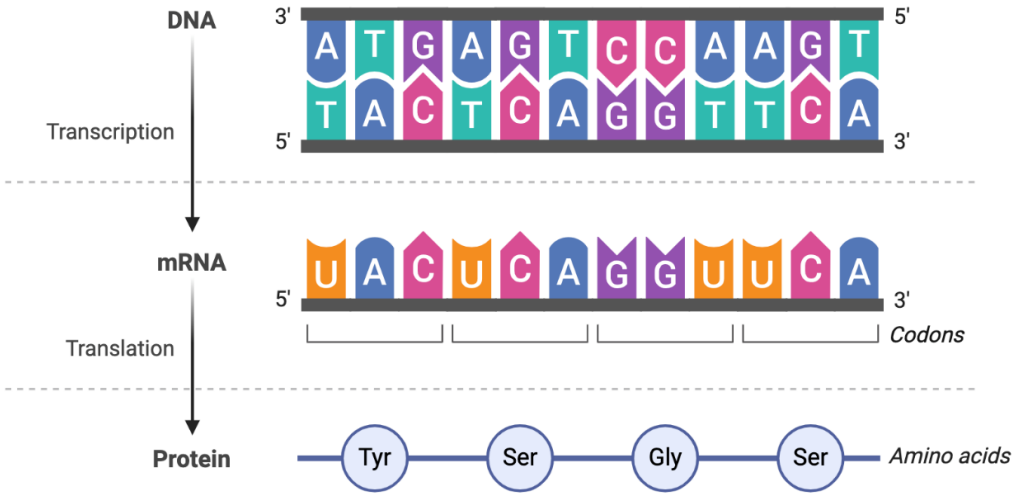

Fig 1: A graphical representation of central dogma in eukaryotes as it is typically understood. A single DNA sequence produces a single mRNA through

transcription, which in turn produces protein through the process of translation. (Image from BioRender)

Translation, the central dogma's final stage, entails converting RNA sequences into amino acid sequences, the building blocks of proteins. Ribosomes, complex molecular machines composed of RNA and proteins, facilitate this process by decoding mRNA sequences and assembling amino acids into polypeptide chains according to the genetic code. Proteins are synthesised with precise amino acid sequences that dictate their structure and function through the coordinated action of transfer RNA molecules, which recognise specific codons on the mRNA and deliver the corresponding amino acids (Alberts B, 2002).

1.1 Ribosomes and their functions

The ribosome is basically the cellular machinery necessary for protein production. It is an assembly of proteins and RNA. Ribosomes are comprised of two subunits that scan through the messenger RNA and read them. It also serves as a docking station of tRNA, which produces the peptide sequence or proteins. Proteins are key functional outputs of the transcriptomes. Proteins' identity, amount and activity are the final determinants of how cells and tissue function. Historically, fluctuations in gene expression have been understood through the lens of transcriptional and post-transcriptional regulation, which are mediated by the 5' and 3' Untranslated Regions (UTRs). (Opron & Burton, 2019; Wu & Bazzini, 2018)

As generally understood, the most relevant function of the ribosome is translating mRNA sequences into proteins. However, the ribosome also has a fundamental role as the mRNA quality checker through translation (Shoemaker & Green, 2012). Various mRNA quality control mechanisms are facilitated by the ribosome. Nonsense-mediated decay (NMD) targets mRNA containing premature termination codons (PTC), non-stop mediated decay (NSD) eliminates mRNA lacking a stop codon, and no-go decay (NGD) is associated with degradation of mRNA with stalled ribosomes (Wu & Bazzini, 2018). These pathways rely on translation activity to identify abnormal mRNA, playing a crucial role in preventing the production of aberrant proteins and aiding in ribosome rescue. (Radhakrishnan & Green, 2016).

In addition to these functions related to quality control, ribosomes can also regulate gene expression for the normally processed mRNAs for either mRNA stability or translation efficiency. Translation impacts the stability of properly processed mRNA through codon composition, introducing a concept known as codon optimality. This mechanism adds another dimension to the traditional understanding of mRNA translation, expanding its biological implications. (Bazzini et al., 2016; Mishima & Tomari, 2016; Wu & Bazzini, 2023). Ribosomes could also regulate gene expression in a codon-independent manner. For example, the translation of upstream ORF (uORF, small ORF in 5'UTR) represses the translation of the main CDS across species (Kute et al., 2022). uORF regulation is modulated by the translation activity itself, not through the encoded peptide (Johnstone et al., 2016).

1.2 Open Reading Frames (ORF)

Open Reading Frames are defined as a series of nucleotides that have a start codon followed by a downstream in-frame stop codon (Kute et al., 2022; Sieber et al., 2018). ORFs are read in a specific reading frame, meaning the nucleotide sequence is divided into sets of three consecutive bases (codons) without any gaps or overlaps (Kute et al., 2022; Mignone et al., 2002a). The correct reading frame is crucial for accurate protein synthesis. They then terminate with the stop-codon. In eukaryotic messenger RNAs, the primary protein-coding sequence (CDS) usually consists of a single main open reading frame (ORF). Although the CDS is typically the longest ORF within the mRNA, numerous shorter ORFs are frequently found in the transcript, some of which have the potential to undergo translation. (Fig2).

Although the mere existence of an ORF does not guarantee the synthesis of protein from the genetic sequence, it is often the case. ORFs, therefore, make it easier to predict protein-coding sequences from a genetic sequence for genomics studies.

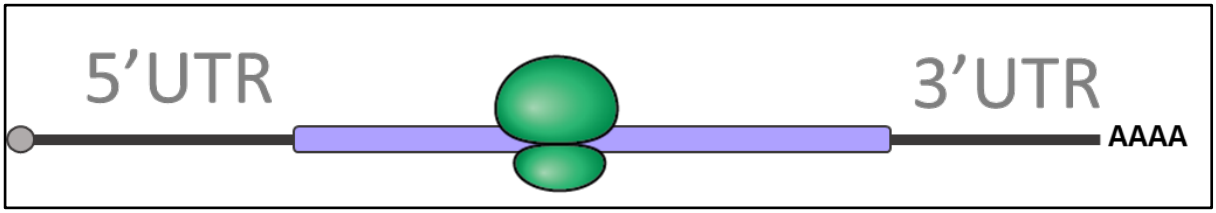

Fig 2: A basic structure of an eukaryotic mRNA. It is thought to contain one ORF that codes for a single peptide. It is flanked by untranslated regions (UTRs) on both its 5' and 3' ends (Venter et al., 2001).

However, the concept that a single protein is encoded by a single mRNA has undergone revisions (Venter et al., 2001). Genome-wide studies revealed the presence of many unconventional small ORFs (sORF) in a variety of transcripts (Xiao et al., 2018). Some of them even lie in regions previously annotated as untranslated regions (UTRs) that were presumed to be non-coding transcripts (Johnstone et al., 2016; Kute et al., 2022; Mignone et al., 2002).

The exploration of ORFs has unveiled a fascinating landscape of gene expression, evolutionary dynamics, and functional genomics, offering profound insights into the complexity and diversity of living organisms. The discovery of non-canonical ORFs, such as small Open Reading Frames (sORFs), has challenged traditional paradigms of gene expression and expanded our conceptual framework of genetic regulation and protein diversity. The functional implications of ORFs extend far beyond their role as mere templates for protein synthesis. Emerging evidence suggests that ORFs play multifaceted roles in gene regulation, RNA metabolism, and cellular signalling pathways. We have evidence of the presence of a plethora of non-coding RNAs encoded within ORFs that have diverse roles, including RNA interference, riboswitches, and RNA editing (Chatterjee et al., 2021).

Moreover, ORFs have been implicated in the generation of bioactive peptides and small proteins, which modulate a wide array of physiological processes, from development and immunity to stress responses and metabolism (Albuquerque et al., 2015).

1.3 Untranslated Regions (UTRs)

As per the analysis of the Human Genome and other higher eukaryotes, only a small fraction of the human genome, approximately 1.5% of it, codes for protein (Venter et al., 2001). The remaining, which is the major part, is involved in regulation at either the transcriptional or post-transcriptional level (Mignone et al., 2002).

Various transcriptional regulators are at play, resulting in the formation of a mature mRNA. It generally consists of a tripartite structure with 5’ and 3’ untranslated regions (UTRs) and a series of triplet coding sequences in between (Kute et al., 2022; Venter et al., 2001).

Untranslated Regions (UTRs) are major post-transcriptional regulators. They modulate mRNA transport and regulate the efficiency of translation, subcellular localisation and stability of the mRNA (Mignone et al., 2002).

UTRs can also vary highly among different transcripts of the same gene. Alternative splicing and alternative polyadenylation sites can give rise to multiple mRNA variants with distinct UTRs. These UTR variations can impact how a gene responds to different cellular conditions or developmental stages. Adopting such a protein-centric perspective, it was unexpected to discover comparable numbers of protein-coding genes in the human genome and in comparatively simpler eukaryotic organisms (Lander et al., 2001; Mayr, 2017). Furthermore, there is substantial uniformity in protein size across various organisms (Milo & Phillips, 2015). This underscores the significant conservation of protein sequences and highlights the considerable constraints imposed on proteins. Nevertheless, it prompts inquiry into the factors facilitating the biological complexity observed in higher organisms.

As mentioned, there are UTRs at both the 5’ and 3’ end of the mRNA. They differ in their roles and how they regulate post-transcriptional modifications. Although named as untranslated regions, small translated open reading frames (sORF) were found quite some time back in the 5’ UTR. Those open reading frames are called upstream open reading frames (uORFs), and they function by repressing the translation of the canonical open reading frame (Barbosa et al., 2013; von Arnim et al., 2014). However, the discovery of those led to the modification in the concept and proved that there are indeed elements in the UTRs that get translated (Xiao et al., 2018).

1.3.1 5’ UTR

5’ UTR refers to the sequence upstream of the canonical ORF. It was long thought to be untranslated and lacking any function. However, we have found different regulatory roles for these regions (Mignone et al., 2002). The 5’ untranslated region (UTR) harbours numerous regulatory elements, including small open reading frames (ORFs), internal ribosome entry sites (Kieft, 2008), microRNA binding sites, and structural components crucial for modulating mRNA stability, pre-mRNA splicing, and translation initiation. Disruption of cis-regulatory elements or secondary structures within the 5′UTRs can lead to alterations in gene expression (Johnstone et al., 2016), underscoring the functional significance of the 5′UTR in gene expression regulation. Moreover, mounting evidence indicates that mutations within 5′UTRs frequently correlate with diseases, including cancer (Prensner et al., 2021; Ryczek et al., 2023).

1.3.2 3’ UTR

3’UTR is the denoted region downstream of the stop codon of the main ORF. Like 5’UTR, it was also thought to be untranslated and lacking any function. With the advent of technology, new suitable and robust methods helped study the sequences in the 3’UTR regions and decipher their functions (Mignone et al., 2002). The role of 3’UTR in sub-cellular localisation and mRNA stability has been known for some time now (Mayr, 2019). However, the involvement of 3’UTR in post-transcriptional regulation is gradually becoming better understood (Kurosaki & Maquat, 2013). These roles are mediated by RNA binding proteins (RBP) and microRNAs. Moreover, 3’UTRs are also highly involved in polyadenylation and cleavage that generate mRNA isoforms, which differ only in their 3’UTRs and are major gene regulatory elements (Mishima & Tomari, 2016). Fairly recently, the presence of small ORFs was also found in the 3’UTR region as well, like the 5’UTR, signifying important unknown roles for this region and sequences.

1.4 Small Open Reading Frames

Simply put, small Open Reading Frames (sORFs) are such open reading frames with more than ten amino acids but less than 100 amino acids (10AA $\leq$ sORF $\leq$ 100AA) (Kute et al., 2022; NIH - NHGRI, 2024). sORFs play a very diverse role in genetic regulation. They can either produce functional micro-peptides or have regulatory post-transcriptional roles (Couso & Patraquim, 2017). sORFs can be located in either coding transcripts (5’ UTR, 3’UTR, CDS) or non-coding transcripts like lncRNAs and mitochondrial RNAs. Due to their high abundance and small sizes, it is extremely difficult to identify and annotate sORF in the genome (Mackowiak et al., 2015). Nevertheless, numerous studies have highlighted the significance of small open reading frames (sORFs) in various cellular processes and the regulation of translation within the coding sequence (CDS).

The small open reading frames found in 5’UTR and 3’UTR are majorly involved in regulatory functions (Chen et al., 2020). The sORFs found in 5’UTRs are known as upstream open reading frames (uORFs). Approximately half of the human coding transcripts naturally contain uORFs that repress the translation of the canonical open reading frame (Barbosa et al., 2013; Johnstone et al., 2016).

The presence of small open reading frames was found in the 3’UTR as well. These are called downstream open reading frames (dORFs). The dORFs, however, function

opposite to the uORF and enhance the translation of the canonical ORF (Wu et al., 2020).

Modern high-throughput detection techniques have made it possible to distinguish and identify the presence of sORFs throughout the genome and, in turn, made it possible to discover more of their roles (Chen et al., 2020).

