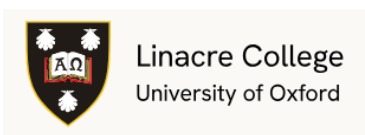
Investigating 5’ untranslated regions and translational regulation across human genes

Nechama Tatz-Wieder
Linacre College, University of Oxford

DPhil Clinical Medicine

Supervisors: Associate Professor Nicola Whiffin, Associate Professor Robert Davies

Word Count: 34,160
Submitted: 20 June 2024



Acknowledgements

This thesis would not have been possible without the contribution, encouragement and support of many people.

First and foremost, I would like to thank my primary supervisor, Assoc Prof. Nicky Whiffin for taking me on as her first DPhil student. I will always be endlessly grateful for being given this opportunity to spend some years training in the most supportive and intellectually stimulating group I could ask for. I have enjoyed the DPhil process immensely and have learnt a tremendous amount. This would not have been possible without Nicky's unwavering support, encouragement and tremendous supervisory skills. I feel extra lucky that Nicky not only allowed but actively encouraged me to pursue my other interests throughout my DPhil, and these activities have been greatly enriching. I am honoured to have been supervised by Nicky, an exemplary role model and scientist, and will try to emulate her empathetic, kind and brilliant management skills in my future roles.

I would like to thank my secondary supervisor Assoc Prof. Robbie Davies, who generously gave me his time and knowledge to perfect some of the techniques and methods in this thesis; it was a pleasure working with you.

Many thanks to Prof Jim Hughes, Prof Andrew Wilkie and Dr Calli Dendrou who acted as the assessors for my transfer of status and confirmation of status examinations, for their valuable time and constructive feedback on my work.

This work would not have been possible without the help and support of my dear friends and colleagues in the lab. I value your tremendous time and efforts you took in providing feedback and help to me over the years. I especially want to mention Dr Alex Martin-Geary, who was there from the very beginning and gave me a tremendous amount of help to get me off my feet. Alex was always there, enthusiastically willing to help wherever she could and was always looking out for me. Many thanks to Elston D'Souza for the many hours he put in helping me with coding troubles and sharing of his amazing knowledge; I would not have managed this without his brilliant input. To Maria, Fred, Ruby, Hyung-Chul, Yuyang, Eloise, Alex B, Anthony and all the other past members of the group, my deepest thanks for always being happy and willing to help me out, for your time, camaraderie and constructive input which greatly improved my work.

I want to thank the Nuffield Department of Medicine for the funding for this DPhil, making it possible for me to do this DPhil. It also enabled me to have the best equipment and opportunity to travel to many conferences, which were highly enjoyable and a tremendous learning experience.

On a personal note, I feel very lucky to be surrounded by a big group of supportive and loving friends and family, who helped make this journey an enjoyable one. Much gratitude to our dear friends Rabbi Eli and Freidy Brackman, for all that you have done to enable Jewish life in Oxford. We are forever indebted to all the help and support you gave us, not least in running a beautiful nursery. Our time spent in Oxford with you is a time we will always treasure. Deepest thanks to my

wonderful parents in law, Danny and Adele Wieder, for their endless support, it means the world to me. Many thanks to my parents, Rabbi Dr Akiva and Suzanne Tatz, for their unwavering support and encouragement in my pursuing higher education, and for all the practical ways in which they help me achieve my goals. Many thanks to my father for sparking my interest in the medical sciences, and for the many valuable formative opportunities which have led to me where I am today. My daughter Minde, whilst doing a PhD with you as a feisty little thing had its challenges, was only made better having you in my life and to share the special opportunities this journey gave us.

Finally, to my husband Gidon. I would not be in this position today, without the endless, unwavering love, enthusiasm and belief in me. There are no words to thank you enough, and I am eternally grateful to have gone on this journey with you by my side.

Statement of Originality

I declare that this thesis constitutes original work.

Throughout the thesis, any data processing or code that has been conducted by others are acknowledged as such. The work presented in this thesis including all the code and analyses to generate text, figures, and tables are a result of my original work, except as specified in the text.

Publications

Content from three of my thesis chapters (chapters 3-5) has been published, and content from chapter 1 is in preparation for submission to a journal. I have received written confirmation from all co-authors of each publication and my supervisor, Assoc Professor Nicola Whiffin, to reuse parts of previously published text and figures in my thesis.

First author

Chapter 1

Wieder, N., D'Souza, E.N., Dawes, R, *et al.* The role of untranslated region variants in Mendelian disease (in preparation)

Chapters 3-5

Wieder, N., D'Souza, E.N., Martin-Geary, A.C. *et al.* Differences in 5'untranslated regions highlight the importance of translational regulation of dosage sensitive genes. *Genome Biol* **25**, 111 (2024). https://doi.org/10.1186/s13059-024-03248-0

Co-author

Martin-Geary AC, Blakes AJM, Dawes R, Findlay SD, Lord J, Walker S, Talbot-Martin J, Wieder N, D'Souza EN, Fernandes M, Hilton S, Lahiri N, Campbell C, Jenkinson S, DeGoede CGEL, Anderson ER, Burge CB, Sanders SJ, Ellingford J, Baralle D, Banka S, Whiffin N. Systematic identification of disease-causing

promoter and untranslated region variants in 8,040 undiagnosed individuals with rare disease. medRxiv [Preprint]. 2023 Sep 12:2023.09.12.23295416. doi: 10.1101/2023.09.12.23295416. PMID: 37745552; PMCID: PMC10516070.

Hardcastle F, Lyle K, Horton R, *et al*. The ethical challenges of diversifying genomic data: A qualitative evidence synthesis. *Cambridge Prisms: Precision Medicine*. 2024;2:e1. doi:10.1017/pcm.2023.20

Conference oral presentations

Translation UK, 2022 - Systematic characterisation of 5'untranslated regions reveals key differences between and across gene-sets

Genomics of Rare Disease, 2023 - The role of start-stop elements in rare disease

Table of Contents

Figures and Tables

Abbreviations

5'TOP	5’terminal oligopyrimidine
ACMG	American College of Medical Genetics
AF	Allele frequency
APA	Alternative cleavage and polyadenylation
API	Application programming interface
ARE	AU-Rich Elements
ASO	Antisense oligonucleotides
bp	Base pairs
CADD	Combined Annotation Dependent Depletion
CAGE	Cap Analysis of Gene Expression
CDS	Coding sequence
CLIP-Seq	Cross-linking and immunoprecipitation followed by sequencing
CNV	Copy-number variants
COSMIC	Catalogue Of Somatic Mutations In Cancer
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DDD	Deciphering Developmental Disorders
DDG2P	Developmental Disorders Genotype-Phenotype Database
DNA	Deoxyribonucleic acid
dORF	Downstream open reading frame
EBI	European Bioinformatics Institute
eIF	Eukaryotic initiation factors
EMBL	European Molecular Biology Laboratory
FA	Fanconi Anaemia
FACS-seq	Fluorescence-activated cell sorting and high-throughput DNA sequencing
FANTOM	Functional ANnoTation Of the Mammalian genome
GEL	Genomics England Limited 100,000 Genomes Project
gnomAD	Genome Aggregation Database
GTEx	Genotype-Tissue Expression project
GWAS	Genome wide association study
HGNC	HUGO Gene Nomenclature Committee
HHCS	Hyperferritinemia/cataract syndrome
HPO	Human Phenotype Ontology

HS	Triplosensitive
IRE	Iron response element
IRES	Internal ribosome entry site
IRP	iron-regulatory protein
LOEUF	Loss-of-function observed/expected upper bound fraction
LoF	Loss of function
MANE	Matched Annotation from the NCBI and EMBL-EBI
MAVE	Multiplexed assays of variant effect
MFE	Minimum free energy
miRNA	microRNA
mRNA	Messenger ribonucleic acid
NCBI	National Center for Biotechnology Information
NEM8	Nemaline myopathy-8
NF1	Neurofibromatosis Type 1
NF2	Neurofibromatosis Type 2
NHS	National Health Service
NMD	Nonsense mediated decay
NTE	N-terminal extension
Onc	Oncogene
oORF	Overlapping open reading frame
ORF	Open reading frame
PAS	polyA site
PCH	Pontocerebellar hypoplasia
PhyloP	Phylogenetic P-values
pLoF	Predicted loss of function
PTB-1	Polypyrimidine tract-binding protein 1
RAN	Repeat associated non-ATG
RBP	RNA binding protein
Ribo-seq	Ribosome sequencing
RNA	Ribonucleic acid
RPF	Ribosome protected fragment
SBP2	SECIS RNA binding protein
SCA8	Spinocerebellar ataxia type 8
SECIS	Selenocysteine insertion sequence
SEP	smORF encoded peptide
smORF	Small open reading frame
TAD	Topologically associated domains
tAI	tRNA adaptive indices
TE	Translational Efficiency
TF	Transcription factor

TPM	Transcript per million
tRNA	Transfer ribonucleic acid
TS	Haploinsufficient
TSG	Tumour suppressor gene
TSS	Transcription start site
uORF	Upstream open reading frame
UTR	Untranslated region
VEP	Variant effect predictor
VUS	Variant of unknown significance
XMEA	X-linked myopathy with excessive autophagy
YB-1	Y-box binding protein 1

Abstract

Untranslated regions (UTRs) are important mediators of post-transcriptional regulation. The length of UTRs and the composition of regulatory elements within them are known to vary substantially across genes, but the reasons remain poorly understood in humans. Here, I investigated whether this variation, specifically in 5'UTRs, correlates with gene dosage sensitivity and whether it differs between different categories of disease genes.

I investigated 5'UTR length, the number of alternative transcription start sites (TSS), the potential for alternative splicing, the number and type of upstream open reading frames (uORFs) and the propensity of 5'UTRs to form secondary structures. I explored how these elements vary by gene tolerance to loss-of-function (LoF; using the LOEUF metric), and in genes where changes in dosage are known to cause disease. I found that LOEUF correlates with 5'UTR length and complexity. Genes that are most intolerant to LoF have longer 5'UTRs ($P<1x10^{-15}$), greater TSS diversity ($P<1x10^{-15}$), and more upstream regulatory elements compared to LoF tolerant genes. These differences are evident in disease gene sets but not in recessive developmental disorder genes where LoF of a single allele is tolerated.

Finally, I focused on a single category of 5'UTR regulatory elements, start-stops, which have only recently been described. I analysed variants within them using

ClinVar and the Genomics England 100,000 Genomes Project (GEL). I identified candidate variants for future functional studies and highlight the complexity of interpreting UTR variants.

These results underscore the importance of 5'UTRs in post-transcriptional regulation through tightly regulating mRNA and protein levels, particularly in genes susceptible to changes in dosage. Including UTR analysis in diagnostic pipelines should enhance diagnoses, though challenges in interpreting UTR variants remain. This work enhances our understanding of 5'UTR features and their role in disease.

Chapter 1
An Introduction to UTRs: features within them and their role in rare disease

1.1 Non-coding regions and Gene Expression

The central dogma of DNA, which encompasses the process from DNA to RNA to protein, is a highly regulated, complex and tightly controlled process. It involves the DNA sequence itself and external factors that control gene expression. The gene sequences that encode directly for protein are known as the coding regions, whereas the other essential DNA sequences which do not directly give rise to the protein are known as the non-coding regions. Coding and non-coding regions, as well as external factors like proteins that bind and influence specific regions of DNA, are all essential for controlling gene expression and leading to the correct and timely amount of protein a cell needs for essential function. These sections describe function in humans unless explicitly stated otherwise.

1.1.1 Transcription

Facilitating transcription of DNA into RNA requires several essential factors. Promoter elements are directly upstream of a gene [1], and there are enhancer and silencer elements which range from being proximal to very distant from the gene, all of which bind transcription factors (TFs) [2,3]. TFs are proteins that, once bound, stimulate or block transcription depending on their type and the location they bind to. Insulator

elements are DNA-protein complexes that are able to block enhancers or have silencing effects [4]. Each gene is also situated within a broader regulatory framework; chromosomes fold within cell nuclei leading to the genome being organised into domains of preferential chromatin interactions called topologically associated domains (TADs) [5]. Together, this organisation into TADs with specific enhancer and insulator elements ensures the correct quantity of gene is transcribed into RNA at the correct time and in the right tissue.

1.1.2 Post-transcriptional regulation and splicing

Once transcribed into pre-messenger RNA (pre-mRNA), introns within the gene require removal to form mature mRNA, which comprises only the exonic regions. Splicing of introns is directed by specific sequences within the intronic sequence and the sequences near the exon boundary [6]. Mature mRNA comprises the coding sequence of the gene (CDS) flanked by the untranslated regions (UTRs) which play a vital role in regulation of protein production by controlling transcript stability and the rate and location of protein synthesis [7]. mRNA has specific modifications such as the 5'cap which prevents it from being degraded and recruits the ribosomes to bind to the mRNA to translate the CDS into protein product [8]. The addition of the poly(A) tail to the 3' end of the mRNA has an important role in nuclear export and affects the stability of mRNA [9]. Various elements such as enzymes and non-coding RNA (ncRNA) within the nuclei also exert effects on the fate of mRNA by either trafficking it to specific locations or by degrading it [10,11].