1.5 Ribosome Profiling

Ribosome profiling, also called Ribo-seq, is a cutting-edge technique that enables the comprehensive investigation of translation dynamics across the transcriptome at single-nucleotide resolution. This method relies on the deep sequencing of ribosome-protected mRNA fragments, providing a snapshot of the actively translating ribosomes and elucidating the spatial and temporal aspects of protein synthesis within cells (Ingolia et al., 2012; Koehbach & Jackson, 2015).

To conduct ribosome profiling, cells are treated with a translation inhibitor to halt ribosome movement, thereby preserving the position of ribosomes along mRNA transcripts. Subsequently, the ribosome-mRNA complexes are lysed, and ribosome-protected mRNA fragments, typically 28-30 nucleotides in length, are isolated and subjected to deep sequencing (Xiao et al., 2018) (Fig3).

The resulting ribosome profiling data offer insights into ribosome occupancy, reading frame usage, and translation efficiency across the transcriptome. By analysing the ribosome footprints' distribution along mRNA sequences, researchers can infer the location and abundance of actively translating ribosomes, as well as detect translational pauses and ribosome stalling events that may reflect regulatory mechanisms or mRNA secondary structures (Ingolia, 2016; Xiao et al., 2018).

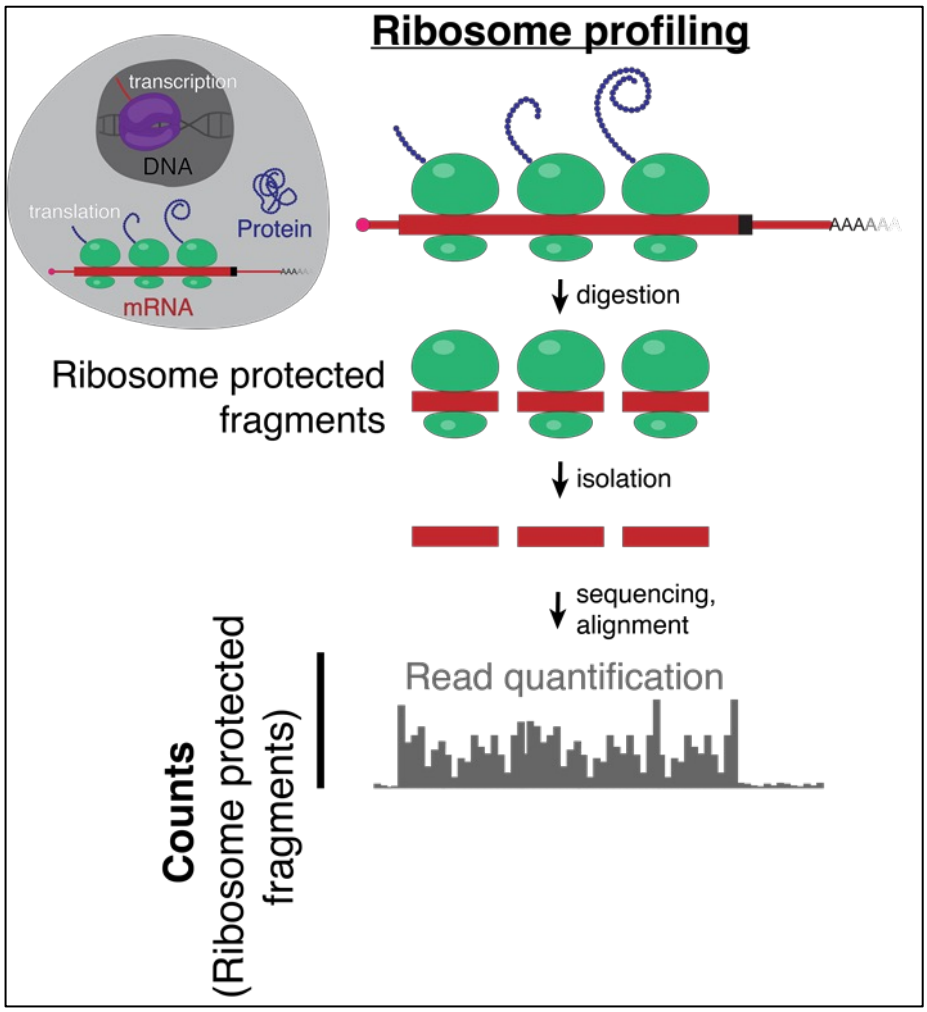

Fig 3: Graphical representation of ribosome profiling (Created using Illustrator based on Ingolia et al., 2012).

sORFs typically exhibit a characteristic pattern of ribosome occupancy, with ribosome footprints concentrated around the initiation and termination codons (Meindl et al., 2023). By quantifying the density of ribosome footprints along mRNA sequences, researchers can identify regions enriched in ribosome binding, thereby pinpointing potential sORFs within the transcriptome (Brar & Weissman, 2015).

Ribosome profiling data can be leveraged to map translation initiation sites (TISs) within mRNA transcripts, which serve as the starting points for protein synthesis. Computational algorithms, such as RiboTaper and RUST, have been developed to predict TISs based on the distribution of ribosome footprints around translation start sites (Bazzini et al., 2014; Calviello et al., 2016; O'Connor et al., 2016).

The advent of high-throughput sequencing technologies has facilitated the comprehensive analysis of entire genomes, enabling researchers to systematically annotate ORFs and especially sORFs across diverse organisms with unprecedented accuracy and resolution. Coupled with sophisticated bioinformatics algorithms and

comparative genomics approaches. These tools have empowered scientists to elucidate the functional significance and evolutionary conservation of ORFs across species boundaries. By analysing the ribosome profiling data, researchers can identify candidate sORFs with evidence of translation initiation and ribosome recruitment (Bazzini et al., 2014; Calviello et al., 2016; Meindl et al., 2023).

The identification and prediction can also be validated with data from mass-spectrometry (Bazzini et al., 2014; Schwaid et al., 2013; Slavoff et al., 2013). Although it is very difficult to identify such small peptides using mass-spectrometry, with proper predictions, they can be planned better. Mass-spectrometry has been used widely to detect small peptides and, hence, small open reading frames (Schwaid et al., 2013; Slavoff et al., 2013). However, the advent of high-throughput ribo-seq data and better analysis makes it better suitable with higher resolution to identify sORFs, which can be individually validated using mass-spectrometry (Bazzini et al., 2014).

1.6 Upstream Open Reading Frames

Upstream Open Reading Frames are sequences of in-frame nucleotides present upstream of the canonical open reading frames (Somers et al., 2013). The regulation of gene expression at the post-transcriptional level is increasingly acknowledged as a fundamental mechanism through which cells and organisms modulate their gene expression patterns. uORFs reduce the protein expression level by reducing the efficiency of translation initiation of the canonical Open Reading Frame (Johnstone et al., 2016) (Fig4).

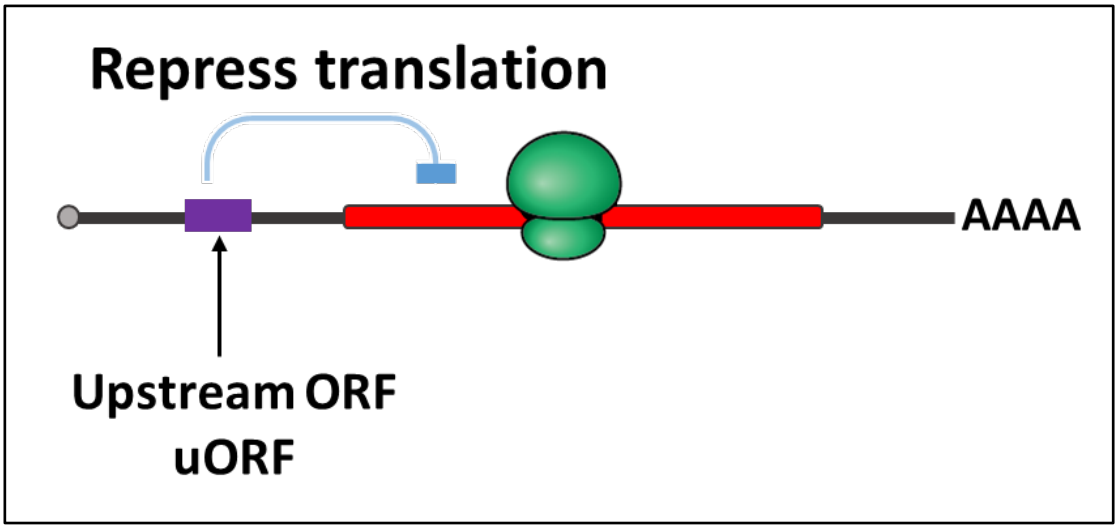

Fig 4: uORFs repress translation of the canonical ORF.

The 5' UTR region contains small open reading frames that have been shown to repress the translation of the main ORF. Figure created using PowerPoint with concept from (Barbosa et al., 2013; Johnstone et al., 2016).

However, uORFs have specific roles under different conditions as well. For example, the presence of uORFs can promote the increased expression of certain stress-related mRNAs in response to cellular stress (Barbosa et al., 2013). It has been shown that 49% of the human transcriptome contains uORF and that they are conserved among species (Barbosa et al., 2013). Genes such as CD36, MDM2, ERBB2, SOC1, and RARB possess conserved and experimentally characterised upstream open reading frames (uORFs) that play a regulatory role in translation. (Johnstone et al., 2016). Moreover, uORFs can also trigger mRNA decay under specific conditions as another means of post-transcriptional control (Somers et al., 2013).

1.7 Downstream Open Reading Frames

Similar to upstream open reading frames, some ORFs were discovered in the 3'UTRs as well and are named downstream Open Reading Frames. Their presence was initially detected by ribosome profiling (Bazzini et al., 2014; Wu et al., 2020; Xiao et al., 2018). After isolating mRNAs, they are digested to produce ribosome-protected fragments known as footprints. These footprints are then sequenced to determine such protected sequences signifying translation. Such sequences were also found in the 3' UTRs, signifying translations in the previously denoted untranslated regions (Bazzini et al., 2014; Ingolia, 2016).

However, there were no known functions for these translated elements in the 3'UTR. Fairly recently, it was shown that downstream open reading frames work in contrast to the downstream open reading frames and enhance the translation efficiency of canonical open reading frames (Wu et al., 2020) (Fig5).

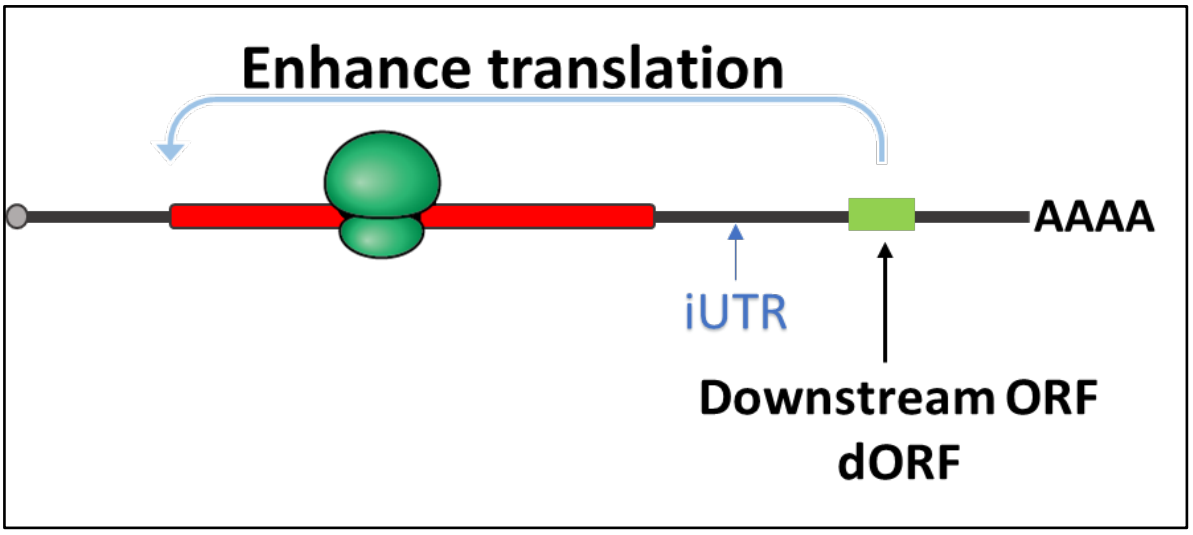

Fig 5: dORFs enhance translation of the canonical ORF.
Similar to uORFs, small ORFs were found in the 3'UTRs through ribosome profiling data that were shown to enhance the translation of the main ORF. The region in between the stop codon of the main open reading frame and the start codon of the dORF is called iUTR. This region is thought to recruit ribosomes to drive the translation of the dORF, which is similar to viral IRES. Figure constructed using PowerPoint with concepts from (Wu et al., 2020).

The presence of dORFs is prevalent in vertebrates and conserved in orthologous genes. This evolutionary conservation means that dORFs are necessary for the functioning of these genes. This is similar to what is seen in uORFs, indicating a selective evolutionary pressure to maintain these dORFs. However, current data suggests that the amino acid sequence is not conserved, signifying that the peptides are not really relevant for their functioning (Wu et al., 2020). As per the available data, the total number of genes that contain dORFs in humans is 1453 (Wu et al., 2020).