1.1.3 Translation

Translation of mRNAs into protein is a complex, not fully elucidated, process requiring many molecules. Translation typically begins with recruitment of the ribosome at the 5'cap of the mRNA, ribosomal scanning along the mRNA in a 5' to 3' direction (i.e. through the 5'UTR), and initiation of translation at a start codon, which is usually an AUG [12]. The ribosome is a macromolecular complex with both RNA and protein components, consisting generally of a large and small subunit. Eukaryotic initiation factors (eIFs) are involved in ribosomal initiation on the mRNA. eIF4F, along with other initiation factors, recognises the 5'cap to recruit the small ribosomal subunit to form the pre-initiation complex. Other eIFs deliver the methionine-bound tRNA to the ribosome. The ribosome then scans in a 5' to 3' direction until a start codon is identified. The large ribosomal subunit then joins the complex and elongation commences. Elongation involves the sequential addition of amino acids as per the codons the ribosome is scanning, and the accurate pairing of complementary tRNA molecules with these codons within the ribosomal complex. When the translating ribosome encounters a stop codon, specific proteins called release factors recognise these and trigger termination, where the ribosome undergoes conformational change to release the synthesised protein and disassembly of the translation complex (reviewed in [13]).

1.2 UTR sequences are important regulatory elements

Protein production is a fine-tuned process in cells; too much or too little can disrupt cellular processes and lead to disease. 5' and 3'UTR sequences together play a vital

role in regulation of protein production by controlling transcript stability and the rate and location of protein synthesis. 5' and 3' UTRs are not translated into part of the protein but mediate post-transcriptional regulation via linear and structural elements, controlling RNA-stability, cellular localisation, and translation. A diagram of UTRs and elements that can be found within them is provided in Figure 1.1.

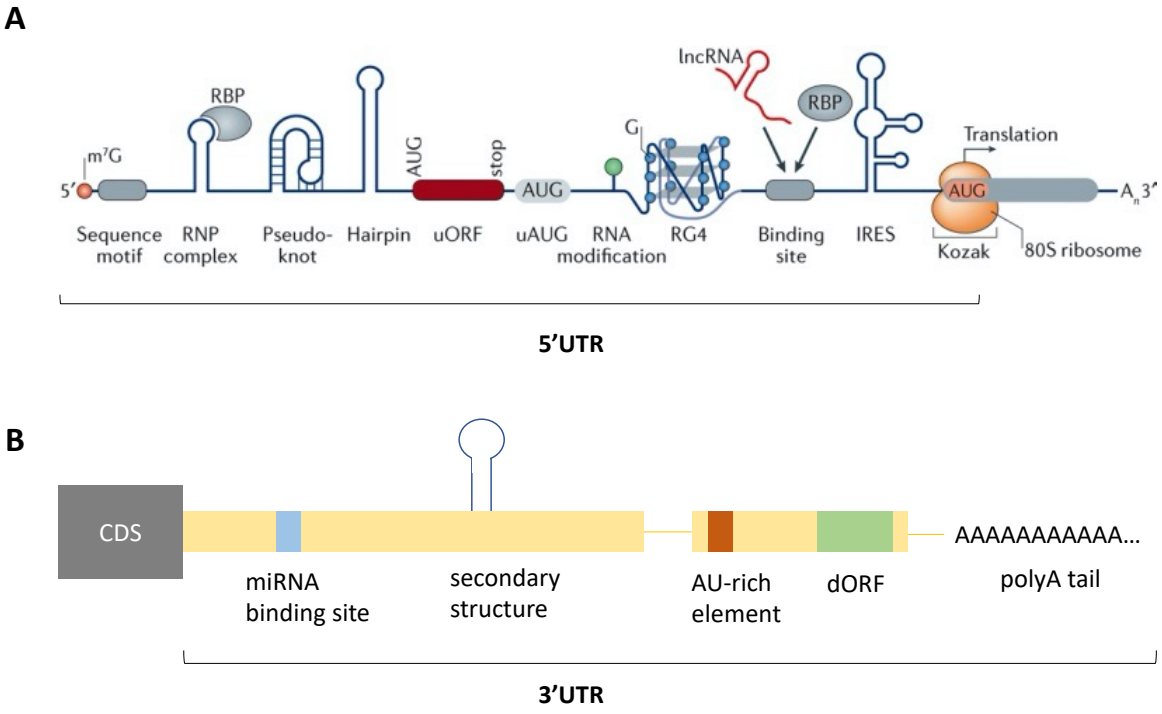


Figure 1.1: UTR structure and features that can be found within them
A) *5'UTR structure. Abbreviations L-R (m7G, 7-methyl-guanosine cap; hairpin, hairpin-like secondary structures; uORF, upstream open reading frame; IRES, internal ribosome entry site; UTR, untranslated region). Image obtained from [14], Creative Commons Attribution 4.0 International License.* **B)** *3'UTR structure. Abbreviations L-R (CDS, coding sequence; miRNA binding site, microRNA binding site; dORF, downstream open reading frame; polyA tail, polyadenylation tail).*

UTRs contribute to mRNA diversity through alternative splicing and alternative transcription start site (TSS) usage. The definition of UTRs varies depending on the

transcript; there can commonly be multiple isoforms with distinct UTRs but identical CDS [15,16]. Compared to CDS, UTRs generally are not well conserved across species; however, there can be regions which are highly conserved [17,18]. Additionally, 3'UTR sequence space has substantially expanded during the evolution of higher organisms and correlates with cellular complexity of organisms. This suggests that 3'UTRs play an important role in regulation of biological complexity [18]. While UTRs have long been named "untranslated", we now know that regulatory elements within both 5' and 3' UTRs may undergo active translation. Consequently, these regions may more appropriately be referred to as 'leader' (5'UTR) and 'trailer' (3'UTR) sequences [19]. However, for this thesis I use the term UTRs.

Below I detail the role of different regulatory elements in UTRs.

1.3 5'UTRs

The primary function of 5'UTRs is in translational regulation. This process is regulated through key sequence elements such as upstream open reading frames (uORFs) and structural features which influence the amount and speed of protein production, usually by interacting with, or pausing the scanning ribosome [20]. The length of 5'UTRs can vary between genes and they can be as long as over 2,000 base pairs (bp) [7]. There is decreased translational efficiency of mRNA's with long 5'UTRs, including oncogenes, transcription regulators and chromatin modifiers. These genes may require tighter control of their expression [21,22]. An in vitro experiment demonstrated that 5'UTR length limits mRNA translational output in HeLa cells [23].

The length of the 5'UTR as well as the amount and type of various translational control elements varies between genes [24], providing each gene with the correct combination of required elements for normal protein production. Approximately 38% of 5'UTRs contain introns, with the number of introns ranging up to 11 in some 5'UTRs [24]. Alternative splicing within the UTRs occurs in transcripts of at least 13% of mammalian genes [25,26] and alternative splicing of the 5'UTR is known to impact mRNA stability and translation [27,28]. Alternative splicing and different TSS usage can lead to diverse transcript 5'UTR isoforms containing different combinations of translational control elements, enabling different levels of protein translation across tissues and developmental stages [29,30].

1.3.1 5'cap

mRNAs contain an N7-methylated guanosine linked to the first nucleotide, known as the 5'cap. It is essential in cap-dependent initiation of translation (the ribosome recognises it to bind), and it also protects from exonuclease cleavage [8].

1.3.2 uAUGs and upstream open reading frames (uORFs)

Upstream AUG (uAUG) triplets are commonly observed within 5'UTRs. uAUGs, or other near-cognate codons (most commonly CUG), may be recognised by the scanning ribosome as start codons which can initiate translation [31]. Translation from a uAUG may have one of multiple effects (Figure 1.2). uORFs are encoded when a uAUG has an in-frame stop codon within the 5'UTR. Translation of a uORF can be followed either by ribosome dissociation from the mRNA (therefore decreased CDS translation), or continued scanning before reinitiation of translation at the

downstream CDS [32,33]. There is some evidence from mass spectrometry that some uORFs encode a detectable peptide product (SEPs; smORF encoded peptides) [34], but further work needs to be done to find and curate these and to understand their role. If there is no in-frame stop codon, translation from a uAUG would overlap the CDS. If this translation is in-frame with the CDS, it will result in an N-terminally elongated protein (N-terminal extension, NTE), also known as an in-frame overlapping ORF (oORF). Alternatively, if the uAUG is out-of-frame with the CDS, an out-of-frame oORF will be translated out-of-frame with the CDS, terminating within the CDS or past the CDS stop codon [35–37]. This would yield a non-functioning peptide. A start-stop is an upstream start codon immediately followed by a stop codon, with no codons in between. They have also been referred to as the "minimum ORF" [38–40]. They are a less known and characterised 5'UTR feature but initial work has shown that start-stops are thought to cause ribosome pausing and decrease CDS translation [39]. uAUGs are generally assumed to repress CDS translation, and active translation of uORFs can reduce downstream translation by up to 80% [37]. The prospect of a uAUG initiating translation by influencing the recognition of the start codon by the ribosome, is dependent on the local sequence context, known as the Kozak consensus sequence [41,42]. The consensus sequence is the specific nucleotide sequence surrounding the start codon. A strong Kozak sequence includes a purine (A or G) at the -3 position and a G at the +4 position, relative to the A of the start codon. Around 43% of genes have one or more uAUGs in their 5'UTR [24] and these uAUGs are conserved to a significantly greater degree than any other triplet in 5'UTRs [43].

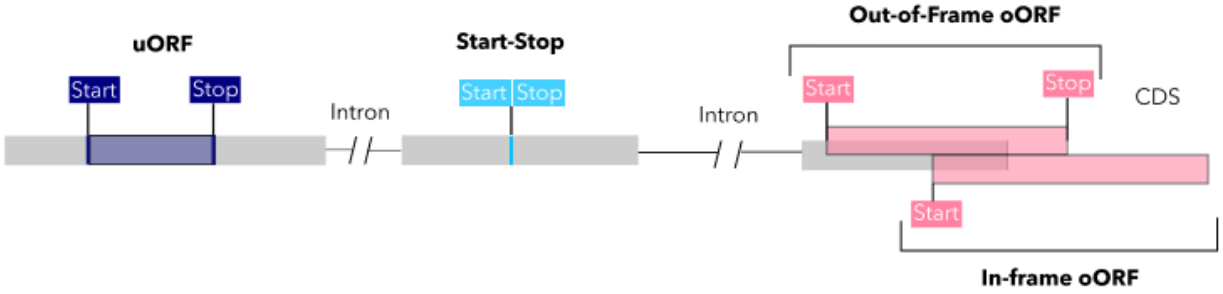


Figure 1.2: *Illustration of the different types of upstream start codons. uORFs: upstream open reading frame; oORF: overlapping open reading frame. Image created by Elston N. D'Souza and used with permission.*

smORFs (small open reading frames), of which uORFs are a subset, have been more extensively studied in organisms such as yeast, bacteria and plants, and experimental characterisation has shown that hundreds of smORFs can produce a phenotype [44–46]. A very well studied example of uORF activity is in the stress response gene *GCN4* in yeast (*ATF4* in vertebrates). The two uORFs in *GCN4* repress translation of the CDS until the cell encounters stress conditions. Under stress, eIF2 (an initiation factor that is part of the ribosome) becomes phosphorylated and therefore reduces its levels. This increases the time for the scanning ribosomes to become competent to initiate translation. Therefore, after translating the first uORF, the ribosomes cannot reinitiate at the second and scan through it, instead reinitiating at the CDS, leading to increased expression. *GCN4* is involved in repairing cellular stress damage [47].

1.3.3 Internal Ribosome Entry Sites

A subset of mRNAs can initiate translation in a cap-independent manner via internal ribosome entry sites (IRESs). These are specialised sequences within the 5'UTR that can directly recruit 40s ribosomal subunits to initiate scanning from within the 5'UTR independently of the 5'cap, which allows for translation to occur under stress conditions where cap-dependant translation is suppressed [48,49]. IRESs were first discovered in picornaviruses and subsequently found in other viruses. Some IRESs have been reported in eukaryotic mRNAs, however the evidence supporting their widespread existence in humans is limited [50]. An example of an IRES studied in humans is in *p53*. There are two IRESs; one in the 5'UTR and one in the CDS. These are purported to mediate translational regulation of p53 mRNA and its differing isoforms which determine different signalling pathways [51]. Notably, IRES motifs differ between genes so are very difficult to predict, limiting our ability to annotate and interpret variants in these elements.

1.3.4 Secondary Structure

mRNA is a single stranded RNA sequence capable of folding and forming secondary structures through complementary base-pairing with itself. These structures can occur within the 5'UTR and can impact translation by causing inefficient ribosomal scanning [14,52] and are negatively correlated with mRNA stability [53]. There are several types of secondary structures, including pseudoknots, G-quadruplexes and hairpin structures (also referred to as stem-loops), which occur when the mRNA strand folds and base pairs with an adjacent section. It is difficult to predict the exact

conformation and structure of mRNA based on sequence alone as secondary structures are dynamic [54]. One of the best studied examples of an important stem-loop is the iron responsive element (IRE) which controls translation of mRNAs important for iron homeostasis. Cellular iron uptake must be tightly regulated as insufficient or excess levels of iron can be damaging. A single conserved IRE stem-loop close to the 5'cap of mRNA in some iron-related genes such as the L-ferritin (*FTL*) gene is bound by iron-regulatory protein 1 or 2 (IRP1/IRP2). IRP binding represses translation initiation by blocking ribosome access to the 5'UTR. The IRPs register iron availability through direct interactions with iron in the cytosol and alter their binding and hence translational inhibition in response [55].