The concept of dORFs as a post-transcriptional regulator that enhances the translation of canonical ORF is, however, quite new. Their small sizes, combined with their very little translation, make it difficult for them to be identified. dORFs as a new stable post-transcriptional regulatory mechanism are established now. But, very little is known about their mechanism of function. We know that dORFs are conserved across homologous genes. The present working hypothesis is that the dORF sequence possibly forms a loop and recruits translation initiation factors at the start of the main ORF, increasing its translation. Genes containing multiple dORFs also enhance more than genes containing a single dORF. Since we know that the

sequence of the dORF is not relevant for its functioning, we can replace that and use any translated small ORF as a dORF. We use this feature of dORF by replacing the dORF with fluorescent tags for various assays. This enables us to determine the translation aspects of dORFs by using these fluorescence signals.

Until now, we have not had any data about how the structural components and their properties affect the translation of the dORF or its enhancement activity. With this background information, we aimed to characterise dORFs in human cells. In this project, we established how the various structural aspects like dORF length, the 3'UTR length and the iUTR length affect its enhancement activity, which could serve as a basis to identify a functional model for dORF translation and activity.

Chapter 2

2.1 Materials

High-Efficiency DH5α cells were purchased from NEB(Cat No. C2987I). The mini-prep kit for Plasmid isolation (Cat No. 27106), PCR purification (Cat No. 28106) and gel extraction (Cat No. 28704) kits were from Qiagen. Nuclease-free water is from Ambion (AM9937). All the bacterial media (LB and SOC) and LB-Agar plates were supplied by the media-prep of Stowers Institute. Ampicillin purchased from the cube of the institute was used as the antibiotic selection marker at 100 mg/ml stock concentration (Cat No. 76807). Agarose was purchased from ThermoFisher (Cat No. 17850). All the agarose gel was made with 1x TAE. DNA gel was run in a Bio-Rad gel apparatus system at 90-110 volts for 50 – 70 minutes. Varied gel percentage (0.8%-2%) depending on need was used. 1kb+ DNA ladder from NEB (Cat No. N3200L) or Invitrogen (Cat No. 10488090) was used. 6x Gel Loading Dye was from NEB (Cat No. B7024S). Gels were imaged in BioRad GelDoc Go. BioRad C-1000 Touch thermocycler was used for all the PCRs. Qubit Flex by Invitrogen was used for DNA quantitation. DNA HS kit from Invitrogen was used for this purpose (Cat No. Q32851). 2X Hi-Fi DNA Assembly Master Mix for Gibson cloning was from NEB (Cat No. M5520AA), T4 DNA Ligase (Cat No. M0202L), Phusion Polymerase (M0530L), and Restriction enzymes SnaB1, Xba1, and Xho1 are all from NEB. Lipofectamine 3000 from Invitrogen was used for DNA transfection (Cat No. L3000-001).1X DMEM (15-013-CV) and 1X PBS (21-040-CV) was from Corning. Trypsin-EDTA (Cat No. 25200-056), Penicillin-Streptomycin (Cat No. 15240-062) and OMEM (Cat No. 31985-062) are purchased from Gibco. Primers are from IDT. RLM-RACE kit for 3'RACE was purchased from Invitrogen (Cat No. AM1700). Cytek Aurura Flow Cytometer was used to acquire the fluorescent intensities. HEK293T cells are from the American Type Culture Collection with identifier number #CRL-11268. FCS Express by DeNovo Softwares was used for cytometric analysis. R software was used for calculation and subsequent plotting.

2.2 Methods

2.2.1 Tissue culture

HEK293T cells were cultured with DMEM media, supplied with 10% FBS. The cell culture media also contains L-glutamine and penicillin/ streptomycin. The cells were ordered from the tissue culture facility from the Stowers Institute for Medical Research at a relatively low passage, lower than passage 12. Washes were given with Phosphate Buffer Saline (PBS) with low $MgCl_2$ and $CaCl_2$ - dPBS and trypsinised using 0.25% Trypsin-EDTA. The cells were incubated at 37^0 C at 5% CO_2. For transfection, 0.5 x 10^5 cells were seeded in each well of a 48-well plate.

2.2.2 Cloning

Cloning was a major part of the project. To address the first objective, a wild-type reporter was available in a pCS2 vector that had a mCherry sequence as the main ORF followed by an iUTR and then a ZsGreen as the dORF. This is termed as the wild type dORF reporter. The second amino acid of this ZsGreen was then mutated to generate a stop codon so that translation does not happen for this ZsGreen, and this group was termed as the dMut reporter.

The ZsGreen was then truncated to generate different lengths of the dORFs. Nine different lengths, namely 180AA, 90AA, 50AA, 30AA, 20AA, 10AA, 6AA, 5AA, and 3AA, were cloned for both the wild-type and mutant reporters. This was done using PCR amplification where one primer had 5' phosphorylation and hence re-ligated after PCR amplification. This truncated ZsGreen does not show any fluorescence but acts as the dORF.

To test iUTR independence, iUTR from a different gene called CENP-A, which has a translated dORF, was cloned into the wild type and mutant dORF reporter containing full-length ZsGreen. This ZsGreen was then truncated as before. But we generated thirteen different lengths, namely, 180AA, 120AA, 90AA, 60AA, 50AA, 40AA, 30AA, 20AA, 10AA, 7AA, 6AA, 5AA, and 3AA, to get a more continuous distribution.

To address the 3'UTR length dependence, the 3'UTR sequence was amplified from a zebrafish gene called PPARAA. This was the longest available 3'UTR sequence available to us in the lab. The full length of this sequence was 2062 bp. A shorter length of 1000 bp was also amplified. This sequence was then inserted into the enhancement plasmid of 50AA length downstream of the dORF using Gibson cloning.

Two different restriction enzyme sites for Xba1 and Xho1 were also cloned at the 5' and 3' positions of the 3'UTR sequence. This was done for the 50AA truncated ZsGreen plasmids containing CCDC167 and CENP-A iUTR.

To address the iUTR dependence of dORF enhancement activity, the two different 3'UTR lengths were cloned before the iUTR of the wild-type and mutant reporter of 50AA truncated ZsGreeen using a Gibson assembly reaction.

All these plasmids were then transfected into the 293T cells to check for the relative expression of mCherry.

2.2.3 Transfection

HEK 293T cells were transfected with lipofectamine 3000 based on the manufacturer's instruction in 48-well plates. The plate is set overnight before transfection in 48-well plates to achieve 70% confluency on the day of transfection. 0.5 x 10^5 cells were plated. 500ng total DNA was added with transfection reagents per well. DNA was diluted to be 125ng/µl, and 2µl of the reporter plasmid was transfected with 2µl of transfection control, which was eGFP in this case. For dORF reporters, 24h post-transfection, cells were trypsinised and collected in DMEM to be analysed using Cytometry.

2.2.4 Cytometric Analysis

Cells after transfection were run through a flow cytometer (Cytek Aurora) to check for fluorescence intensity. Cells showing both the fluorescence (double positives) were considered for calculation using lasers of eGFP (488/510) and mCherry (587/610). For running cytometry, cells were suspended in DMEM with 10%FBS. Cells were not fixed. The median intensity of mCherry and GFP from the double positive population were extracted for each of the transfected plasmids and used for further calculation. The relative expression of mCherry with respect to GFP was used for quantification. This relative expression was plotted against the dORF length. This ratio was normalised to be 1 for all the mutants, and the corresponding ratio for the wild-type reporters was plotted to get the fold change of mCherry expression in wild type reporters compared to their corresponding mutant reporters.

2.2.5 Data analysis and plotting

Cytometry data from the .fsc files were extracted using FCS express software. The calculation was done in Microsoft Excel. R (version 4.3.3, 2024-02-29) was used for plotting, mean fold change calculation and significance analysis. A two-tailed t-test was used to calculate the significance of the data.

Chapter 3

Results

3.1 The effect of the length of the downstream Open Reading Frames

3.1.1 Enhancement of the main Open Reading Frame translation is dependent on the length of the dORF

The idea of this experiment was to deduce whether the enhancement of the expression of the main open reading frame is dependent on the length of the dORF. As explained in the methods section, nine different lengths of dORF were used with the CCDC167 iUTR (Fig6A). For every length of the dORF reporter, there was a dMut reporter that had a stop codon at the second amino acid position. The plasmids were co-transfected with GFP in HEK293T cells for transfection and normalisation control (Fig6B). Cytometry was done 24 hours post-transfection.

We measured the mCherry expression of this WT dORF reporter and the corresponding dMut and then normalised it to the GFP expression for each. This relative expression of mCherry/GFP was further normalised to 1 for the dMut reporters, and the corresponding expression of the WT dORF reports was calculated and plotted.

We observed that the length of the dORF was indeed important for translation enhancement (Fig6C). Starting from 6AA, the wild type dORF reporters showed significantly increased mCherry expression compared to the mutant once, suggesting that the shortest dORF to show enhancement was 6AA in length. dORF, which has less than six amino acids, did not demonstrate this effect. Starting from 6AA enhancement gradually increased with increasing length of the dORF and peaked at 50AA length of the dORF. dORFs of length 30 and 50 AA showed an approximately two-fold increase in the translation of the main ORF compared to their respective Mutant reporters. The best enhancement was seen for dORFs with 50 amino acids

length. After which the enhancement showed a decrease. dORFs of length 180AA did not enhance the translation of the main open reading frame.

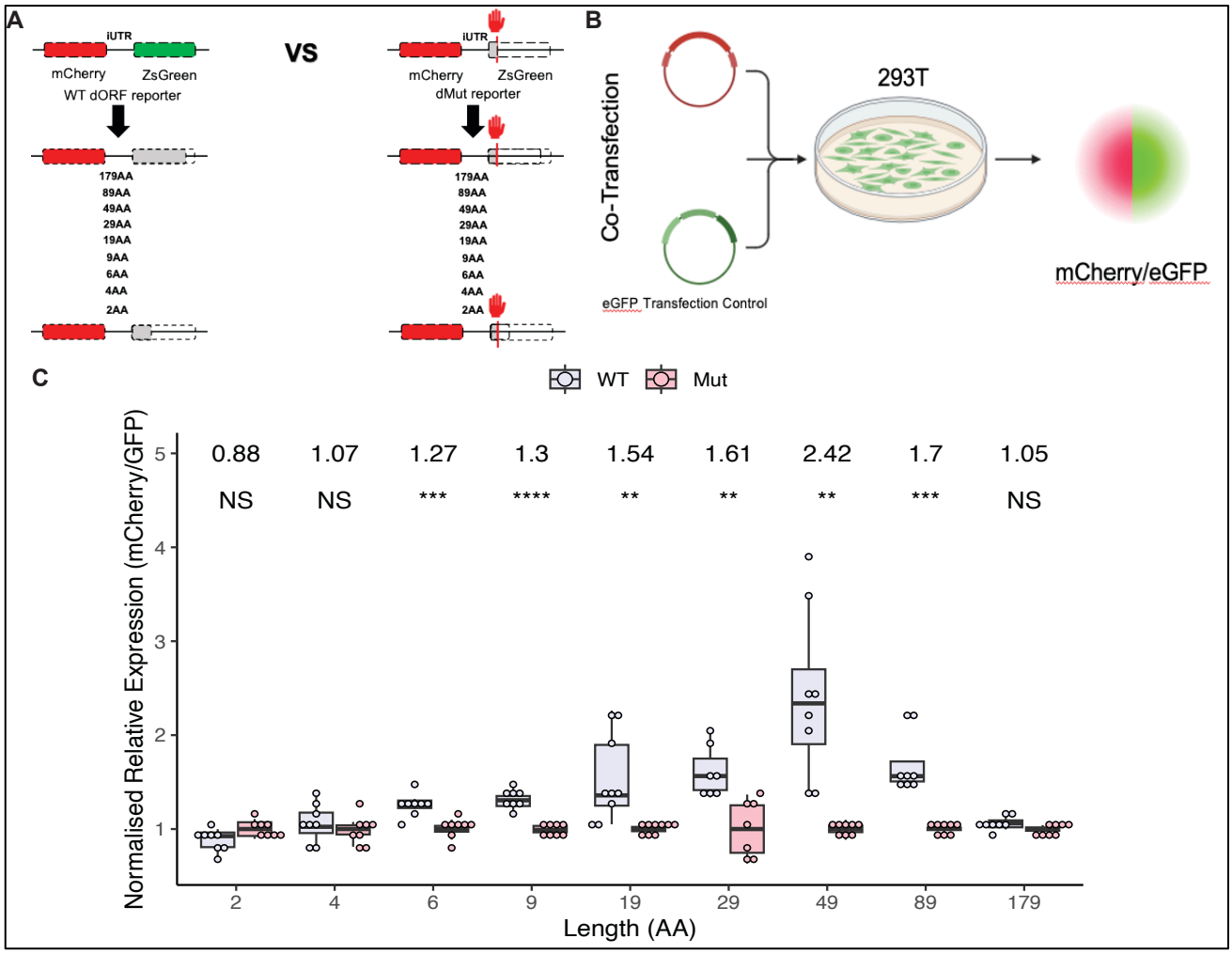

Fig 6: Enhancement of the main Open Reading Frame is dependent on the length of the dORF.

A. Schematic showing cloning procedure of the WT and Mut dORF reporters. These reporters contained iUTR from the CCDC167 gene, and we had nine different dORF lengths.
B. Schematic showing the workflow of the experiment.
C. Boxplot showing the effect of translation enhancement of the main open reading frame depending on the length of the dORF. The numbers on the top indicate the mean fold change of mCherry expression in the WT dORF reporters compared to the mutant reporters. The experiment has four biological replicates with two technical replicates each time.

3.1.2 The length dependence of the main ORF enhancement is independent of the iUTR.

To check whether this enhancement effect was dependent on the specific iUTR we replaced the iUTR. This was especially important because we think that the iUTR functions similar to a viral IRES and hence recruits ribosomes for dORF translation. Initially, all the plasmids of 9 different dORF lengths contained the iUTR from a gene called CCDC167. This iUTR was replaced with another iUTR from the gene CENP-A, which also contains a translated dORF (Fig. 7A). Moreover, four extra lengths of dORF were cloned to obtain a more continuous distribution of length. These 13 different dORF and dMut reporters were then transfected. mCherry and GFP expression were measured, and we observed the same enhancement activity of the dORF. Plasmids containing dORF of length six amino acids and above showed enhanced translation of the main open reading frame. Similar to CCDC167, dORF reporters of length 30AA and 50AA showed approximately 2-fold increased mCherry expression in the WT reporter compared to mutant ones (Fig7B). Moreover, this result also confirms that translation of the dORF is necessary for it to be enhancing, as the only difference between the WT and dMut reporters is the stop codon at the second AA position of dMut reporters, preventing it from getting translated.

We conclude from these experiments that the enhancement of the translation of the main ORF is dependent on the length of the dORF. For enhancing the main ORF expression, dORF can neither be too short (less than 6AA) nor too long. Moreover, this enhancement is not dependent or specific to the iUTR.

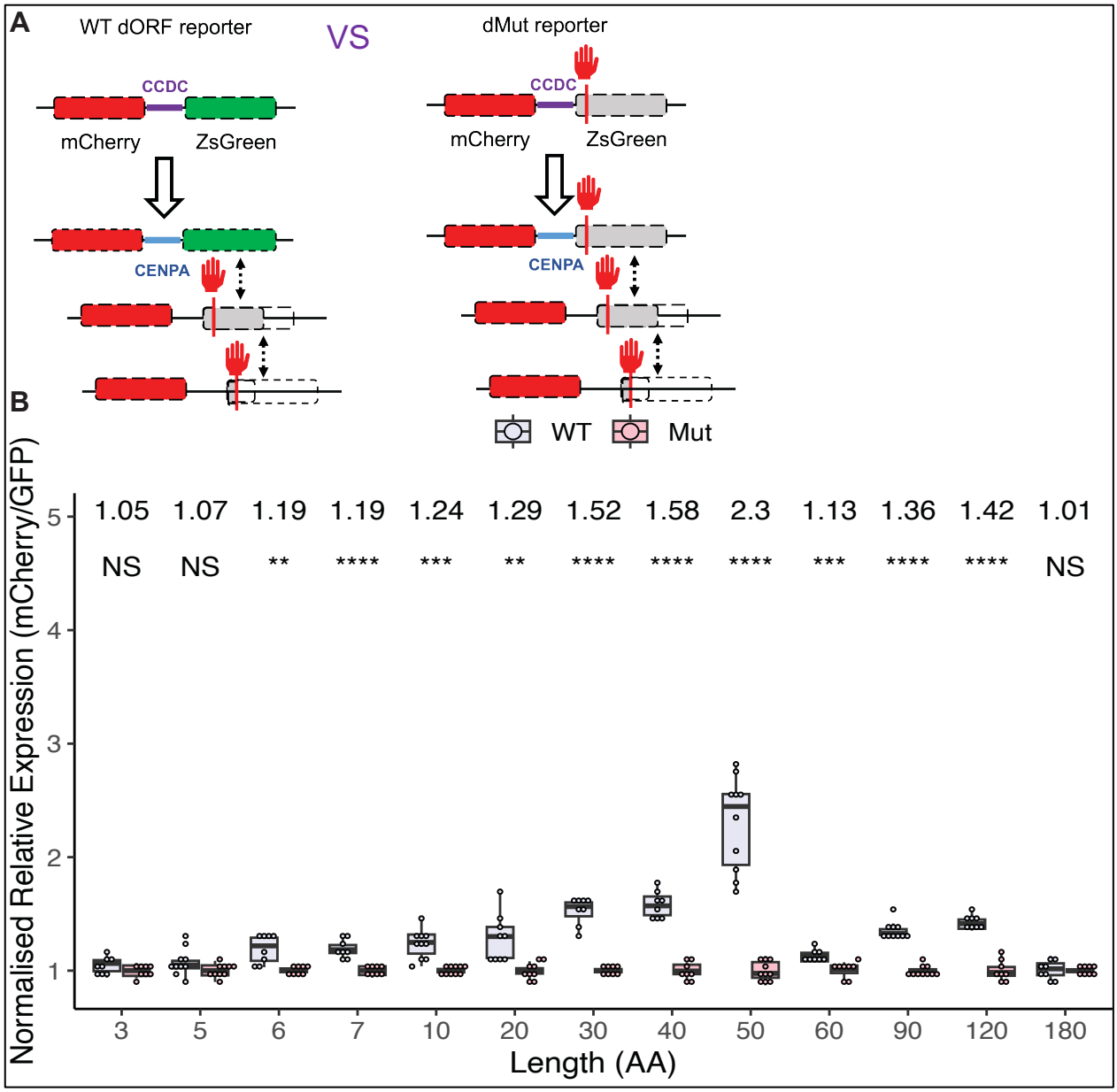

Fig 7: Enhancement of translation of main ORF by dORF is iUTR independent

A. Schematic showing cloning strategy of the WT dORF and dMut reporters.
B. Boxplot showing the effect of translation enhancement of the main ORF depending on the length of the dORF. These reporters contain 13 different lengths of dORF and their corresponding dMut. The numbers on the top indicate the mean fold change of mCherry expression in the WT dORF reporters compared to the mutant reporters. The experiment has four biological replicates with two technical replicates each time.

3.1.3 dORFs less than 6AA show possible translation enhancement

Earlier small open reading frames less than 10AA were not considered for analysis. It was partly due to the fact that shorter ORFs are difficult to detect by ribosome profiling because the number of ribosome profiling correlated with the length of the coding sequence. All the analyses to identify and characterise dORFs were based on the assumption that they would be more than 10AA. Hence, our universe of dORFs or small ORFs, for that matter, was limited to up to ten amino acids only. However, after finding that dORF as short as 6AA can enhance translation (Fig: 6C,7B), it becomes important to check their possible biological role with more importance. This opens up many new possibilities about even smaller ORFs and their potential functions that might be regulatory and not necessarily peptide-dependent.

3.1.4 Identification of short downstream Open Reading Frames

Based on the evidence we found before, we rerun the entire ribosome profiling pipeline to identify possible small dORFs in the 6-9 amino acids range that we might have missed earlier. We found new possible 78,571 dORFs. This implies that the total number of genes containing dORFs could be much higher than we currently think. However, this huge number is just a possibility and demonstrates how vast the scope is, but is not possible to experimentally validate. Therefore, based on criteria like the number of times they are detected in ribo-seq data, median coverage, and their median ORF score, we prepared a ranking (Fig: 8A). We selected the top 50 candidates and checked for their presence in the metagene plots (Fig: 8B). We wanted to confirm whether these metagene plots contain characteristic features of the known translated element of the genome. We observed that these metagene plots possess properties like three-nucleotide periodicities and the presence of nucleotides upstream and downstream. This indicates that these very short open reading frames, i.e., short dORFs in this case, can indeed be translated. Based on these evidences, we are designing a library containing the sequences of these fifty short dORFs and their corresponding iUTRs so that they can be individually validated for translational activity. The library also includes frameshift reporters of these 50 short dORF sequences to experimentally check their correct start site. Therefore, this library will be inserted into 293T cells using lentiviral transduction and checked whether these short dORFs get translated or not using reporter expression (Fig. S2).

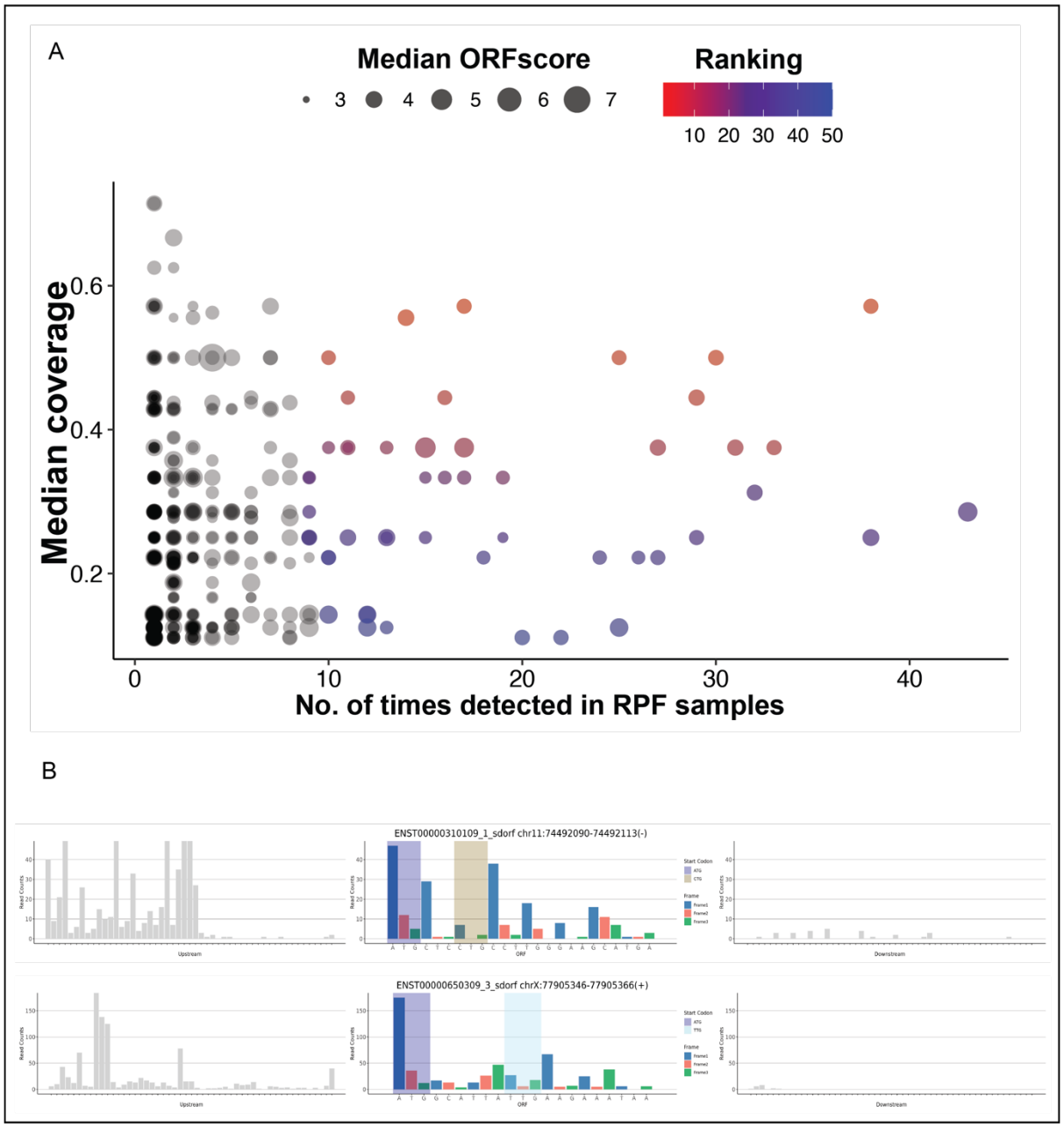

Fig 8: Identification of short downstream Open Reading Frames.

A. Scatter plot showing the distribution of all the 78,571 possible newly identified short dORFs. They were ranked based on median coverage, the no. of times they were detected and their median coverage. The best 50 candidates are coloured, and the red indicates the best, and the blue indicates the 50th ranked short dORF.
B. Metagene plots as obtained from ribosome profiling for two of the top 50 short dORFs. They demonstrate clear three-nucleotide periodicity and the presence of reads upstream and downstream. The coloured region indicates the dORF.

3.2 The effect of the length of the 3'UTR on the enhancement effect of dORFs on main ORFs

3.2.1 Enhancement of the main Open Reading Frame translation is independent of the 3'UTR length

To check whether the length of the 3'UTR has any effect on the translation enhancement capability of the dORF, two different lengths of 3'UTR were cloned downstream to the 50AA dORF reporter plasmid to create a total of 3 variants of reporter plasmids. This was based on the fact that the 50AA dORF showed the best enhancement. These reporter plasmids were then co-transfected with GFP in HEK293T cells, and the fluorescent intensity was measured using cytometry. The normalised expression of mCherry was plotted for all the WT reporters along with their corresponding mutant reporters containing these three different lengths of 3'UTR.