1.3.5 5'TOP motifs

A small subset of mRNAs have specific motifs within the 5'UTR called the 5'terminal oligopyrimidine (5' TOP) motif, which is 4-15 pyrimidine bases at the 5'end of the mRNA immediately after the 5'cap. Most genes containing the 5'TOP motif encode ribosomal proteins and translation factors; proteins that are essential for protein synthesis. Ribosomes and translation factors are synthesised according to cellular demands but under suboptimal conditions translation is paused. The 5'TOP motif allows cells to react to adverse conditions and quickly modulate expression of ribosomes and translation factors (reviewed in [56,57]). There is a complex regulatory circuit for 5'TOP transcripts [56]. 5'UTRs can also contain other binding sites for microRNAs (miRNAs) [58] and RNA binding proteins (RBPs) but these are mainly a feature of 3'UTRs so are discussed in the following section.

1.4 3’UTRs

The primary roles of 3’UTRs are regulating the stability of the mRNA molecule, the rate at which it is degraded, and where it is located within the cell, although, similar to 5’UTRs, they do also play a role in regulation of translation (reviewed by Mayr [59]). 3’UTRs help stabilise the mRNA and prevent mRNA from being degraded prematurely [18]. Much of this regulation by 3’UTRs is mediated through interactions with additional factors, namely miRNAs and RBPs that bind to motifs and structural elements within 3’UTRs to mediate their effect. While miRNAs are generally repressive, RBPs have a range of diverse regulatory roles [60]. As with 5’UTRs, the precise combinations of regulatory elements vary widely between genes and between different transcript isoforms of the same gene. Alternative polyadenylation creates distinct transcript isoforms with 3’UTRs of differing length, containing different numbers of binding sites for regulatory miRNAs and RBPs, which adds another layer of control [61]. While splicing is common in 5’UTRs, only around 6% of 3’UTR sequences contain introns [62]. The vast majority of these introns are very close to the CDS stop codon as introns >50/55 bps downstream of the CDS are predicted to cause transcript degradation through the nonsense mediated decay (NMD) pathway [63].

1.4.1 Polyadenylation

For most protein coding genes, the 3’ end of the pre-mRNA is formed by consecutive cleavage and polyadenylation that occurs co-transcriptionally. This is a widely studied phenomenon and is reviewed by Curinha *et al.* [64], and summarised here. 3’UTRs

contain polyadenylation sites (polyA site (PAS)) that directs addition of several hundred adenine residues to create the polyA tail at the end stage of transcription. The polyA tail has an important role in nuclear export, translation, and the stability of mRNA. Usage of a specific PAS in the pre-mRNA is directed by RNA cis-elements and several trans-acting factors. The most important cis-element is the polyA signal, a hexamer (usually AAUAAA or a close variant such as UAUAAA) [65] located ~10-35 bp upstream of the PAS. Many genes use alternative cleavage and polyadenylation (APA) which generates transcripts with differing 3'UTRs but producing the same protein [66,67]. An example where APA yields mRNAs with distinctly different demonstrated biological function is in the *CD47* gene. The longer 3'UTR isoform enables efficient cell surface expression of CD47 protein and the shorter 3'UTR primarily localises CD47 protein to the endoplasmic reticulum. Therefore, the CD47 protein has different functions depending on which 3'UTR isoform generated it [68]. APA adds to functional diversity of the proteome without changing the amino acid sequence.

1.4.2 microRNAs binding sites

microRNAs (miRNAs) are RNAs ~22 bp long and are involved in RNA silencing and post-transcriptional regulation of gene expression. miRNAs base pair to complementary sequences on the mRNA (usually the 3'UTR) and can cleave the mRNA strand in two or shorten the poly(A) tail, both leading to silencing and inhibition of protein production [69]. An example of this is miR-196 which has complementarity to a group of *HOX* genes and exerts translation effect by inhibiting protein expression

and cleaving the mRNA. This is a miRNA-mediated posttranscriptional restriction of *HOX* gene expression, which is important in early development [70].

1.4.3 Secondary Structures

Secondary structures can also be present in 3'UTRs; however, in 3'UTRs they are associated with longer half-lives and higher abundance of mRNAs [53].

1.4.4 AU-Rich Elements (AREs)

A small subset of 3'UTRs also contain AU-rich elements (AREs). AREs are regions with adenine and uridine bases which bind various proteins; some of which when bound stabilise the mRNA and some destabilise it leading to rapid degradation [71,72].

1.4.5 Downstream ORFs

Downstream open reading frames (dORFs) are open reading frames within the 3'UTR. Some studies have shown that these are translated [31] but little is known about them. There is some evidence that translation of a dORF enhances upstream translation of the main CDS [73]. How translation occurs past the stop codon of the CDS is intriguing and there is much to understand here still.

1.5 Techniques to annotate UTR regulatory elements

There are several techniques to functionally evaluate UTR regulatory elements.

1.5.1 Ribosome Sequencing (Ribo-seq)

Ribosome sequencing (Ribo-seq) is a ribosome profiling technique which chemically immobilises translating ribosomes in situ on the mRNA. This provides a “snapshot” of all active ribosomes in a cell at a specific time. The mRNA fragments, known as ribosome protected fragments (RPF), within the bound ribosome are then sequenced. Using computational tools, the RPF are then aligned to reference genomic sequences to identify their position in the genome. This provides a translational profile - what parts of the genome at that time point underwent translation [74]. Ribo-seq has been an essential tool in detecting translation of small open reading frames such as uORFs [31]. It is worth noting that this will only capture translating ribosomes at a specific time point in a specific cell type.

1.5.2 Detecting IRES

Bicistronic reporters can be used to confirm or validate IRES activity. They typically contain two reporter genes, one upstream and one downstream of the IRES of interest. The upstream gene is translated via a cap-dependent mechanism and the downstream gene is translated via a cap-independent mechanism, relying on the presence of the IRES. The two reporter genes used are usually genes that encode fluorescent proteins such as luciferase, which can then be quantified using a

luminescence assay. Depending on which of the fluorescent proteins are expressed will indicate whether there is IRES activity or not [75].

1.5.3 Identifying polyA site usage

There are many techniques to identify poly(A) sites. Most isolate mRNA and identify at which position the poly(A) tail was added. mRNA with poly(A) tails are differentiated from other RNAs by using chemicals that specifically adhere to poly(A) tails. Reverse transcription is then performed to generate the complementary DNA using a poly(T) primer which binds to the poly(A) tail. The mRNA can then be sequenced and bioinformatically aligned to a reference sequence to identify where in the genome the poly(A) site is [76,77].

1.5.4 Detecting miRNA binding sites

To determine miRNA binding site, the 3'UTR of the mRNA containing the putative miRNA binding site can be cloned downstream of a luciferase reporter gene. This is transfected into cells containing the miRNA of interest. If there is a reduction in luciferase activity this indicates miRNA mediated repression [78]. Another in vivo method is cross-linking and immunoprecipitation followed by sequencing (CLIP-Seq). This method involves cross-linking (chemically fixing) RNA binding proteins to their RNA targets [79]. miRNAs are often associated with certain proteins, for example Argonaute proteins which repress expression [80]. These associated proteins are targeted in immunoprecipitation with antibodies, which isolates and purifies proteins of interest from a mixture of molecules. Sequencing of the RNA bound by the proteins

can be carried out and mapped to the genome to determine the position of the binding sites.

1.6 UTRs in rare disease

Ascertaining the genetic basis of disease is important in the clinical management of patients, potentially provides therapy targets, and informs patients and their families about reproductive decisions. To date, clinical genetics has mostly focused on variants in the coding regions of the genome, with associated genetic diagnostic rates for severe rare diseases of only ~30-50% [81–83]. There are numerous potential areas where diagnostic gaps might occur. Some are inherent to the sequencing methods used (such as exome sequencing, short or long read sequencing), and the quality and biases of the reference genome sequences are aligned to. Additionally, interpreting variants can be challenging as our understanding of all genes' roles and mechanisms are not complete. Although it is known that variants in non-coding regions can also cause disease [84], this is a comparatively understudied area. Indeed, only recently have recommendations been published that enable routine clinical classification of variants in non-coding regions [85]. To aid in this effort, my research focuses on UTRs. Understanding functional regulatory elements within UTRs sequences is pertinent in understanding their role in translation and function, and how perturbation can cause disease. By increasing knowledge of these regions, we can begin to incorporate analysis of them in clinical genetics and improve diagnostic rates [86,87].

Given UTRs' crucial roles in gene regulation, perturbing one or more of the regulatory elements within UTR sequences through genetic variation can have a dramatic impact on protein production and lead to severe disease (Figure 1.3). The various underlying mechanisms that have been uncovered to date are reviewed in the following sections. All example variants are listed in Table 1.1, in section 4.11.

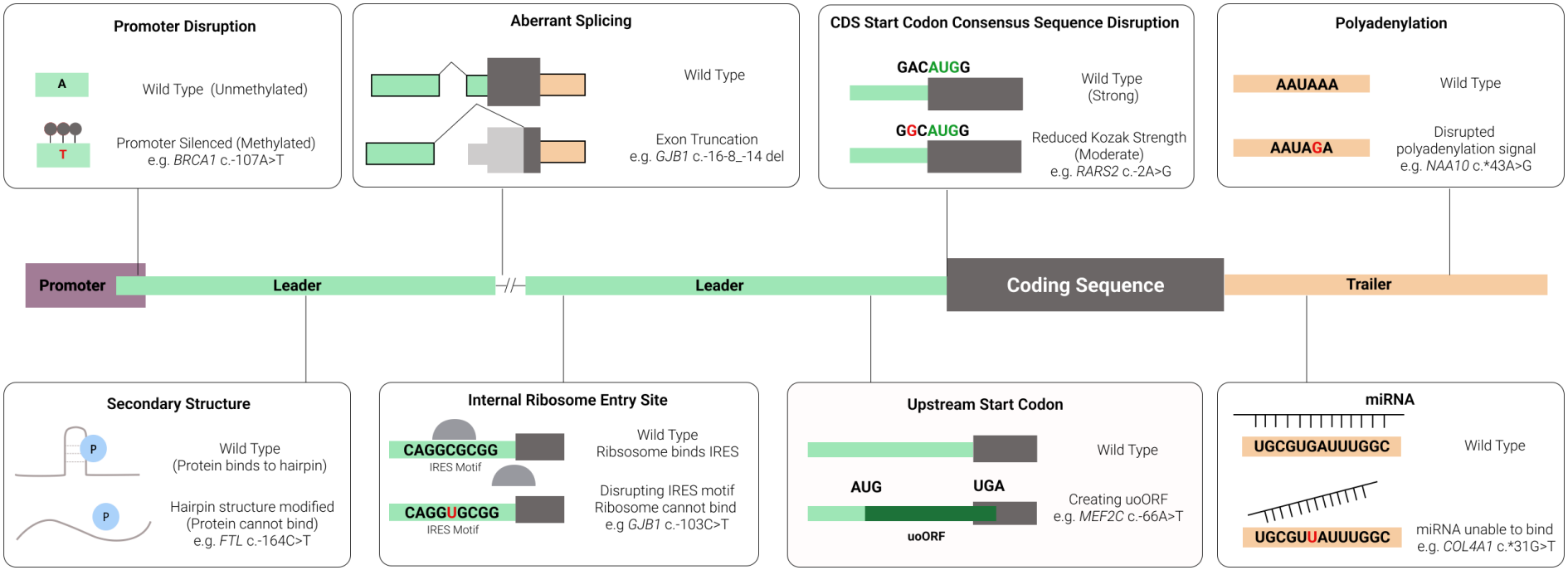


Figure 1.3: Illustration of UTR regulatory elements and examples of disruptions that lead to disease
Gene and protein expression is a tightly controlled process. Elements on the DNA level (e.g. promoter) and mRNA level (e.g. uAUG codon) influence this process. When these elements are disrupted, they can lead to disease.

1.6.1 Variants that create upstream start codons decrease normal CDS translation

Creation of new uAUGs has been shown to be under strong negative selection. In particular, variants that create oORFs or NTEs are on aggregate as deleterious as missense variants [88].

Variants that create out-of-frame oORFs have been found across a range of diseases, including in the recessive gene *SLC22A* causing carnitine deficiency where multiple patients were compound heterozygous, with one uAUG-creating variant in the 5'UTR and one CDS variant both predicted to reduce protein levels [89]. Other examples include *MEF2C* and *NF1*, where oORF-creating variants were found to cause severe developmental disorders [86] and neurofibromatosis type I [88], respectively. In each case, translation of an oORF leaves the ribosome unavailable to translate the CDS. If the oORF is translated at high levels due to creation of an upstream start codon into a favourable consensus sequence context, this can cause a complete loss of CDS translation. Conversely, if initiation of translation at the upstream start codon is incomplete, due to a process termed 'leaky scanning' [20], the variant will only result in a partial decrease in protein levels. Whether such hypomorphic variants have a large enough effect to cause disease is gene dependent, complicating variant interpretation.

uORF-creating variants have also been reported to cause disease, again acting to lower CDS translation. Examples include variants in *NIPBL* causing Cornelia de

Lange syndrome [90] and in *TWIST1* causing Saethre-Chotzen syndrome [90,91]. uORF-creating variants are, however, even more difficult to interpret as the effect of introducing a new uORF on CDS translation is hard to predict. Not only does the uAUG context impact the strength of translation of the uORF, but also the length of the uORF created, and the distance from the end of the uORF to the CDS affects the chance of the ribosome re-initiating translation downstream at the CDS start codon. Notably, the interpretation of all uAUG-creating variants is further complicated by the complex makeup of 5'UTR sequences. For example, if a uAUG is created within an already highly translated uORF, it will likely not be "seen" by scanning ribosomes as a potential start site.