We observed no significant difference in the mCherry expression with different 3'UTR lengths. The dORFs were still enhancing the mCherry expression as expected, and the change in the length did not cause any difference in the enhancement capability. The experiment was performed for two independent iUTRs from CCDC (Fig. 9B) and CENP-A (Fig. 9C), and the results were consistent. For all the reporters containing 50, 1000 or 2000 bp of 3'UTR, the WT dORF reporters showed approximately two-fold increased mCherry expression compared to the mutant reporters. There was no significant difference in mCherry expression depending on the 3'UTR length, indicating that the length of the 3'UTR is independent of the dORF enhancement activity.

To validate that the 3'UTR was not getting spliced during RNA processing, we extracted total RNA from the cells after transfection and ligated specific adapters to the end of the mRNAs. 3'RACE and Northern Blots will be performed to check whether the transcript lengths are three different as expected or not. We can only attribute our observed result to the length of the 3'UTR only after we confirm that the 3' UTRs do not get spliced. These experiments are currently being performed. We observed the expected bands in each of these conditions corresponding to 1kb and 2kb 3'UTR length.

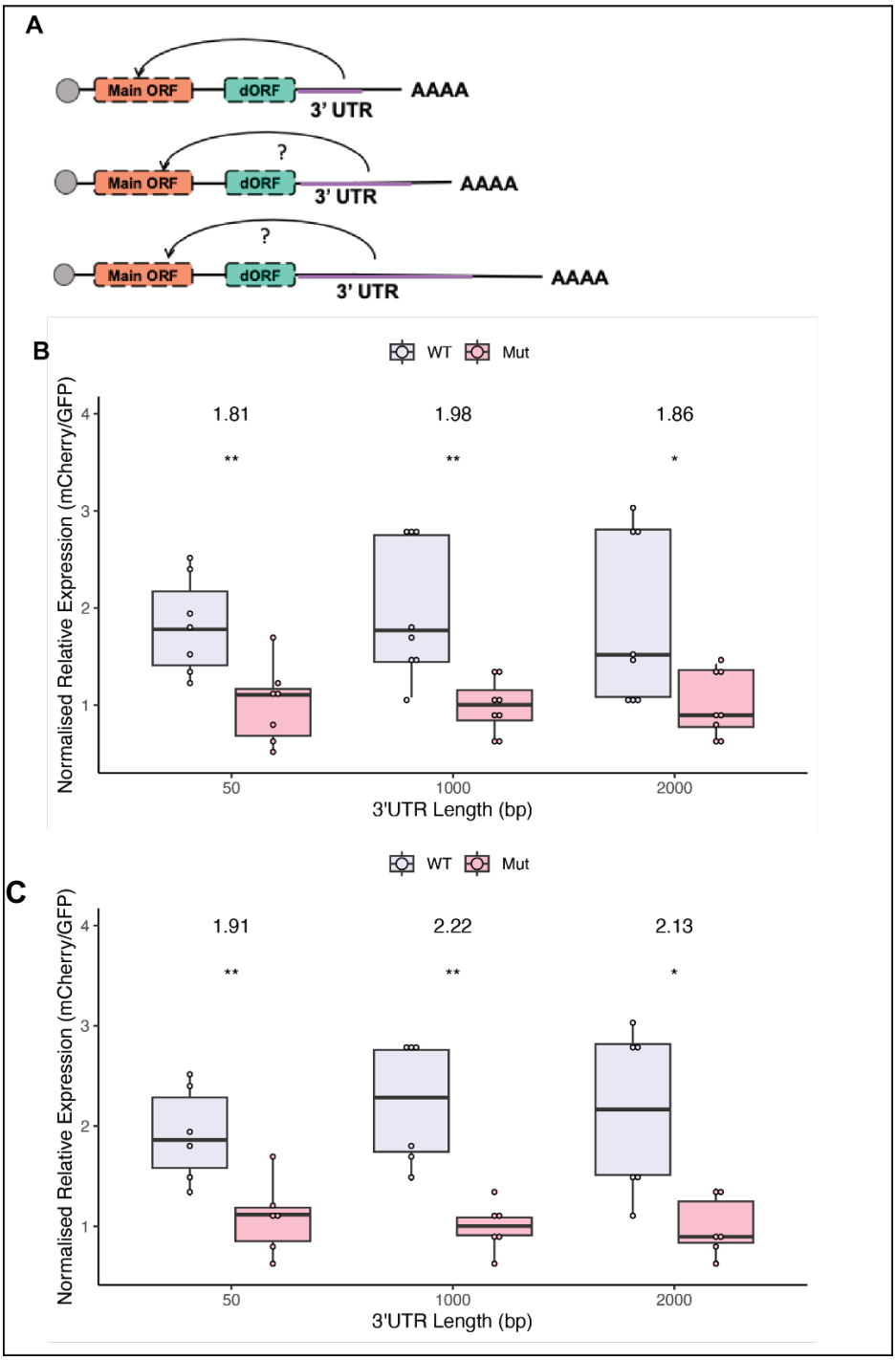

Fig 9: Enhancement of the main Open Reading Frame translation is independent of the 3'UTR length.

A. Schematic showing the experimental plan for checking the correlation between 3'UTR length and main ORF translation enhancement by dORF.
B. Boxplot showing the effect of translation enhancement of the main ORF by dORF irrespective of the 3'UTR length. This reporter contains iUTR from CCDC167. This experiment has three biological replicates with two technical replicates each.
C. Boxplot showing the effect of translation enhancement of the main ORF by dORF irrespective of the 3'UTR length. This reporter contains iUTR from

CENP-A. This experiment has three biological replicates with two technical replicates each.

3.3 The effect of the iUTR length on the enhancement effect of dORFs on main ORFs

3.3.1 Enhancement of the main Open Reading Frame is dependent on the internal UTR length.

Next, we investigated how changing the length of the iUTR affects this enhancement capability of dORF. As mentioned, iUTR is the distance between the stop codon of the main open reading frame and the start codon of the downstream Open Reading Frame. We suspected that changing this length might affect the enhancement as it might cause disruptions in ribosome recruitment. The 3'UTR that we cloned before was cloned in between the iUTR to elongate it (Fig 7A). Therefore we had two more lengths of iUTR of 1,000bp and 2,000bp along with the previous endogenous iUTR. Following the previous protocol as before, we transfected these plasmids in HEK293 cells along with GFP as a transfection and normalisation control. The median mCherry and GFP intensity were used for calculations.

We found that with the increasing length of iUTR, the enhancement capability of the WT reporters decreased. There was no significant increased mCherry expression in wild type reporter compared to the mutant one for both the reporters containing either 1000bp or 2000bp elongated iUTR. Moreover, the fold change was near one for the longest iUTR of 2000 bp for both the CCDC and CENPA iUTR (Fig10 B, C).

We suspected that this could be due to two reasons. Either the length of the iUTR is related to the enhancement capacity of the dORF, or the dORF itself was not getting translated; as we know, translation is necessary for enhancement. To validate our concern about translation, these iUTR-dORF pairs were cloned into the bi-cistronic reporters already available in the lab to check for translation (Fig. S1). These reporters will express both red and green colours if the dORF is translated and only red if it is not translated. However, this bi-cistronic system is in a lentiviral construct and the experiments are yet to be performed.

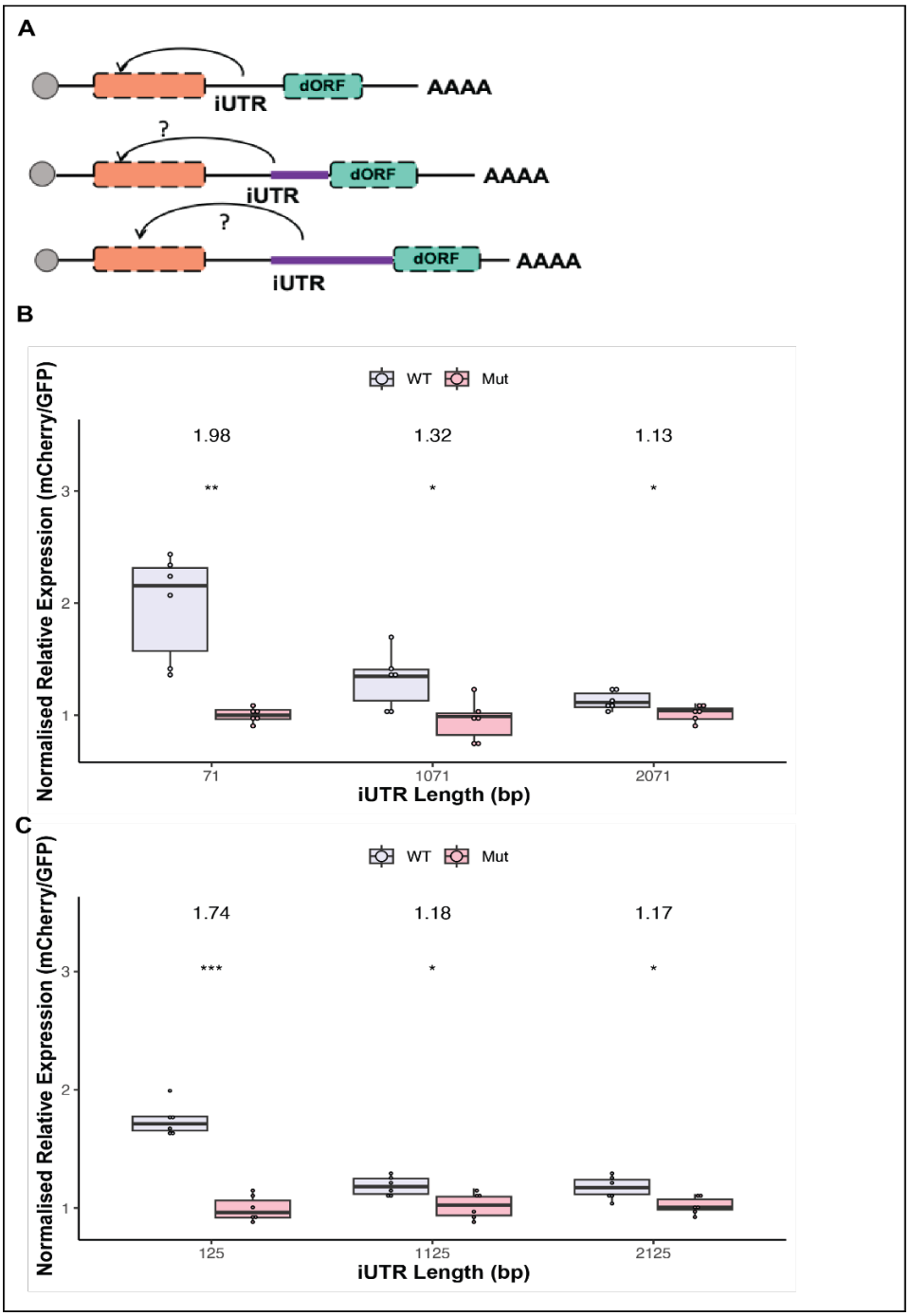

Fig 10: Enhancement of the main Open Reading Frame is dependent on the internal UTR length.

A. Schematic showing the experimental plan for checking the correlation between iUTR length and main ORF translation enhancement by dORF.
B. Boxplot showing the effect on translation enhancement of the main ORF by dORF depending on the iUTR length. This reporter contains iUTR from the gene CENP-A. This experiment has three biological replicates with two technical replicates each.
C. Boxplot showing the effect on translation enhancement of the main ORF by dORF depending on the iUTR length. This reporter contains iUTR from the gene CCDC. This experiment has three biological replicates with two technical replicates each.

3.4 Validation of experimental results using *in silico* data

As a validation of our experimental findings, we went on to look into previously published experimental datasets to see whether our results are supported by *in silico* data or not. We found excellent agreement between our findings and the previous data. Fig 11A demonstrates how dORFs look in a ribosome profiling analysis. We can see that we already knew that dORFs are generally short with a median of around 30AA in length in endogenous conditions. (Fig. 11B). Moreover, we can also see that the iUTR lengths are also generally very short in translated dORFs when compared to untranslated ones (Fig. 11B). This validates our finding that dORFs lose enhancement capability when the iUTR is elongated (Fig10). As mentioned, we can find the reason for this with more experiments. However, this could very well be due to the fact that a long iUTR does not allow the dORF to be translated, which is necessary to enhance the main ORF translation.

Moreover, the fact that there is no significant difference in 3'UTR length in translated and non-translated dORFs (Fig. 11C) is in agreement with our findings that dORF enhancement of the main ORF is independent of the 3'UTR length (Fig. 9).

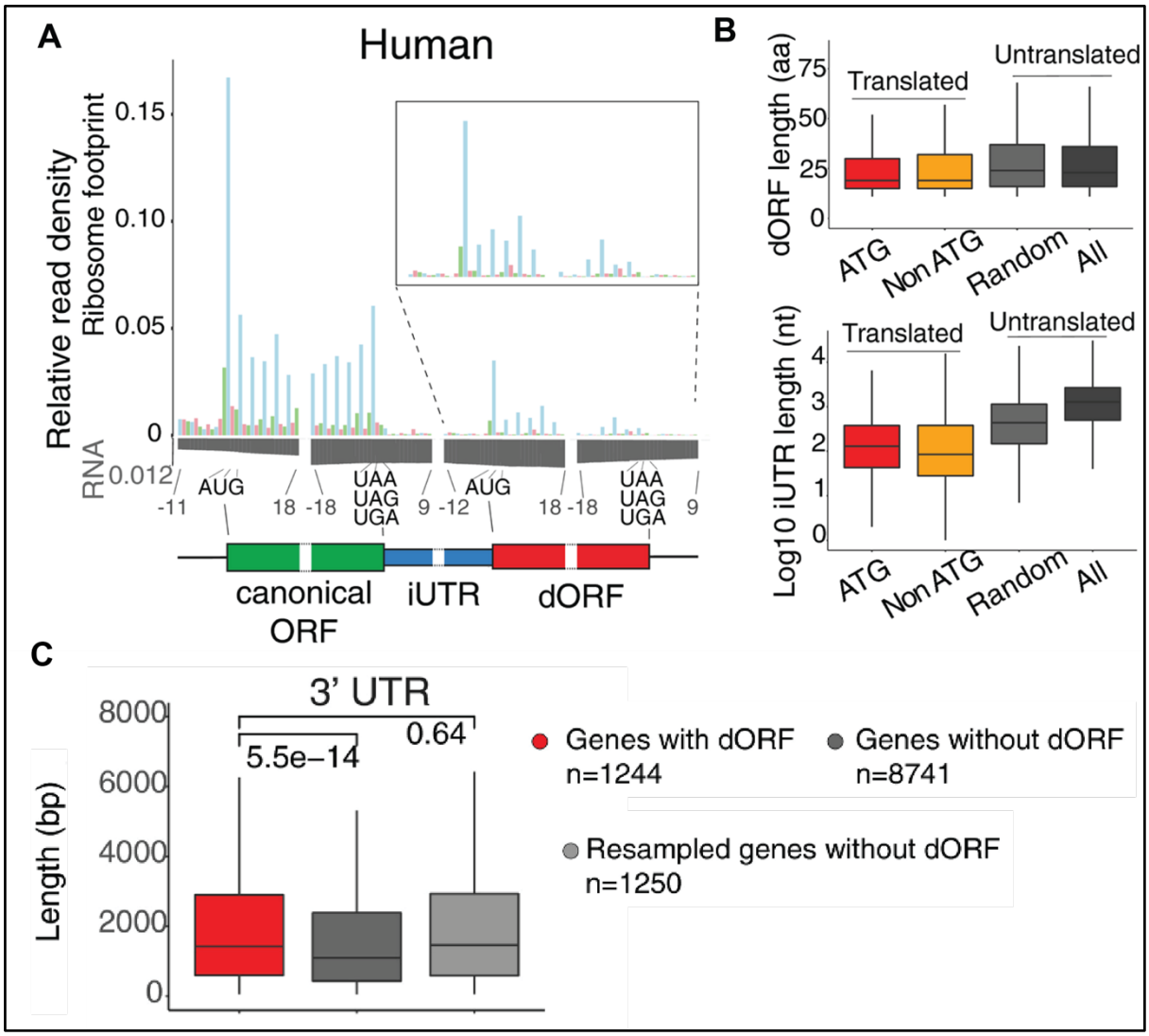

Fig 11: *In silico* data supports experimental findings of 3'UTR and iUTR lengths

A. Schematic showing how dORF looks in a ribosome profiling dataset
B. Boxplot showing that endogenous dORFs are short in length with a median of around 30AA. Translated dORFs also possess a short iUTR
C. Boxplot showing that there is no significant difference in 3'UTR length in groups containing translated or untranslated dORFs. Data Credit, (Wu et al., 2020)

Chapter 4
Discussion

The findings in this thesis further establish dORFs as a novel post-transcriptional regulator. The result further supports previously published data from the lab that showed that the translation of dORF enhances the translation of the main ORF. Moreover, the level of regulation, which is a two-fold increase in this case, is also consistent with previously observed data. However, our findings establish new characteristic characterisations of dORF that opens up avenues to understand how their structure either restricts or enables them to functionally enhance the expression of the main Open Reading Frame. We believe that dORFs are translated by new ribosome recruitment through the iUTR based on some preliminary data. The ribosomes that translate the main ORF are not responsible for translating the dORF.

The loop hypothesis of mRNA structure talks about the 3' and 5' interaction of mRNA that enables ribosome recruitment for mRNA translation. If that is true, then the translation of dORF by ribosomes might signal the ribosomal recruitment factor to recruit more ribosomes to the main ORF for increased translation.

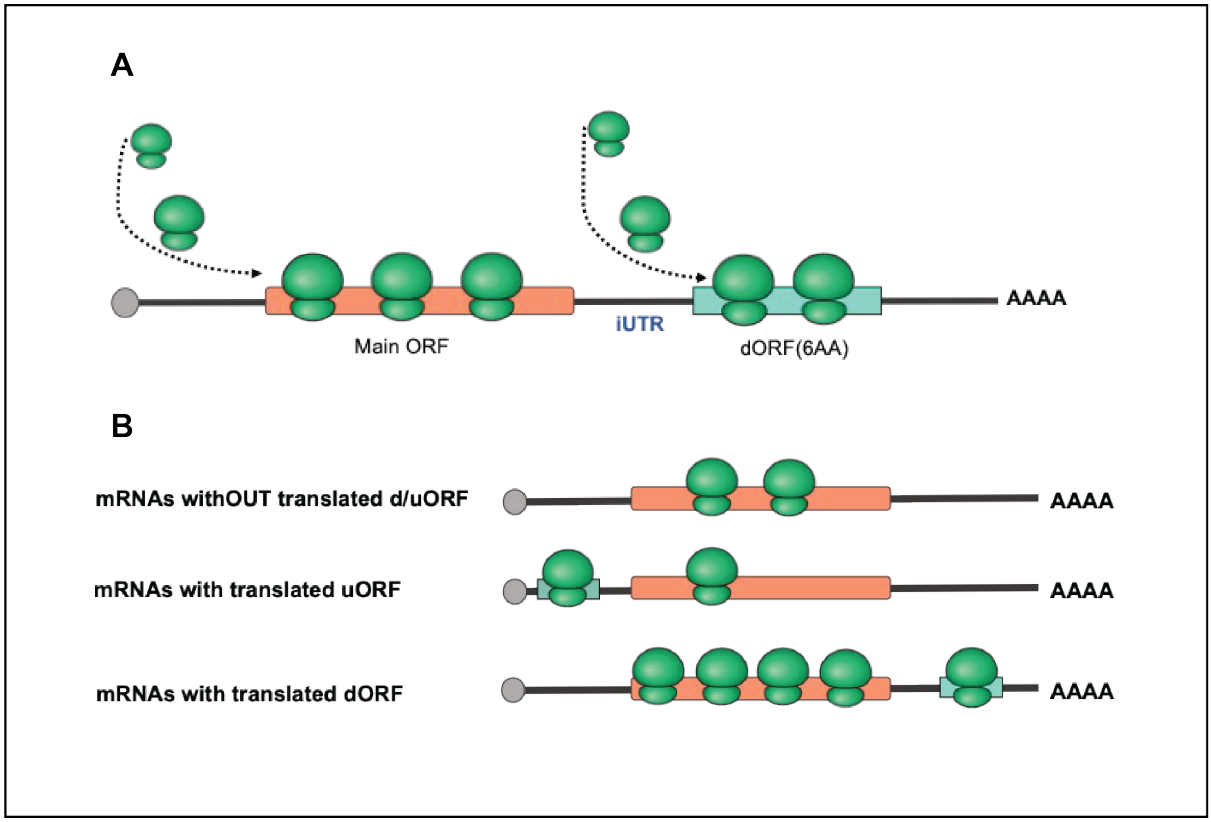

Fig 12 : Schematic highlighting translation of dORFs (Proposed model using data from (Wu et al., 2020) and this work).

A. **dORFs are not translated by the same ribosomes that translate the main ORF. iUTRs are thought to recruit new ribosomes for dORF translation.**
B. **mRNAs with translated dORFs recruit more ribosomes to the main ORF, enhancing their translation.**

The data proves that in order to enhance, the dORFs necessarily need to get translated. We believe that six amino acids is the minimum length required to recruit two ribosomes, which might be necessary to recruit ribosomal recruitment factor to the main ORF. The fact that at least six amino acids are needed to significantly enhance the main ORF translation can be explained by this. It is also exciting to note that there is an optimal length of dORF that can enhance the most. With a very long dORF, the ribosome might just pass through without ever being able to signal recruitment at the main ORF.

Moreover, the idea of short dORFs is going to open a new universe of short ORFs in general and highlight newer functions of such short sequences, altering the definition of small open reading frames in general. It demonstrates a yet unknown world that we did not pay attention to signifying; such short sequences might have important regulatory roles that might have been overlooked. And hence, better functional characterisation is required for these short dORFs or small ORFs in general.

Moreover, with the finding of these short dORFs, we can delve deeper into other species to find similar occurrences in them. This might enable us to find evolutionary conservation and elucidate why dORFs were necessary in the first place. Moreover, this might highlight how these lengths vary across species and show if there are any correlations.

The fact that 3’UTR length does not alter the enhancement of translation of the main Open Reading Frame by the dORF is more validatory in nature. We already knew that the length of 3’UTR does not differ in the sets of genes that either contain or do not contain dORFs. Hence, it can be concluded that the effect of dORF on the main ORF expression does not involve the 3’UTR in its molecular mechanism.

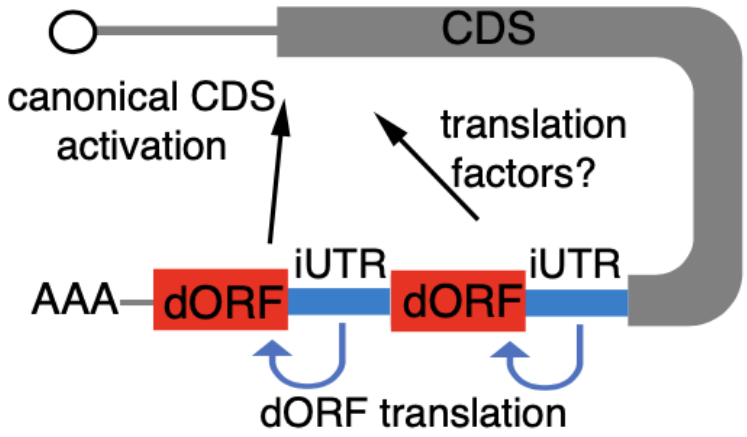

Fig 13 : Schematic depicting proposed hypothesis of main ORF translation enhancement by dORF. dORF translation might activate canonical CDS, leading to the recruitment of new ribosomes using signalling by translation factors (Wu et al., 2020).

The iUTR length dependence is very crucial in deciphering a functional model, though. Until now, we have not had a proposed mechanism for how this dORF actually functions. We know that it is regulatory through the sequence and not peptide-dependent. Therefore, the length dependence of the iUTR highlights that the mode of action is more structural in nature. The iUTR is the distance between the main ORF stop and the dORF start. Furthermore, increasing that length might hinder efficient ribosome recruitment, creating an inefficient translation of the dORF. Moreover, the distance might also restrict interaction between the main ORF and the dORF which might be necessary for the dORF to enhance.

The work successfully characterises different structural elements related to dORF and its associated regulatory mechanism. However, it also highlights the scope of a plethora of studies that is essential to completely understand how dORFs function in a cellular context. Experiments need to be performed in endogenous conditions to validate the results and provide more context and conclusions. The work establishes a groundwork based on which further investigations into the functional significance of dORFs and their implications for cellular physiology and disease pathogenesis could be conducted. uORFs have already been shown to be associated with different diseases where they alter gene expression by expression repression. We can also speculate that dORFs might be associated with similar expression regulation. The field

is quite new and requires more investigation for better functional characterisation. Gene Ontology analysis might be performed to check if there is any significant enrichment across the genes containing dORFs. The evolutionary context of dORF also needs to be looked into, which might highlight other functions associated with dORFs. The presence of dORF might provide stability to the mRNA transcript to prevent it from going into degradation, like in the context of non-sense mediated mRNA degradation (NMD). Since iUTR functions similar to viral IRES, the effect of dORFs in specific contexts like viral infection might also be looked into.