While uORF and oORF creating variants impact translational regulation, variants that create uAUGs in-frame with the CDS that result in NTE may be more likely to disrupt protein function. For example, a variant that extends the N-terminus of *IFITM5* by five amino acids was found repeatedly in individuals with osteogenesis imperfecta type V [92]. In this case, the addition of the extra amino acids renders the protein non-functional. Similarly, two distinct variants that add three and nine amino acids, respectively, to the start of MEF2C protein cause MEF2C haploinsufficiency and severe developmental disorders [86]. *MEF2C* cannot tolerate addition of amino acids at the N-terminus as this disrupts the binding of this transcription factor to DNA, abolishing its function.

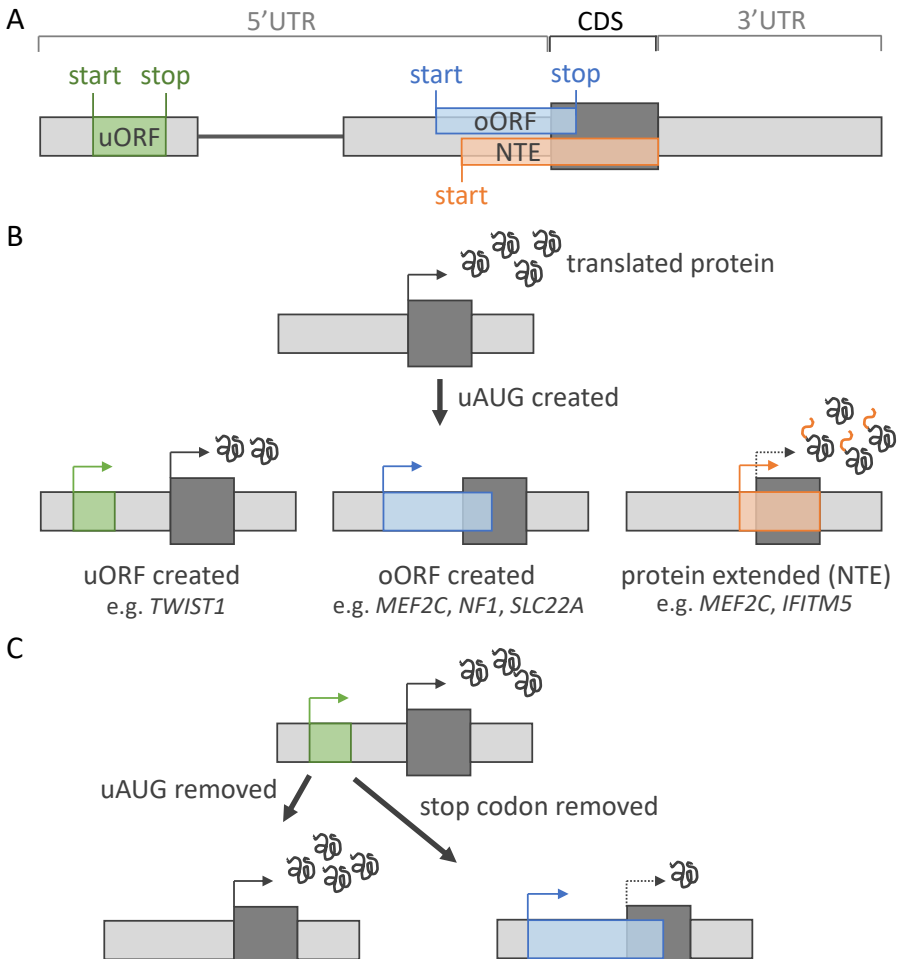


Figure 1.4: Creation and disruption of upstream open reading frames (uORFs) in disease

A) *Schematic of the different elements encoded by an upstream start codon. uORFs are encoded by an upstream start with an in-frame stop codon also within the 5'UTR. When there is no in-frame stop within the 5'UTR either an upstream overlapping ORF (oORF) is formed when the start codon is in a different reading frame to the coding sequence (CDS), or the CDS is extended at the N-terminus (N-terminal extension; NTE).* ***B)*** *Depiction of the different impacts of variants that create uORFs, oORFs, or NTEs on CDS translation.* ***C)*** *Depiction of two types of uORF-perturbing variants on CDS translation. Removing the start codon of a uORF (left) is predicted to increase CDS translation whereas removing the stop codon of a uORF resulting in its extension to form an oORF (i.e. there is no alternative in-frame stop codon) is predicted to decrease, or abolish, CDS translation. Figure created by Nicola Whiffin and used with permission.*

1.6.2 Variants that impact existing uORFs or uoORFs disrupt translational regulation

The above examples of uAUG-creating variants are all predicted to decrease CDS translation. Conversely, removing an existing upstream start codon is more likely to result in an increase in protein levels and have a gain-of-function effect. Upstream start-codon removing variants in *EPHB1* (breast and colon) and *MAP2K6* (colon) in cancer samples (removing an oORF and uORF, respectively) were found to be associated with enhanced translation, suggesting that loss-of-uAUG mediated translational increase of the downstream main protein coding sequence may contribute to carcinogenesis [93].

Variants may also alter the inhibitory effect of uORFs or oORFs. For example, a variant in the 5'UTR of *THPO* turns an oORF into a uORF through the creation of a stop codon within the oORF sequence. The native oORF is strongly translated, leading to necessary low production of *THPO. THPO* encodes thrombopoietin, required for normal functioning of the pathway controlling the production of platelets. Turning the oORF into a uORF (with the possibility of reinitiation) increases protein levels causing hereditary thrombocythemia [94,95]. Contrastingly, changing a highly translated uORF into an oORF can increase its inhibitory effect. This can be achieved through two different mechanisms. An example of the first is seen in *NF1*, where two different variants remove the stop codon of a native uORF, transforming it into an oORF (as there are no alternative in-frame stop codons before the CDS), have been observed in patients with neurofibromatosis type I [96]. The second mechanism is

seen in *NF2,* the 5'UTR of which contains a native uORF with prior evidence of translation and a strong predicted Kozak consensus. A single base insertion changes the frame of the uORF, causing it to bypass the downstream stop codon and create an out-of-frame oORF. Translation of this oORF is predicted to lower translation of *NF2*, consistent with haploinsufficiency causing neurofibromatosis type 2 [88]. Similarly to uAUG-creating variants, interpretation of variants that alter existing uORFs can be complex. The difference in translational repression conveyed by uORFs and oORFs, and uORFs of different lengths, is difficult to predict.

1.6.3 Altering UTR splicing can cause disease

Several variants affecting splicing within the 5'UTR of *PAX6* are considered pathogenic for aniridia [97]. These variants induce exon skipping or splicing errors around exon 2 and 3 of the 5'UTR. The hypothesised mechanism of disease is through uORF dysregulation (Figure 1.5A); there is a uORF crossing these exons, and the variants reportedly change the uORF frame, which, similar to the *NF2* example above, turns the uORF into a more inhibitory oORF, leading to loss-of-function and disease [97].

Altering 5'UTR splicing can also impact the CDS sequence. For example, variants that disrupt the final acceptor site in the 5'UTR can lead to skipping or truncation of the exon containing the canonical CDS start codon. One such example is in *GJB1*, where this mechanism causes a significant amount (278 bps) of the following exon to be removed, including 262 bps of the CDS (31%; Figure 1.5B). *GJB1* only has two exons, one that is entirely 5'UTR and the other encoding a short segment of 5'UTR

plus the entire CDS and 3'UTR sequence. This variant is reported to lead to Charcot-Marie-Tooth disease [98].

Variants that create new introns into 3'UTR sequences can trigger NMD and result in loss of protein expression. An example is a 3'UTR variant in *KLHL40* that was found in a compound heterozygous state with a *KLHL40* CDS variant to cause a form of nemaline myopathy-8 (NEM8). The exonic 3'UTR variant creates a cryptic donor splice site leading to creation of a cryptic intron and resulting in decreased protein in the patient [99]. A second example is a variant in *F8* that creates a new donor splice site resulting in a 159 bp deletion in the 3'UTR, demonstrated to reduce expression and is therefore thought to cause haemophilia A [100].

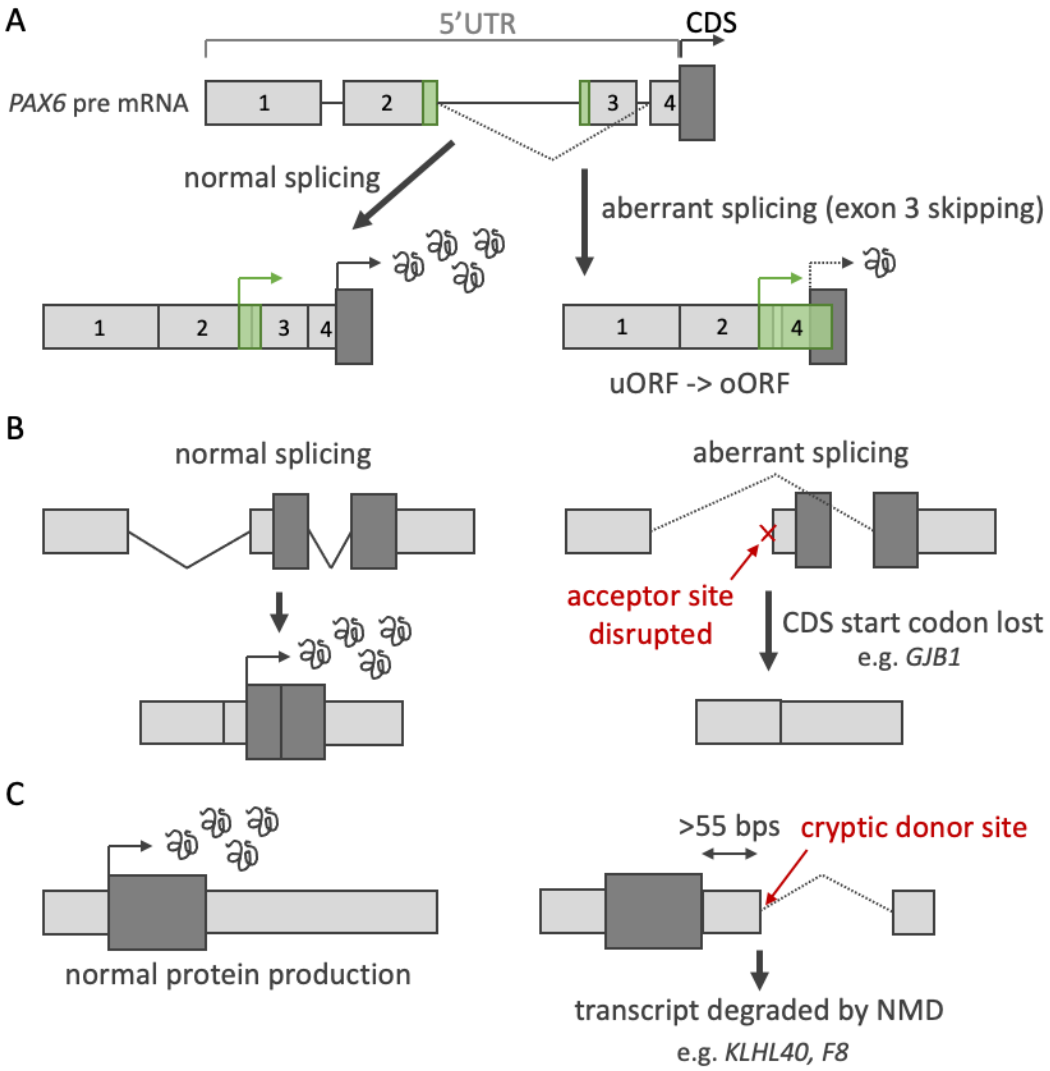


Figure 1.5: Schematic showing how UTR variants can cause disease through disrupting RNA splicing

A) *5'UTR variants in PAX6 are a common cause of aniridia. These variants are thought to cause skipping of exon three which contains the stop codon of an upstream open reading frame (uORF), converting it into an upstream overlapping ORF (uoORF) and resulting in lower translation of the PAX6 coding sequence (CDS).* ***B)*** *Variants that disrupt the acceptor splice site of the final exon in the 5'UTR remove the start codon of the CDS.* ***C)*** *3'UTR variants that create cryptic donor splice sites further than 55 bps downstream of the end of the CDS are predicted to lead to transcript degradation through nonsense mediated decay (NMD). Figure created by Nicola Whiffin and included with permission.*

1.6.4 Variants in Internal Ribosome Entry Sites (IRES) impact ribosomal recruitment

IRES motifs differ between genes so are very difficult to predict, limiting our ability to annotate and interpret variants in these elements. Potentially as a consequence, there are very few mentions of IRESs role in disease in the published literature. Exceptions include a variant in *GJB1* segregating with Charcot-Marie-Tooth disease. GJB1 encodes the protein connexin-32, which is also known as gap junction beta 1. The native IRES of *GJB1* is essential for translation of connexin-32 mRNA in nerve cells, and the variant in this case was reported to abolish the IRES function leading to no translation, as was discovered via in-depth in vivo analysis using bicistronic reporters [101]. However, the existence of this IRES was thrown into doubt by more recent experimental work [102], and an alternative proposed mechanism for this variant is that it is predicted to significantly alter the mRNA secondary structure (98). Another example is in the proto-oncogene c-myc, where in patients with multiple myeloma, a C>T substitution within a reported IRES causes higher IRES activity and increased production of c-myc protein. The underlying mechanism is not fully understood, but experimental data has shown that c-myc IRES trans-acting factors (Y-box binding protein 1 (YB-1) and polypyrimidine tract-binding protein 1 (PTB-1)) bind more strongly to the mutated version of c-myc IRES [103].