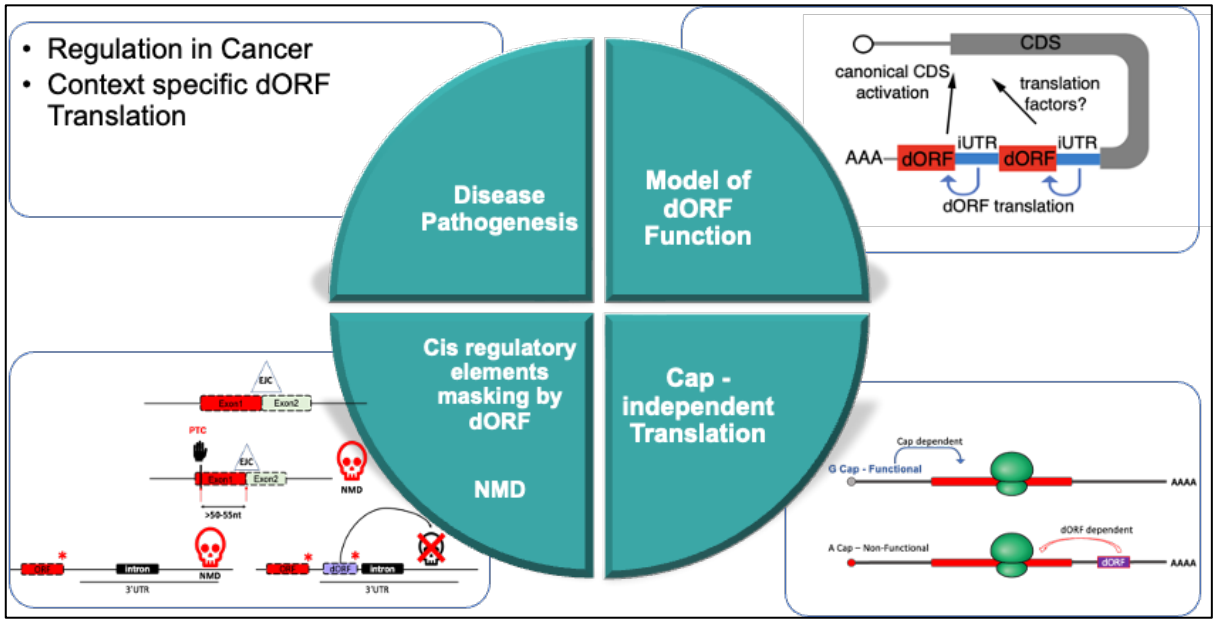

Fig 14 : Schematic depicting multi-faceted approaches and future perspectives based on this thesis or in relation to dORF in general.

Although this study answers some basic questions related to characterisation of dORFs, it opens up more questions requiring more investigations. Our result provides insights that will be very essential for further investigations that might answer multiple open questions relating to the molecular mechanism and biological impact of dORFs.

References

Alberts B, J. A. L. J. et al. (2002). *Molecular Biology of the Cell : From DNA to RNA.* (4th ed.). Garland Science.

Albuquerque, J. P., Tobias-Santos, V., Rodrigues, A. C., Mury, F. B., & Da Fonseca, R. N. (2015). small ORFs: A new class of essential genes for development. In *Genetics and Molecular Biology* (Vol. 38, Issue 3, pp. 278–283). Brazilian Journal of Genetics. https://doi.org/10.1590/S1415-475738320150009

Barbosa, C., Peixeiro, I., & Romão, L. (2013). Gene Expression Regulation by Upstream Open Reading Frames and Human Disease. In *PLoS Genetics* (Vol. 9, Issue 8). https://doi.org/10.1371/journal.pgen.1003529

Bazzini, A. A., del Viso, F., Moreno-Mateos, M. A., Johnstone, T. G., Vejnar, C. E., Qin, Y., Yao, J., Khokha, M. K., & Giraldez, A. J. (2016). Codon identity regulates mRNA stability and translation efficiency during the maternal-to-zygotic transition . *The EMBO Journal*, *35*(19), 2087–2103. https://doi.org/10.15252/embj.201694699

Bazzini, A. A., Johnstone, T. G., Christiano, R., MacKowiak, S. D., Obermayer, B., Fleming, E. S., Vejnar, C. E., Lee, M. T., Rajewsky, N., Walther, T. C., & Giraldez, A. J. (2014). Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO Journal*, *33*(9), 981–993. https://doi.org/10.1002/embj.201488411

Brar, G. A., & Weissman, J. S. (2015). Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nature Reviews Molecular Cell Biology*, *16*(11), 651–664. https://doi.org/10.1038/nrm4069

Calviello, L., Mukherjee, N., Wyler, E., Zauber, H., Hirsekorn, A., Selbach, M., Landthaler, M., Obermayer, B., & Ohler, U. (2016). Detecting actively translated open reading frames in ribosome profiling data. *Nature Methods*, *13*(2), 165–170. https://doi.org/10.1038/nmeth.3688

Chatterjee, S., Chauvier, A., Dandpat, S. S., Artsimovitch, I., & Walter, N. G. (2021). A translational riboswitch coordinates nascent transcription-translation coupling. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(16). https://doi.org/10.1073/pnas.2023426118

Chen, J., Brunner, A.-D., Cogan, J. Z., Nuñez, J. K., Fields, A. P., Adamson, B., Itzhak, D. N., Li, J. Y., Mann, M., Leonetti, M. D., & Weissman, J. S. (2020). Pervasive functional translation of noncanonical human open reading frames. *Science*, *367*(6482), 1140–1146. https://doi.org/10.1126/science.aay0262

Cobb, M. (2017a). 60 years ago, Francis Crick changed the logic of biology. *PLoS Biology*, *15*(9). https://doi.org/10.1371/journal.pbio.2003243

Cobb, M. (2017b). 60 years ago, Francis Crick changed the logic of biology. *PLoS Biology*, *15*(9). https://doi.org/10.1371/journal.pbio.2003243

Couso, J. P., & Patraquim, P. (2017). Classification and function of small open reading frames. In *Nature Reviews Molecular Cell Biology* (Vol. 18, Issue 9, pp. 575–589). Nature Publishing Group. https://doi.org/10.1038/nrm.2017.58

CRICK, F. (1970). Central Dogma of Molecular Biology. *Nature*, *227*(5258), 561–563. https://doi.org/10.1038/227561a0

Ingolia, N. T. (2016). Ribosome Footprint Profiling of Translation throughout the Genome. In *Cell* (Vol. 165, Issue 1, pp. 22–33). Cell Press. https://doi.org/10.1016/j.cell.2016.02.066

Ingolia, N. T., Brar, G. A., Rouskin, S., McGeachy, A. M., & Weissman, J. S. (2012). The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nature Protocols*, *7*(8), 1534–1550. https://doi.org/10.1038/nprot.2012.086

Johnstone, T. G., Bazzini, A. A., & Giraldez, A. J. (2016). Upstream ORF s are prevalent translational repressors in vertebrates . *The EMBO Journal*, *35*(7), 706–723. https://doi.org/10.15252/embj.201592759

Kieft, J. S. (2008). Viral IRES RNA structures and ribosome interactions. In *Trends in Biochemical Sciences* (Vol. 33, Issue 6, pp. 274–283). https://doi.org/10.1016/j.tibs.2008.04.007

Koehbach, J., & Jackson, K. A. V. (2015). *Unravelling peptidomes by in silico mining*. *1*(1). https://doi.org/doi:10.1515/ped-2015-0002

Kurosaki, T., & Maquat, L. E. (2013). Rules that govern UPF1 binding to mRNA 3′ UTRs. *Proc Natl Acad Sci U S A*, *110*(9), 3357–3362. https://doi.org/10.1073/pnas.1219908110

Kute, P. M., Soukarieh, O., Tjeldnes, H., Trégouët, D. A., & Valen, E. (2022). Small Open Reading Frames, How to Find Them and Determine Their Function. In *Frontiers in Genetics* (Vol. 12). Frontiers Media S.A. https://doi.org/10.3389/fgene.2021.796060

Lander, S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., … Yeh, R.-F. (2001). Initial sequencing and analysis of the human genome International Human Genome Sequencing Consortium* The Sanger Centre: Beijing Genomics Institute/Human Genome Center. In *NATURE* (Vol. 409). www.nature.com

Mackowiak, S. D., Zauber, H., Bielow, C., Thiel, D., Kutz, K., Calviello, L., Mastrobuoni, G., Rajewsky, N., Kempa, S., Selbach, M., & Obermayer, B. (2015). Extensive identification and analysis of conserved small ORFs in animals. *Genome Biology*, *16*(1). https://doi.org/10.1186/s13059-015-0742-x

Mayr, C. (2017). *Regulation by 3-Untranslated Regions*. https://doi.org/10.1146/annurev-genet-120116

Mayr, C. (2019). What are 3′ utrs doing? *Cold Spring Harbor Perspectives in Biology*, *11*(10). https://doi.org/10.1101/cshperspect.a034728

Meindl, A., Romberger, M., Lehmann, G., Eichner, N., Kleemann, L., Wu, J., Danner, J., Boesl, M., Mesitov, M., Meister, G., König, J., Leidel, S. A., & Medenbach, J. (2023). A rapid protocol for ribosome profiling of low input samples. *Nucleic Acids Research*, *51*(13), E68–E68. https://doi.org/10.1093/nar/gkad459

Mignone, F., Gissi, C., Liuni, S., & Pesole, G. (2002a). *Untranslated regions of mRNAs*. http://genomebiology.com/2002/3/3/reviews/0004.1http://genomebiology.com/2002/3/3/reviews/0004

Mignone, F., Gissi, C., Liuni, S., & Pesole, G. (2002b). *Untranslated regions of mRNAs*. http://genomebiology.com/2002/3/3/reviews/0004.1http://genomebiology.com/2002/3/3/reviews/0004

Mignone, F., Gissi, C., Liuni, S., & Pesole, G. (2002c). *Untranslated regions of mRNAs*. http://genomebiology.com/2002/3/3/reviews/0004.1http://genomebiology.com/2002/3/3/reviews/0004

Mignone, F., Gissi, C., Liuni, S., & Pesole, G. (2002d). *Untranslated regions of mRNAs*. http://genomebiology.com/2002/3/3/reviews/0004.1http://genomebiology.com/2002/3/3/reviews/0004

Milo, R., & Phillips, R. (2015). *Cell Biology by the Numbers.* (1st ed.). Garland Science. .

Mishima, Y., & Tomari, Y. (2016). Codon Usage and 3' UTR Length Determine Maternal mRNA Stability in Zebrafish. *Molecular Cell*, *61*(6), 874–885. https://doi.org/10.1016/j.molcel.2016.02.027

NIH - NHGRI. (2024). *Open Reading Frames*. https://www.genome.gov/genetics-glossary/Open-Reading-Frame

O'Connor, P. B. F., Andreev, D. E., & Baranov, P. V. (2016). Comparative survey of the relative impact of mRNA features on local ribosome profiling read density. *Nature Communications*, *7*(1), 12915. https://doi.org/10.1038/ncomms12915

Opron, K., & Burton, Z. F. (2019). Ribosome structure, function, and early evolution. In *International Journal of Molecular Sciences* (Vol. 20, Issue 1). MDPI AG. https://doi.org/10.3390/ijms20010040

Prensner, J. R., Enache, O. M., Luria, V., Krug, K., Clauser, K. R., Dempster, J. M., Karger, A., Wang, L., Stumbraite, K., Wang, V. M., Botta, G., Lyons, N. J., Goodale, A., Kalani, Z., Fritchman, B., Brown, A., Alan, D., Green, T., Yang, X., … Golub, T. R. (2021). Noncanonical open reading frames encode functional proteins essential for cancer cell survival. *Nature Biotechnology*, *39*(6), 697–704. https://doi.org/10.1038/s41587-020-00806-2

Radhakrishnan, A., & Green, R. (2016). Connections Underlying Translation and mRNA Stability. *Journal of Molecular Biology*, *428*(18), 3558–3564. https://doi.org/https://doi.org/10.1016/j.jmb.2016.05.025

Ryczek, N., Łyś, A., & Makałowska, I. (2023). The Functional Meaning of 5′UTR in Protein-Coding Genes. In *International Journal of Molecular Sciences* (Vol. 24, Issue 3). MDPI. https://doi.org/10.3390/ijms24032976

Schwaid, A. G., Shannon, D. A., Ma, J., Slavoff, S. A., Levin, J. Z., Weerapana, E., & Saghatelian, A. (2013). Chemoproteomic Discovery of Cysteine-Containing Human Short Open Reading Frames. *Journal of the American Chemical Society*, *135*(45), 16750–16753. https://doi.org/10.1021/ja406606j

Shoemaker, C. J., & Green, R. (2012). Translation drives mRNA quality control. In *Nature Structural and Molecular Biology* (Vol. 19, Issue 6, pp. 594–601). https://doi.org/10.1038/nsmb.2301

Sieber, P., Platzer, M., & Schuster, S. (2018). The Definition of Open Reading Frame Revisited. *Trends in Genetics*, *34*(3), 167–170. https://doi.org/10.1016/j.tig.2017.12.009

Slavoff, S. A., Mitchell, A. J., Schwaid, A. G., Cabili, M. N., Ma, J., Levin, J. Z., Karger, A. D., Budnik, B. A., Rinn, J. L., & Saghatelian, A. (2013). Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nature Chemical Biology*, *9*(1), 59–64. https://doi.org/10.1038/nchembio.1120

Somers, J., Pöyry, T., & Willis, A. E. (2013). A perspective on mammalian upstream open reading frame function. In *International Journal of Biochemistry and Cell Biology* (Vol. 45, Issue 8, pp. 1690–1700). Elsevier Ltd. https://doi.org/10.1016/j.biocel.2013.04.020

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., … Zhu, X. (2001). *The Sequence of the Human Genome*.

von Arnim, A. G., Jia, Q., & Vaughn, J. N. (2014). Regulation of plant translation by upstream open reading frames. *Plant Science*, *214*, 1–12. https://doi.org/https://doi.org/10.1016/j.plantsci.2013.09.006

Webster, M. W. , & W. A. (2021). The intricate relationship between transcription and translation. . *Proceedings of the National Academy of Sciences of the United States of America* , *118*(21).