1.6.5 Repeat expansions in 5’ and 3’ UTRs can cause disease through multiple mechanisms

Repeat expansions (also known as microsatellites or simple sequence repeats) are a unique class of variation. The nucleotide sequence, location within genes, range of

repeat length and clinical outcomes vary between repeats [104]. Repeat expansions can be pathogenic when located in non-coding regions, causing so called non-coding repeat expansion disorders [105], with many examples within UTRs. The pathogenic mechanism is not known in every case, but the main possible mechanisms are as follows (as reviewed in [106]). Repeat sequences can form intramolecular structures that can influence transcription, translation and binding to various RNA binding proteins. GC-rich repeat sequences are also prone to hypermethylation that can silence the gene. More rarely, through repeat associated non-ATG (RAN) translation, the repeat RNA itself might be unconventionally translated into toxic peptides [107].

CGG repeat expansions in the 5'UTR of *FMR1* leads to hypermethylation and silencing of the gene. This results in insufficient amounts of FMR1 protein that is required for neuronal development and causes Fragile X syndrome [108,109]. A CGG repeat expansion in the *GIPC1* 5'UTR is associated with oculopharyngodistal myopathy. The number of CGG repeats is <30 in controls but >60 in affected individuals [110]. However, genes can be even more sensitive to the number of repeats; insertion of a surplus eighth CGG triplet in the *PTCH-1* 5'UTR represses protein translation drastically compared to the wild-type seven-time repeat sequence. This non-coding variant is reported to predispose to basal cell carcinoma [111]. There are also multiple instances of pathogenic repeat expansions within the 3'UTRs. For example, a large CTG repeat in the 3'UTR of *DMPK* causes myotonic dystrophy type 1 (DM1) via a toxic gain-of-function mechanism [112]. It is hypothesised that the mutant *DMPK* transcripts form aberrant structures and anomalously associate with

RNA-binding proteins [113]. Similarly, spinocerebellar ataxia type 8 (SCA8) is caused by a CTG expansion in the 3'UTR sequence of *ATXN8OS*. Here, there is evidence for both toxic RNA and toxic protein effects as RAN translation can generate a polyglutamate protein from the antisense strand [114,115].

1.6.6 5'UTR variants that overlap the promoter can disrupt transcription initiation

Promoters are sequences of DNA which initiate transcription of a gene. Relevant proteins, termed transcription factors, bind to promoter motifs to initiate transcription [1]. Promoters typically span the transcription start site (TSS) of a gene, which also marks the beginning of the 5'UTR. Hence, the portion of the promoter that is downstream of the TSS may also overlap the 5'UTR. A variant in the 5'UTR-overlapping promoter region of *BRCA1* in two unrelated families associated with epigenetic silencing (allele-specific promoter methylation) was thought to lead to grade 3 breast cancer or high-grade serous ovarian cancer [116]. However, another study questions this. The authors looked for this variant in a larger cohort of patients with *BRCA1* allele-specific promoter methylation and did not find this variant [117]. Another example is a 5'UTR variant reported to overlap the transcription start site of *MERTK* that has been hypothesised to disrupt transcription, alter secondary structure and cause inherited retinal disease [87]. This was supported by reduction in mRNA levels in a luciferase reporter assay.

1.6.7 Variants can cause disease by disrupting UTR secondary structure

The important stem-loop, the iron response element (IRE), in the L-ferritin (*FTL* gene) affects translation of mRNA in response to iron levels. L-ferritin binds and stores iron in the cell, and the quantity of L-ferritin required is tightly controlled. Cellular iron uptake must be tightly regulated as insufficient or excess levels of iron can be damaging. Hyperferritinemia/cataract syndrome (HHCS) is caused by mutations in the IRE in the 5'UTR of the L-ferritin (*FTL*) gene, preventing interaction with the IRPs and leading to dysregulated high levels of L-ferritin production [118,119].

Another example of altered 5'UTR secondary structure involves a heterozygous variant in the 5'UTR of *ADAR1* that reduces gene expression and is reported to cause dyschromatosis symmetrica hereditaria. The variant does not appear to disrupt any known 5'UTR features in that region, however, this single nucleotide change is thought to be sufficient to alter the structure of the mRNA [120]. The mechanism of how this structural change results in reduced gene expression is not currently known.

SEPN1-related myopathy corresponds to four autosomal recessive disorders. *SEPN1* produces selenoprotein which is required for normal muscle development. A unique feature of all selenoproteins is the presence of the amino acid selenocysteine, unusually encoded by the CDS stop codon, UGA. This re-coding of UGA is made possible by a highly-conserved stem-loop structure that starts 6 bp downstream of the CDS UGA codon, in the 3'UTR. This region is known as selenocysteine insertion sequence (SECIS). A SECIS RNA binding protein (SBP2) which binds SECIS, is vital

in the redefinition of the UGA to a selenocysteine codon, preventing termination at the UGA. Termination at the UGA triggers NMD and therefore leads to insufficient protein. Three variants specifically within the stem-loop structure of the SECIS have been linked to SEPN1-related myopathy by interfering with SBP2 binding, significantly reducing both mRNA and protein levels [121–123].

1.6.8 5'UTR variants can alter the efficiency of translation initiation at the CDS start codon

The exact context surrounding the CDS start codon dictates the strength of translation from that AUG and hence the amount of protein produced. Variants within the 5'UTR that disrupt this Kozak consensus sequence can cause disease. For example, a 5'UTR variant in *GATA4* that changes the Kozak sequence of the CDS start codon to a weaker consensus, leads to substantially less protein production and causes atrial septal defect [124]. Similarly, a recent paper by Nicolle *et al.* describes a variant in the 5'UTR of *RARS2* which alters the Kozak sequence of the CDS start codon, reducing protein production and causing pontocerebellar hypoplasia (PCH) [125]. The authors also assessed ClinVar for other variants that are predicted to alter the Kozak sequence of the CDS and found 20, most of which are denoted as variants of unknown significance. They conclude that variants of this class are likely an under-appreciated mechanism of disease.

1.6.9 Variants in 3'UTRs can affect polyadenylation

Multiple variants affecting polyadenylation are associated with disease risk [64,126]. For example, several variants within the PAS of *NAA10* have been linked to

syndromic X-linked microphthalmia. In vitro studies demonstrated that these variants disrupt cleavage and polyadenylation and lead to reduced mRNA levels [127]. A gain-of-function single nucleotide variant at the terminal end of the *F2* 3'UTR causes elevated prothrombin plasma levels. The wild-type polyadenylation cleavage signal is inefficient but the variant increases cleavage site recognition, increased 3' end processing, mRNA accumulation, and protein synthesis, leading to thrombophilia [128].

While not strictly a UTR variant, a single nucleotide variant changing the α-globin CDS stop codon from a UAA to CAA allows the translating ribosomes to proceed into the 3'UTR. This is associated with substantial decrease of α-globin mRNA half-life. Experimental work has elucidated that there are C-rich regions which interact with a ribonucleoprotein complex, the α-complex, containing proteins that prevent degradation of the mRNA. The α-complex is thought to protect the poly(A) tail and stabilise the mRNA. If this interaction is prevented, such as in this case, the poly(A) tail undergoes accelerated shortening; the mRNA is prematurely degraded and α-thalassemia ensues [129–131].

1.6.10 Variants disrupting microRNAs and their binding sites alter RNA silencing

There are multiple examples of pathogenic variants within miRNAs, specifically within the 'seed region' that is responsible for mediating mRNA binding [132–134]. In addition, variants in miRNA binding sites within UTRs can disrupt miRNA-mediated regulation. For example, variants within a 7 bp region of *COL4A1* abolish a binding

site for the miR-29 miRNA in the 3'UTR, leading to increased levels of COL4A1 mRNA, causing cerebral small vessel disease [135], and two variants in the miR-140 binding site in the 3'UTR of *REEP1* are predicted to suppress miRNA-mediated effects on translation, leading to less available REEP1 protein. This causes hereditary spastic paraplegia type 31 [136]. It is worth noting that variants within the miRNA's themselves can lead to disease; they can prevent binding to target mRNA and also incorrectly target other mRNAs which now match the new binding sequence on the miRNA [132–134].

1.6.11 Table of variants

Region	Mechanism	Example (Variant) GRCh38	Example (Disease)	ClinVar ID	Reference
5'UTR	Promoter disruption through methylation	*BRCA1*:NM_007294.4:c.-107A>T chr17:43125358-T-A	Breast and ovarian cancer	1685584	(116)
5'UTR	Altered Transcription Start Site	*MERTK*:NM_006343.3:c.-125G>A chr2:111898611-G-A	Inherited retinal disease	*NA*	(87)
5'UTR	Aberrant Splicing	*GJB1:*NM_000166.6:c.-16-8_-14del chrX:71223684_71223694del	Charcot-Marie-Tooth Disease	*NA*	(98)
5'UTR	Aberrant Splicing	*PAX6*:NM_001368894.2:c.-128-2del chr11:31806926-CT-C	Aniridia	430969	(97)
5'UTR	Create NTE	*IFITM5*:NM_001025295.3:c.-14C>T chr11:299504-G-A	Osteogenesis imperfecta Type V	37143	(92)
5'UTR	Create NTE	*MEF2C*:NM_002397.5:c.−8C>T chr5:88119613-G-A	MEF2C haploinsufficiency	1202675	(86)
5'UTR	Create uoORF	*MEF2C*:NM_002397.5:c.-66A>T chr5:88823854-T-A	MEF2C haploinsufficiency	1272072	(86)
5'UTR	Create uoORF	*SLC22A*:NM_003060.4:c.-149G>A chr5:132369824-G-A	Carnitine deficiency	25340	(89)
5'UTR	Create uoORF	*NF1:*NM_001042492.3(NF1):c.-280C>T chr17:31095030-C-T	Neurofibromatosis type I	2697306	(96)
5'UTR	Create uORF	*TWIST1*:NM_000474.3:c.-263C>A chr7:9117584-G-T	Saethre-Chotzen syndrome	*NA*	(91)

5'UTR	Create uORF	*NIPBL*:NM_133433.3:c.-457_-456delinsAT chr5:36876903-36876904delinsAT	Cornelia de Lange syndrome	1300231	(90)
5'UTR	uoORF removing	*EPHB1*:NM_004441.5:c.-211A>G chr3:134795421-A-G	Breast and colon cancer	*NA*	(93)
5'UTR	uORF removing	*MAP2K6:*NM_002758.4:c.-245T>C chr17:69414740-T-C	Colon cancer	*NA*	(93)
5'UTR	uoORF -> uORF	*THPO:*NM_000460.4:c.-31G>T chr3:184376290-C-A	Hereditary thrombocythemia	9510	(94)
5'UTR	uORF -> uoORF	*NF1*:NM_001042492.3:c.-272G>C chr17:31095038-G-C	Neurofibromatosis type 1	1379698	(96)
5'UTR	uORF -> uoORF	*NF2*:NM_000268.4:c.-66_-65insT chr22:29999922-A-AT	Neurofibromatosis type 2	*NA*	(88)
5'UTR	IRES Disrupting	*GJB1*:NM_000166.6(GJB1):c.-103C>T chrX:71223249-C-T	Charcot-Marie-Tooth Disease	217166	(101)
5'UTR	IRES Disrupting	*c-MYC:NM_002467.6:c.577C>T* *chr8:127738794-C-T*	Multiple myeloma	*NA*	(103)
5'UTR	Repeat Expansion	*FMR1*:NM_002024.6:c.-128GGC[200] chrX:147912049-147912050	Fragile X	183387	(108)
5'UTR	Repeat Expansion	*PTCH1:*NM_000264.5:c.-6_-4dup chr9:95508364-T-TGCC	Basal cell carcinoma	132684	(111)
5'UTR	Secondary Structure Altered	*FTL*:NM_000146.4:c.-164C > T chr19:48965344-C-T	Hyperferritinemia/cataract syndrome	96691	(119)
5'UTR	Secondary Structure Altered	*ADAR1*:NM_001111.5:c.-60A>G chr1:154608066-T-C	Dyschromatosis symmetrica hereditaria	*NA*	(120)
5'UTR	CDS start codon consensus sequence altered	*RARS2*:NM_020320.5:c.2T>C chr6:87589956-A-G	Pontocerebellar hypoplasia	1452675	(125)