Wu, Q., & Bazzini, A. A. (2018). Systems to study codon effect on post-transcriptional regulation of gene expression. *Methods*, *137*, 82–89. https://doi.org/https://doi.org/10.1016/j.ymeth.2017.11.006

Wu, Q., & Bazzini, A. A. (2023). Translation and mRNA Stability Control. *Annual Review of Biochemistry*, *92*(1), 227–245. https://doi.org/10.1146/annurev-biochem-052621-091808

Wu, Q., Wright, M., Gogol, M. M., Bradford, W. D., Zhang, N., & Bazzini, A. A. (2020). Translation of small downstream ORFs enhances translation of canonical main open reading frames. *The EMBO Journal*, *39*(17). https://doi.org/10.15252/embj.2020104763

Xiao, Z., Huang, R., Xing, X., Chen, Y., Deng, H., & Yang, X. (2018). De novo annotation and characterization of the translatome with ribosome profiling data. *Nucleic Acids Research*, *46*(10), E61. https://doi.org/10.1093/nar/gky179

Appendix

Supplementary Figures

Supplementary Figure 1

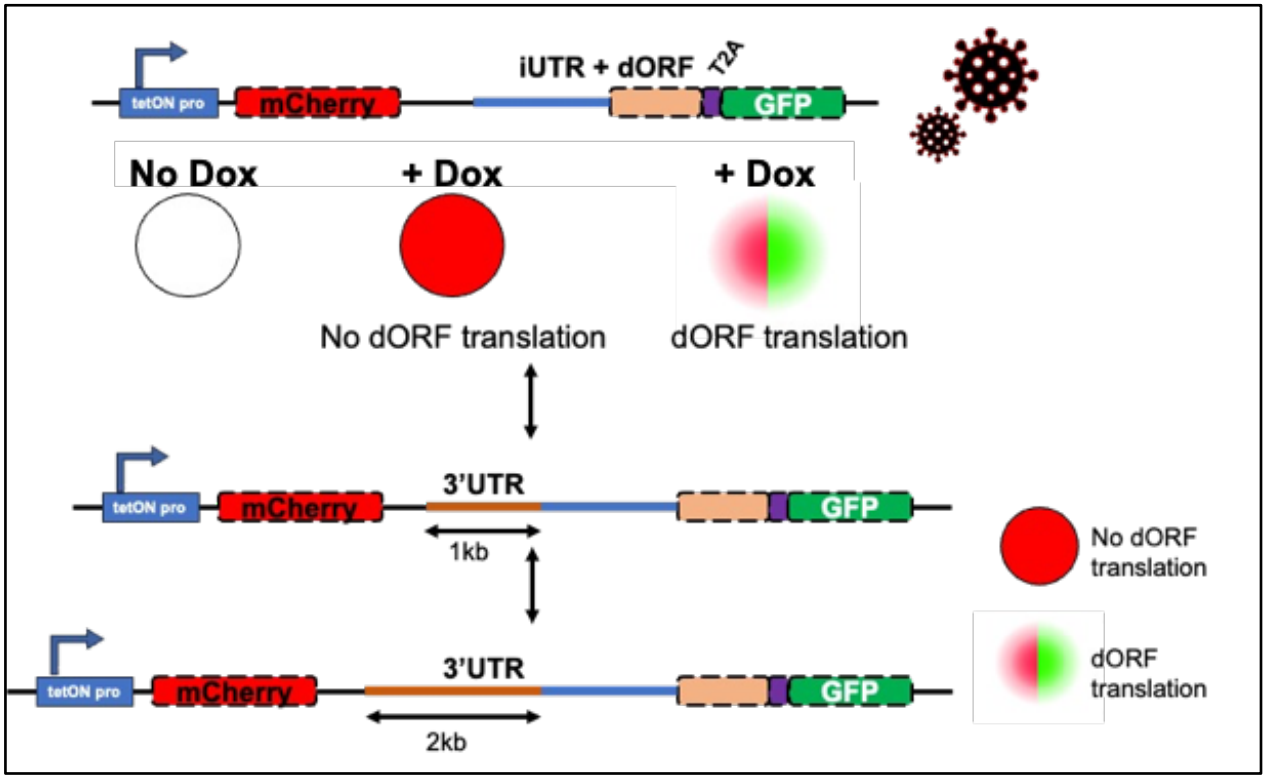

Figure S1 : Schematic explaining the experimental validation of whether long iUTRs get translated or not, to explain result 3.3.

The long iUTRs, along with the dORFs, could be cloned into the lentiviral bi-cistronic plasmids. The plasmid has mCherry as the main ORF and ZsGreen as a system to check dORF translation. It is a tetracycline-inducible system and should not express any colour without doxycycline. However, if it expresses green without doxycycline, that would indicate active transcription in the iUTR.

If it expresses only red with doxycycline, that indicates that the dORF is not getting translated. If it expresses both red and green, that indicates the dORF is translated. In our case, we know that the endogenous iUTR, along with 50AA WT dORF, gets translated and enhances the expression of the main ORF. We also found that when we elongate the iUTR by cloning in either a 1kb or a 2kb fragment of 3'UTR before it. They do not enhance anymore. However, we do not know whether this elongation prevents them from getting translated or not. This losing the enhancement capability

could be due to them not getting translated. However, if they get translated but still do not enhance, that would indicate a structural relevance of this enhancement system, highlighting that enhancement depends on the distance between the main ORF and the dORF.

Supplementary Figure 2

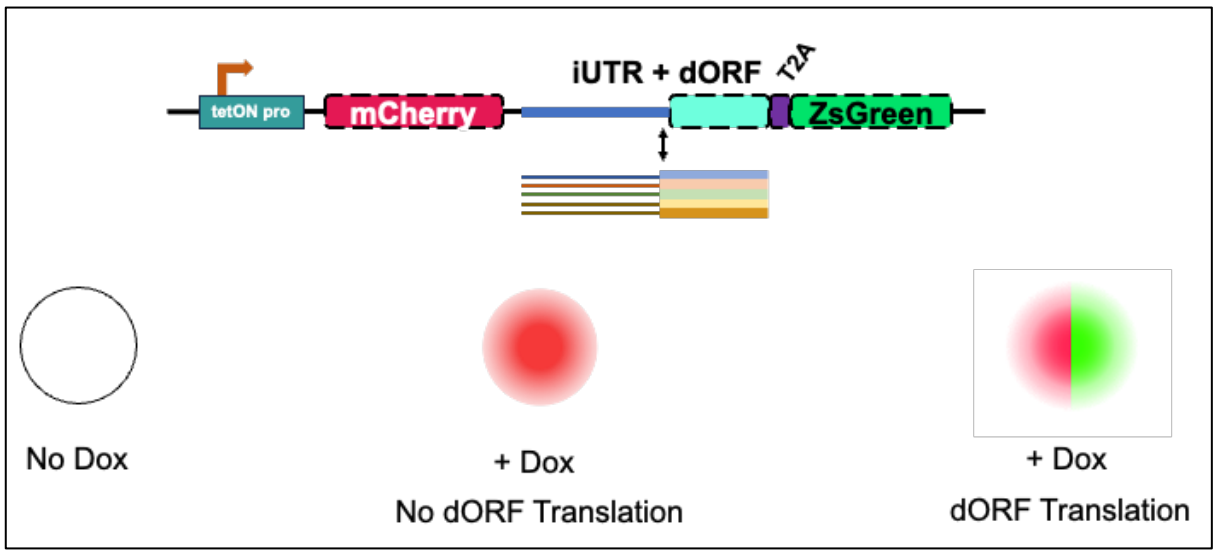

Figure S2 : Schematic explaining the experimental validation of checking whether short dORFs get translated or not.

To check whether the short dORFs that we predicted are getting translated or not, we will order a library containing all the 50 sequences of the short dORFs along with their iUTR. The library will also contain sequences containing mutation at the predicted start site, which will prevent its translation. Now, this library will be cloned into the lentiviral bi-cistronic plasmids. The sequences then can be tested one-by-one in human cells. The dORFs that get translated will express both red and green upon doxycycline treatment. If those positive dORFs lose green when the start site is mutated, then that would validate that the predicted start sites were correct. If that is not the case, then that would mean that the dORF will have alternate start sites that can be validated.

Primers

Serial No.	Name	Sequence
1	4723_RevComp_mCherry_PP1	TCAtttgtaaagttcatccatccc
2	4724_Comp_zsGreen_PP2	ATGgccCAGTCCAAGCACGGCCTGA
3	4725_Comp_mut_zsGreen_PP3	ATGtgaCAGTCCAAGCACGGCC
4	4726_mCherry_CENPAiutr_GibsonFor_PP4	GGATGGATGAACTTTACAAATGACCTGCACCCAGTGTTTCTGTC
5	4727_RevComp_zsGreen_CENPAiutr_Gibson_PP5	GCCGTGCTTGGACTGggcCATGGCTCTGGAGAGTCCCCGG
6	4728_RevComp_MUTzsGreen_CENPAiutr_Gibson_PP6	GCCGTGCTTGGACTGtcaCATGGCTCTGGAGAGTCCCCGG
7	4777_mutGFP_Primer_Rev_PP7	/5Phos/TGAGATCGGAAGAGCACACGTCTG
8	4778_GFP_9bp_PP8	CTGggcCATGGCTCTGGAGAGTCC
9	4779_mutGFP_9bp_PP9	CTGTCACATGGCTCTGGAGAGTCC
10	4780_GFP_15bp_PP10	CTTGGACTGggcCATGGCTCTG
11	4781_mutGFP_15bp_PP11	CTTGGACTGTCACATGGCTCTGGAG
12	4782_GFP_21bp_PP12	GCCGTGCTTGGACTGggcCATG
13	4783_mutGFP_21bp_PP13	GCCGTGCTTGGACTGTCACATGGC
14	4784_GFP_mutGFP_30bp_PP14	CTTGGTCAGGCCGTGCTTGGAC
15	4785_GFP_mutGFP_60bp_PP15	GCCCTCCATGCGGTACTTCATG
16	4786_GFP_mutGFP_90bp_PP16	GGTGATCACGAACTTGTGGCCG
17	4787_GFP_mutGFP_150bp_PP17	CTCCACCACGCACAGGTTGATG
18	4788_GFP_mutGFP_270bp_PP18	GCCGGCGGGGCAGGAGTTCTTG
19	4789_GFP_mutGFP_540bp_PP19	CTGGCAGCGCAAGCGGCC
20	4790_Seq_mCh_CENPA_ZsGreen_PP20	cagaagatggggcacttaaagg
21	4969_GFP_mutGFP_600bp_PP21	GATGAAGTGCCAGTCGGGCATC
22	4970_GibsonFwd_PPARAA_Xba1_50AA_PP22	actatagtgagtcgtattacTCTAGAgctcacacgtatgcattcg
23	4971_GibsonRev_PPARAA_Xho1_50AA_PP23	cttatcatgtctggatctacCTCGAGtgagatgagatgtttctttaag
24	4972_GibsonRev2_PPARAA_1kb_Xho1_50AA_PP24	cttatcatgtctggatctacCTCGAGcccatattaatcaagtagg
25	5098_GibsonFWD_iUTR_long_PP25	ggatggatgaactttacaaaTGAgctcacacgtatgcattcgac
26	5099_GibsonRev_iUTR_long_PP26	GACAGAAACACTGGGTGCAGGtgagatgagatgttctttaag
27	5100_CENPA_iUTR_Rev_PP27	CCTGCACCCAGTGTTTCTGTC
28	5171_Gibson Rev_iUTR_short_PP28	GACAGAAACACTGGGTGCAGGgcccatattaatcaagtagg
29	5213_intron_NMD_Fwd_PP	GTCACACGTCtcagggtgagtttggggaccc
30	5214_intron_NMD_Rev_PP	CGTCTCTAGAcaggacctgtaggaaaaagaag
31	5294_CCDC_LongiUTR_Rev_PP	ggttgtggggaagtgccaggctgagatgagatgtttctttaag
32	5295_CCDC_ShortiUTR_RevPP	ggttgtggggaagtgccaggcgcccatattaatcaagtagg
33	5296_CENPA_ZsG_6AA_PP	GTGCTTGGACTGggcCATGG
34	5297_CENPA_mutZsG_6AA_PP	GTGCTTGGACTGTCACATGGCTCTGG
35	5298_ZsG-mutZsG_40AA_PP	GCCCTTGAAGGGGTAGCCGATG
36	5299_ZsG-mutZsG_60AA_PP	GATGTCCTCGGCGAAGGGCAAG
37	5300_ZsG-mutZsG_120AA	CATGCAGTTCTCCTCCACGCTC
38	5310_CCDC_iUTR_Rev_PP	gcctggcacttccccacaac
39	5681_iUTR_long_short_lenti_Fwd	CTGACGCGTgctcacacgtatgc
40	5682_iUTR_long_short_Lenti_Rev	TCACTGCAGCTCGCCGGTGATCACG