5'UTR	CDS start codon consensus sequence altered	*GATA4* :NM_001308093.3:c.-6G>C chr8:11708307-G-C	Atrial septal defect	*NA*	(124)
3'UTR	Disrupted Polyadenylation and cleavage	*NAA10*:NM_003491.4:c.*43A>G chrX:153929948-T-C	Syndromic microphthalmia	617462	(127)
3'UTR	Disrupted Polyadenylation cleavage signal	*F2*:NM_000506.5:c.*97G>A chr11:46739505-G-A	Thrombophilia	13310	(128)
3'UTR	CDS stop codon readthrough affecting poly(A) tail	*HBA1*:NM_000558.5:c.427T>C chr16:177409-T-C	α-thalassemia	15624	(129)
3'UTR	miRNA binding site altered	*COL4A1*:NM_001845.6:c.*31G>T chr13:110150332-C-A	Cerebral small vessel disease	689433	(135)
3'UTR	miRNA binding site altered	*REEP1*:NM_001371279.1:c.808C>T chr2:86217086-G-A	Hereditary spastic paraplegia type 31	1166866	(136)
3'UTR	Repeat Expansion	*DMPK*:NM_004409.5:c.*224CTG[330] chr19:45770204-45770205[GCA]	Myotonic dystrophy type 1	810828	(112)
3'UTR	Repeat Expansion	*ATXN8OS*:NR_002717.2:n.1103CTG[107_127] chr13:70139384-70139386[CTG]	Spinocerebellar ataxia type 8	562101	(114)
3'UTR	Secondary Structure Altered	*SEPN1*:NM_020451.3:c.*1107T>C chr1:25816825-T-C	SEPN1-related myopathy	844320	(122)
3'UTR	Aberrant Splicing	*KLHL40*:NM_152393.4:c.*152G>T chr3:42692145-G-T	Nemaline myopathy-8	*NA*	(99)
3'UTR	Aberrant Splicing	*F8*:NM_000132.4:c.*56G>T chrX:154837541-C-A	Haemophilia A	1163207	(100)

Table 1.1: Examples of UTR variants in Mendelian disease

1.6.12 Summary

Currently UTR variants are most often annotated with respect to known regulatory elements, however, these elements may be incompletely annotated, especially if they are tissue or temporally specific. In addition, there may be further as yet unknown categories of regulatory elements through which variants may mediate their effect. This is highlighted by reportedly disease-causing UTR variants that act via unknown mechanisms. Some examples include several *ANKRD26* 5'UTR variants which have been linked to thrombocytopenia and functional studies show that these single nucleotide variants in a 19 bp conserved region reduce protein expression. The variants may be interrupting a regulatory motif that has not yet been elucidated [137,138]. A 96 bp deletion in the 3'UTR of *VMA21* segregating with X-linked myopathy with excessive autophagy (XMEA) was demonstrated to reduce mRNA quantity. Again, acting via an unknown mechanism, but by possibly destabilising the mRNA [139].

Disease associated variants in UTRs can act through multiple different mechanisms: impacting RNA-processing, RNA stability and translation. This makes interpretation of variant effects challenging and functional characterisation important. A variant's impact cannot be ruled out unless all of these mechanisms have been captured in the experiments performed [140].

I have focused this introduction on UTR variants in rare Mendelian diseases and not on more common, complex disorders, but I note that UTRs can also be enriched for variants in genome wide association studies (GWAS) [141–143].

1.7 Study aims and scope of enquiry

As discussed above, UTRs play an important role in gene expression and perturbations to them can cause disease. However, while we know that UTRs vary widely across human genes, the reasons for this variation are mostly unclear. To aid in our understanding of 5'UTRs, the overall aim of my thesis was to explore the architecture of 5'UTRs across genes by analysing large genomic datasets. Specifically:

- Chapter 3 presents a comprehensive analysis of the composition of 5'UTRs. This highlights how 5'UTRs vary widely between genes, including in length, the number of introns and the number and type of uAUGs. I also show how uAUGs appear less frequently than would be expected by chance.
- Chapter 4 assesses 5'UTR features by gene tolerance to loss-of-function (LoF), discovering that genes that are sensitive to LoF have longer and more complex 5'UTRs containing more uORFs, with greater propensity to form secondary structures and higher levels of cross-species conservation than genes that are tolerant to LoF.
- Chapter 5 extends the above analyses to sets of genes which cause developmental disorders and cancer, and that are known to be dosage

sensitive. This analysis uncovers a stark difference between genes that act via dominant and recessive mechanisms.

- Chapter 6 explores a novel type of genetic variation, start-stop creating and removing variants, and their role in disease. I examined variants in ClinVar and the Genomics England 100,000 Genomes Project (GEL) to find and assess variants that create or remove start-stops.

Chapter 2

General Datasets

In this chapter I provide a detailed overview of the datasets used throughout this thesis. Additionally, I present a summary table (Table 2.1) that cross-references the datasets with the chapters in which they are used.

2.1 MANE defined UTRs

The MANE (Matched Annotation from the NCBI and EMBL-EBI) project [144] defined a genome-wide set of representative transcripts for human protein-coding genes, and are chosen based on biologically relevant criteria such as transcript expression levels and Cap Analysis of Gene Expression (CAGE) which tags the 5’ ends of mRNA transcripts. As the MANE Select gene set has defined a single biologically relevant transcript per protein coding gene, it provides an optimal dataset for analysing UTRs. It reduces the complexity associated with comparing multiple UTR features in genes that have multiple transcripts.

2.2 Defining a set of uORFs with experimental evidence

Ribosome sequencing (Ribo-seq) is a ribosome profiling technique which chemically immobilises translating ribosomes in-situ on the mRNA. This provides a “snapshot” of all active ribosomes in a cell at a specific time. The mRNA fragments within the

bound ribosome, known as ribosome protected fragments (RPF), are then sequenced. Using computational tools, the RPF are aligned to genomic sequences to identify their position in the genome and filtered to identify regions that are likely actively being translated [74]. I used a Ribo-seq uORF dataset in this thesis from Chothani *et al*. [31] as this is a high confidence set of 5,052 uORFs from multiple human cell types and tissues (human primary atrial fibroblasts, primary human hepatocytes, vascular smooth muscle cells, pluripotent human embryonic stem cells, human umbilical vein endothelial cells, human coronary artery endothelial cells, human aortic endothelial cells, brain tissue and adipose tissue).

2.3 uAUG Translational Efficiencies

The surrounding consensus sequence of the start codon influences its likelihood to be translated by the small ribosome [41,42]. Translational efficiencies (TE) of uAUGs were determined using work by Noderer *et al.*, 2014 [41] by matching to the surrounding sequence context. They used fluorescence-activated cell sorting and high-throughput DNA sequencing (FACS-seq) to determine efficiency of start codon recognition for all possible translation initiation sites using AUG start codons, across a variety of cell lines. They created a quantitative metric, scoring each sequence context according to its efficiency at start-codon recognition; with lower TE scores indicating lower likelihood of translation being initiated from that motif and higher TE scores indicating higher likelihood of translation being initiated from that motif. This is more recent and on a larger experimental scale than the work done by Marilyn Kozak,

who originally found the optimum sequence motif surrounding start codons to allow the ribosome to recognise the translation initiation site [42].

2.4 PhyloP as a measure of conservation

Phylogenetic P-values (PhyloP) [145] is a tool and score to evaluate the evolutionary conservation of each nucleotide position based on the comparison of homologous sequences between different species. Positive PhyloP scores indicate positions that are more conserved than expected and negative scores indicate positions that are evolving faster than expected or positive selection.

2.5 CADD as a measure of deleteriousness

Combined Annotation Dependent Depletion (CADD) scores measure the potential deleteriousness of genetic variants. CADD is an integrative annotation built from more than 60 genomic features, including conservation, epigenetics and genetic context [146].

2.6 Population reference database (gnomAD)

The Genome Aggregation Database (gnomAD), aggregated and harmonised exome and genome sequencing data from a wide variety of large-scale sequencing projects. It is a large genetic variation resource and the data was primarily obtained from case-control studies of common adult-onset diseases with good representation of a diverse range of populations. gnomAD is a valuable and comprehensive population genetics

resource for understanding frequency and distribution of genetic variants and is used to identify and interpret genetic variants [147,148].

Due to the large number of exomes, the gnomAD consortium were able to devise a method to categorise gene tolerance to loss-of-function (LoF). By using this large-scale, population cohort, they developed methods to assess the frequency of deleterious LoF variants within the population compared to what would be expected. Protein-coding genes were scored along a spectrum that represents tolerance to inactivation, termed the “loss-of-function observed/expected upper bound fraction” (LOEUF) [147]. Low LOEUF scores are indicative of genes intolerant to LoF, while high LOEUF scores indicate genes which are tolerant to LoF.

2.7 Accounting of differences in gene expression

The Genotype-Tissue Expression project (GTEx) [149] collects and analyses gene expression levels, quantified from RNA sequencing, from a wide range of human tissue samples. GTEx is valuable for assessing expression patterns of genes across tissues, providing insight into where genes are expressed and their expression levels.

2.8 Assessing transcription start site (TSS) diversity

Cap Analysis of Gene Expression (CAGE) is an experimental technique used to identify and quantify 5’ ends of mRNA transcripts, allowing us to analyse TSS usage. Small fragments from the very beginning of mRNAs are reverse transcribed and sequenced, and then aligned to the reference genome to determine where that

mRNA transcript was transcribed from. The FANTOM project (Functional ANnoTation Of the Mammalian genome) carried out CAGE on many human cell and tissue types [150]. This dataset contains how many TSS sites per gene were determined using this method (CAGE “peaks”), giving insight into transcript diversity.

2.9 Cell Essential Gene Set

A cell essential gene set was derived from CRISPR/SpCas9 studies by Hart *et al.* [151], and was used in the gnomAD flagship paper [147]. CRISPR is a gene editing technique [152] that can be used to create gene knockouts. In this study, CRISPR was used to target human protein-coding genes in human cell lines to identify whether the knockouts caused cellular dysfunction and defined a gene set that are essential to cell function.

2.10 Identifying disease-gene sets

DDG2P is a curated list of genes reported to be associated with developmental disorders, compiled by clinicians as part of the Deciphering Developmental Disorders (DDD) study [153].

The Catalogue Of Somatic Mutations In Cancer (COSMIC) Cancer Gene Census [154] is an expert curated resource detailing genes driving human cancer that is used as a standard in cancer genetics. Each gene is annotated with its mechanism of action, such as whether it functions as an oncogene or a tumour suppressor gene.

Dosage sensitive genes (haploinsufficient and triplosensitive) gene sets were taken from the work by Collins *et al*. [155]**.** Rare copy-number variants (rCNVs) include deletions and duplications that occur infrequently and confer substantial risk for disease. This study quantified the properties of haploinsufficiency (deletion intolerance) and triplosensitivity (duplication intolerance) by analysing rCNVs from nearly one million individuals to construct a genome-wide catalogue of dosage sensitivity across 54 disorders. Using this, they also designed a machine learning model to predict probabilities of dosage sensitivity, which identified 2,987 haploinsufficient and 1,559 triplosensitive genes.

2.11 ClinVar: Clinical genetics variant database

ClinVar is a large public repository of curated clinical variants linked to phenotype [156]. It is a platform where a variety of genetic professionals, such as clinicians and researchers, can add genetic variants along with associations with specific diseases and their clinical interpretations based on literature, functional studies and clinical observations. It is a valuable resource in clinical genetics for variant interpretation.

2.12 Genomics England 100,000 Genomes Project (GEL)

The Genomics England 100,000 Genomes Project was launched in 2012 by the UK government, with the objective to sequence 100,000 whole genomes from National Health Service (NHS) patients with rare diseases or cancer, and often their close family members [157,158]. The sequenced genomes and clinical data can be accessed by healthcare professionals and researchers within a secure environment

controlled by Genomics England, to allow for analysis of the data and translation of genomic insights into clinical benefits for the patients.

Data	Chapter
MANE	3,4,5,6
Ribo-seq uORFs	3,4
uAUG consensus sequence	3,4,6
PhyloP	3,4,5,6
CADD	3,4
gnomAD	4,6
GTEx	4
Codon optimality	4
CAGE	4
Cell essential genes	4
Disease-gene sets	5
ClinVar	6
GEL	6

Table 2.1: *List of datasets and the chapter they are used in.*

Chapter 3

Quantifying 5'UTRs variation across genes

3.1 Introduction

As discussed in Chapter 1, 5'UTRs play an important role in gene expression and alterations to them can cause disease. 5'UTRs vary widely across human genes but the reasons for this variation are mostly unclear. To aid in our understanding of 5'UTRs, it is essential to define their characteristics. In this chapter, I explore the architecture and composition of 5'UTRs and their variance between genes. I analysed 18,764 5'UTRs annotated by the MANE project [144]. I set out to determine the average length of 5'UTRs, intron occurrence and uAUG characteristics. I also explored whether there were the following potential correlations: between 5'UTR length and uAUG number, between 5'UTR length and 3'UTR length, and between 5'UTR length and minimum free energy (MFE) as an indication of secondary structures.

3.2 Results

3.2.1 5’UTR length and transcript diversity varies between genes

The length of 5’UTRs varies widely between genes (Figure 3.1), ranging from 1-3,561 base pairs (bp), with a mean of 202 bp. The majority (88%) of 5’UTRs are 400 bp long or less. 5’UTRs are relatively short compared to the CDS and 3’UTRs; the average CDS length is 1,745 bp and 3’UTRs average length is 1,791 bp.

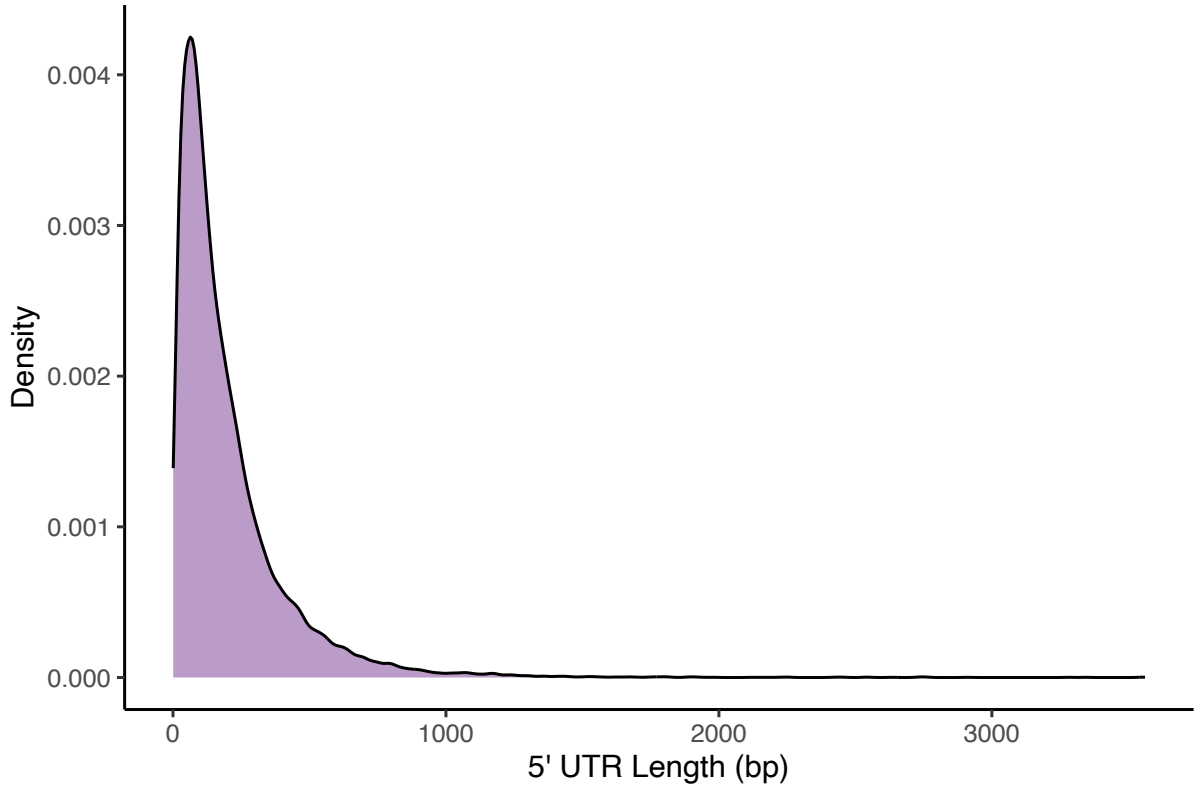


Figure 3.1: *Range of 5’UTR length across 18,764 human genes. 5’UTRs range from 1-3,651 bp long. The majority (88%) of 5’UTRs are 400 bp long or less.*

Although the focus of this thesis is on 5'UTRs, I wanted to test whether 5'UTR and 3'UTR lengths are correlated. I plotted each gene's 5'UTR and 3'UTR lengths (Figure 3.2); no correlation is observed between 5'UTR and 3'UTR length (R^2=0.056). However, there is significant variability, with some genes exhibiting markedly long 5'UTRs and a very short 3'UTR and vice versa.

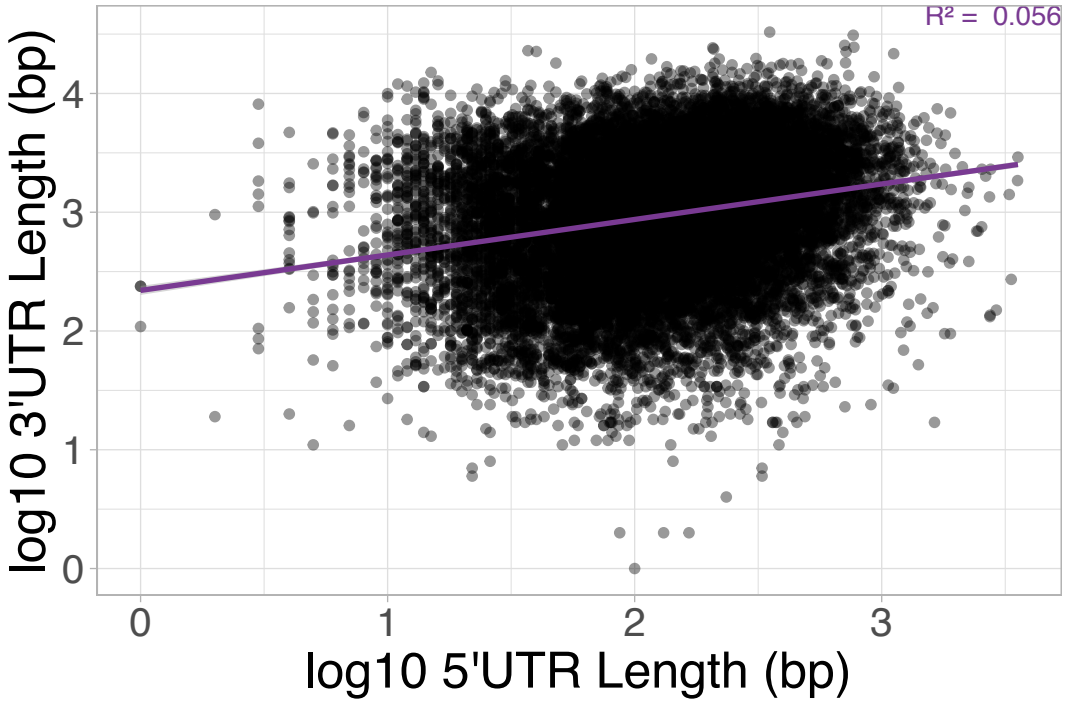


Figure 3.2: *5' and 3' UTR length are not correlated; there is no correlation that genes with longer 5'UTRs have longer 3'UTRs too (R^2=0.056). There is considerable variability with many genes having long 5'UTRs and short 3'UTRs or vice versa.*

To observe 5'UTRs potential for alternative splicing, I explored 5'UTR introns. 5'UTRs can have up to 11 introns, but the majority of 5'UTRs (62.3%) do not have introns (Figure 3.3).

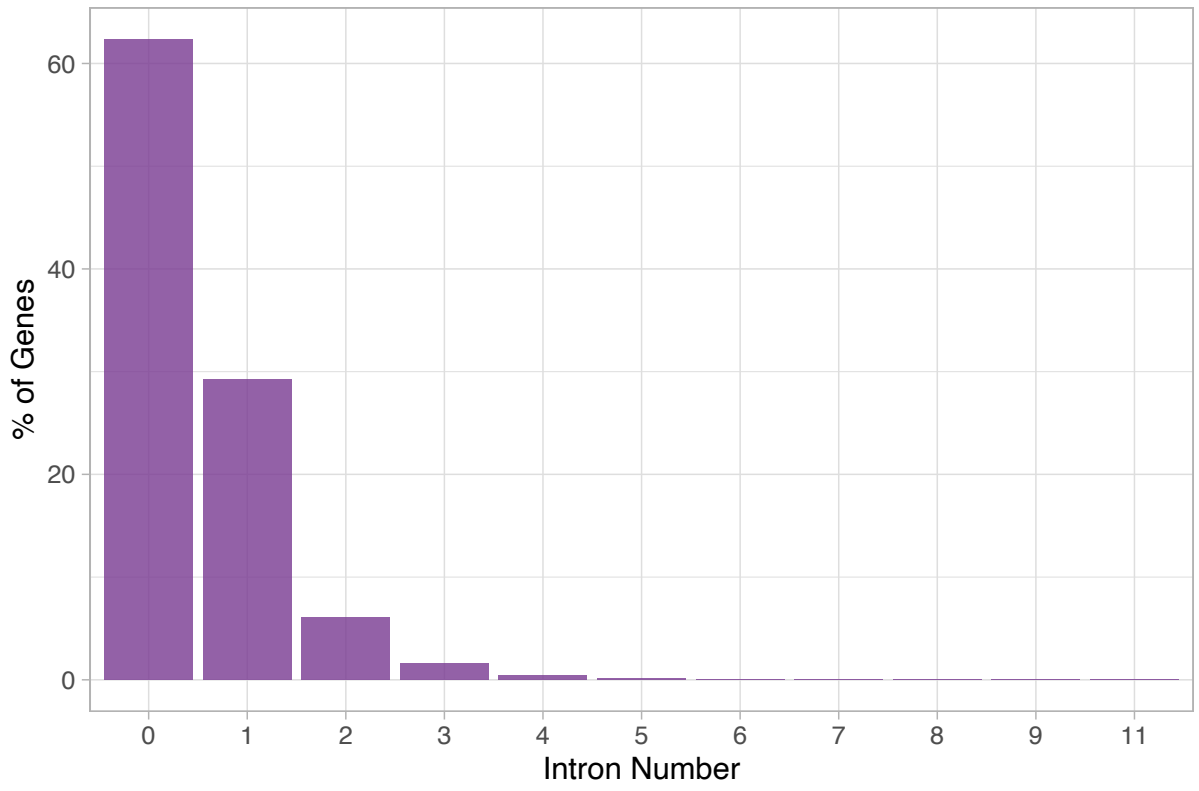


Figure 3.3: *There can be up to 11 introns within a 5'UTR. Majority of 5'UTRs (62.3%) do not have 5'UTR introns.*

3.2.2 uAUGs are a common feature of 5'UTRs

The number of uAUGs ranges from 0-64 per gene, with 42.5% of 5'UTRs having at least one uAUG (Table 3.1). I further classified these uAUGs by effect. A uAUG with a stop codon in-frame within the 5'UTR was considered a uORF. oORFs are uAUGs with no stop codon in-frame within the 5'UTR. oORFs are considered in-frame or out-of-frame depending on if they were in-frame with the CDS. uAUGs that were immediately followed by a stop codon, with no codons in between, were classed as a start-stop. I found that 34.4%, 15.0%, and 5.0% of 5'UTRs contain at least one uORF, oORF, and start-stop element, respectively (Table 3.1). Longer 5'UTRs tend to have

more uAUGs (R^2=0.503) (Figure 3.4), however, there are several exceptions with some longer 5'UTRs having no uAUGs.

Length	Range	1–3561
	Mean	202
	Median	136
Introns	Range	0–11
	Genes with 1 or more introns (%)	37.7
uAUG	Range (per gene)	0–64
	Genes with at least 1 uAUG (%)	42.5
uORF	Range of uORF number per gene	0–61
	Genes with at least 1 uORF (%)	34.4
	Mean uORF length (bp)	57
	Mean uORF to CDS distance (bp)	258
oORF	Range of oORF number per gene	0–8
	Genes with at least 1 oORF (%)	15.0
	oORF in-frame (%)	11.4
	oORF out-of-frame (%)	88.6
Start-Stops	Range of start-stop number per gene	0-8
	Genes with at least 1 start-stop (%)	5.0

Table 3.1: An overview of 5'UTR structure and features
Descriptive statistics of annotated 5'UTRs across 18,764 MANE Select transcripts. The length, number of introns, and number and types of uAUGs range widely across genes. uORFs: upstream open reading frame; oORF: overlapping open reading frame; uAUG: upstream AUG (start codon).

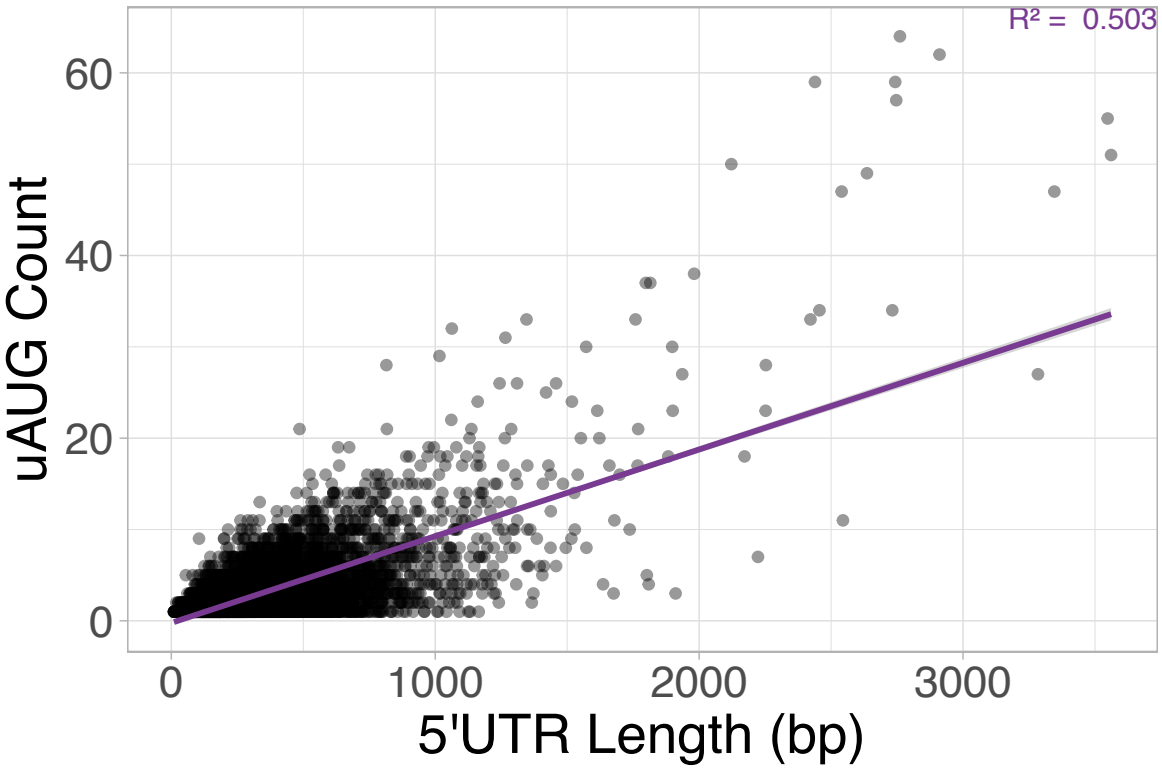


Figure 3.4: *uAUG number per 5'UTR compared to 5'UTR length. Longer 5'UTRs tend to have more uAUGs within them (R^2=0.503).*

Next, I investigated where in the 5'UTR the three kinds of uAUGs fall. I took the uAUG codon position (the first base of the codon) and divided it by the UTR exonic length to get a metric of where in the 5'UTR it was positioned. Figure 3.5 illustrates the distribution of the three types of uAUGs positionally within the 5'UTR - uORFs further away from CDS start, oORFs closer to CDS start and start-stops distributed throughout. This is to be expected as a uAUG closer to the CDS gives less chance of a stop codon to occur between the uAUG and CDS, rendering it an oORF.

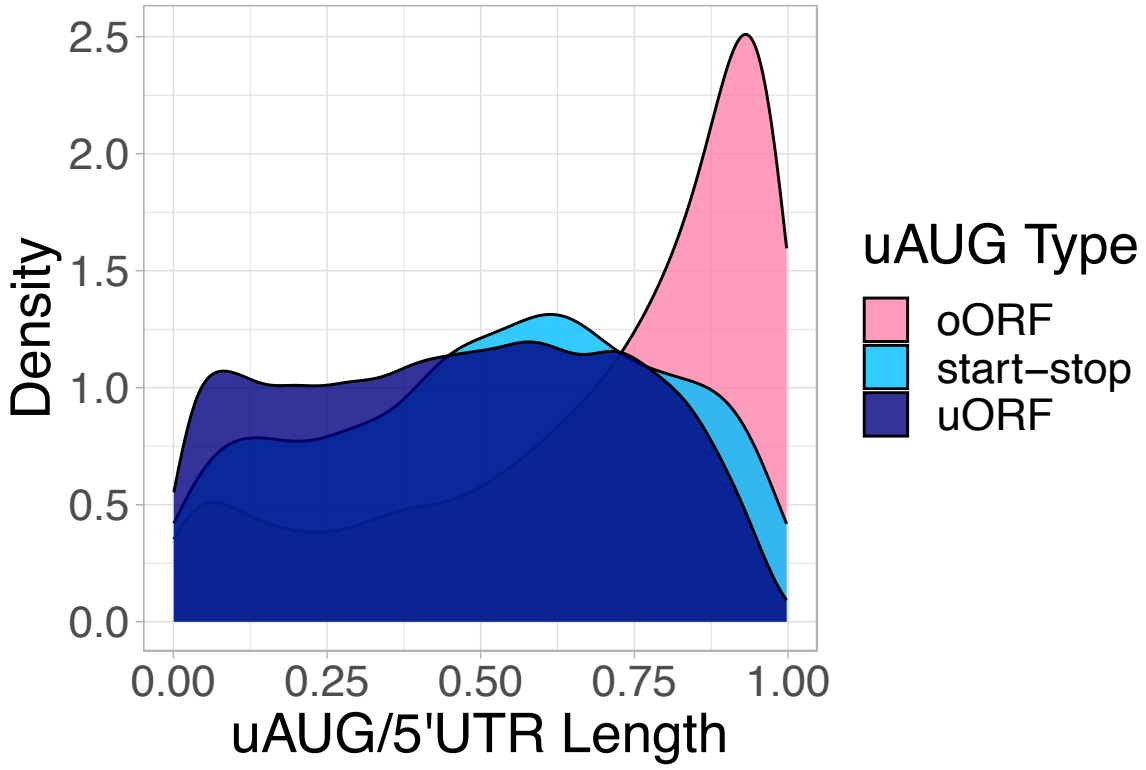


Figure 3.5: *Position of uAUG type within the 5'UTR: oORFs are closer to the CDS and uORFs are further away.*

3.2.3 Characterising translated uORFs from ribosome profiling data

In addition to annotating all occurrences of canonical AUG triplets with an in-frame stop codon in each 5'UTR as 'predicted uORFs', I used a set of 5,052 uORFs detected through ribosome sequencing of six cell types and five tissues ('Ribo-Seq uORFs') [31]. I wanted to see how much overlap between the Ribo-seq uORFs and the predicted uORFs; 1,430 (28.3%) of the Ribo-Seq uORFs have an exact match to the predicted uORFs (Figure 3.6). In addition, the Ribo-Seq uORF set contains 2,288 additional uORFs that start at non-canonical (non-AUG) start-codons (45.3% of the Ribo-Seq uORFs). Overall, 19.3% of MANE transcript 5'UTRs contain one or more Ribo-Seq uORFs (range 1-11).

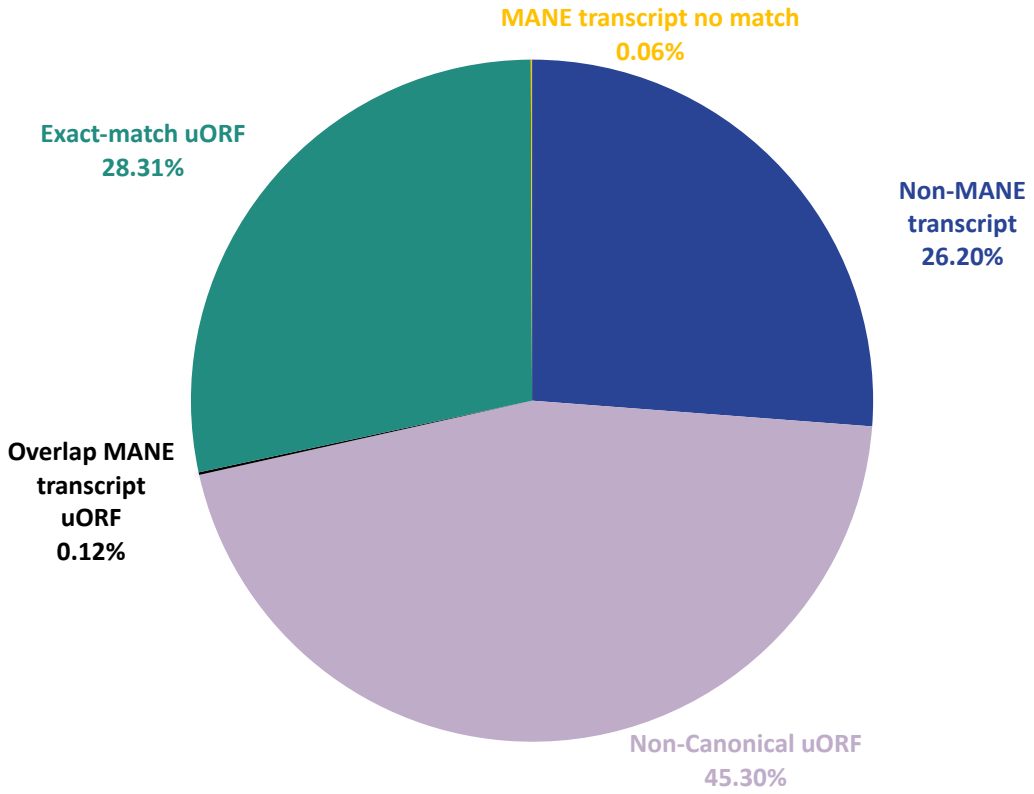


Figure 3.6: *Overlap between Ribo-seq uORFs (n=5,052) and predicted MANE uORFs (n=18,064). 28.3% of Ribo-seq uORFs matched a predicted uORF exactly. A further 0.12% Ribo-seq uORFs overlapped with a predicted uORF. 26.2% of Ribo-seq uORFs did not map to a MANE transcript 5'UTR. 45.3% of Ribo-seq uORFs start with a non-canonical start codon (non-AUG) and hence would not be predicted to match with the predicted uORF set, which only considered uORFs with uAUG start.*

uORFs tend to be short; the predicted uORFs average 57 bp long and Ribo-seq uORFs average 49 bp (Figure 3.7).

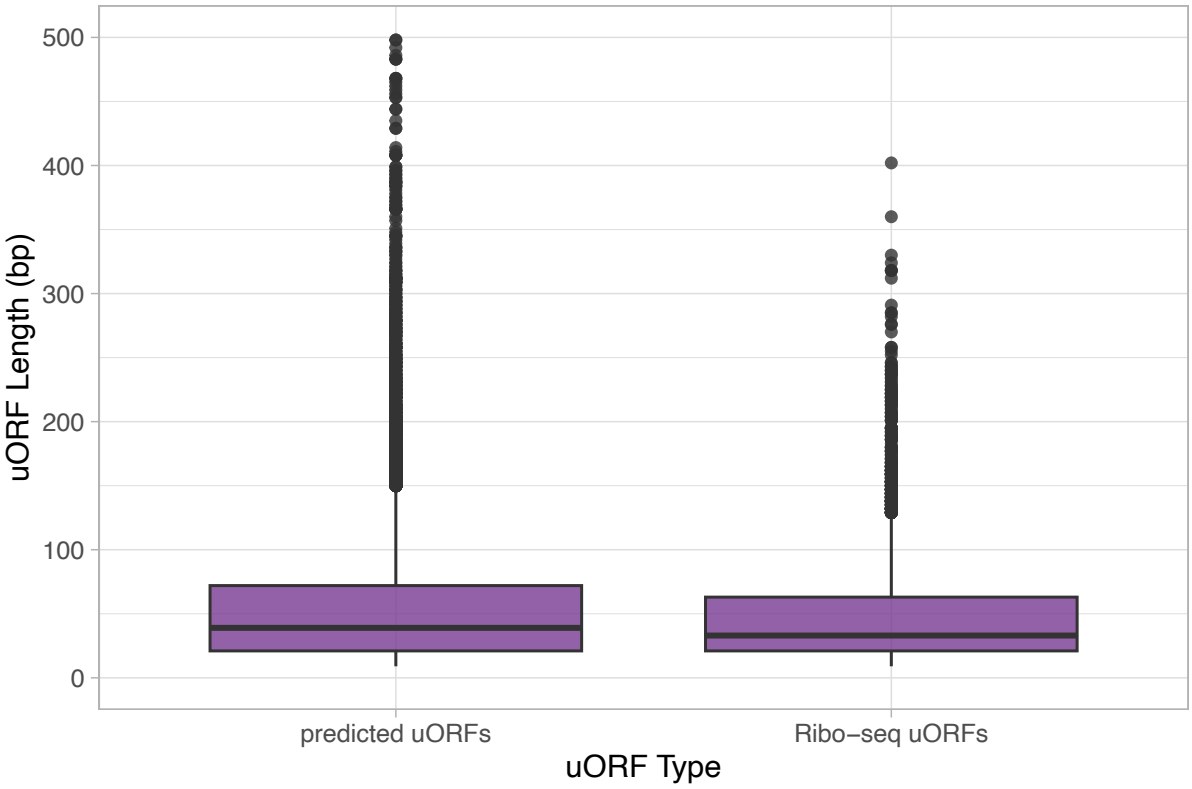


Figure 3.7: *uORF length in predicted uORFs and Ribo-seq uORFs. uORFs tend to be short; the predicted set average 57 bp long and the Ribo-seq uORF set average 49 bp. The y-axis is truncated at 500 bp (21 uORFs were >500 bp).*

The consensus sequence around the start codon influences its likelihood to be translated (translational efficiency, (TE)) [41,42]. Therefore, I analysed the consensus sequence motifs around the predicted uAUGs to determine if they generally exhibit consistent or varied consensus sequences and whether they tend to have strong or weak consensus sequences. This analysis aimed to assess the potential of uAUGs to initiate translation. I used an experimentally validated TE score set [41] and the Kozak consensus sequence [42]. The TE score set is a quantitative metric to determine the efficiency of start codon recognition for all possible translation initiation sites, derived from experimental work carried out across a variety of cell lines. Lower

TE scores indicate lower likelihood of translation being initiated from that motif [41]. Using this, I found that for predicted uAUGs, there is a wide range of translational efficiencies (Figure 3.8). However, oORFs had significantly lower TE scores than uORFs (Wilcoxon, $P<1x10^{-15}$). This is in line with what would be expected as oORFs are more repressive of the main CDS than uORFs, so are likely to be less tolerated.

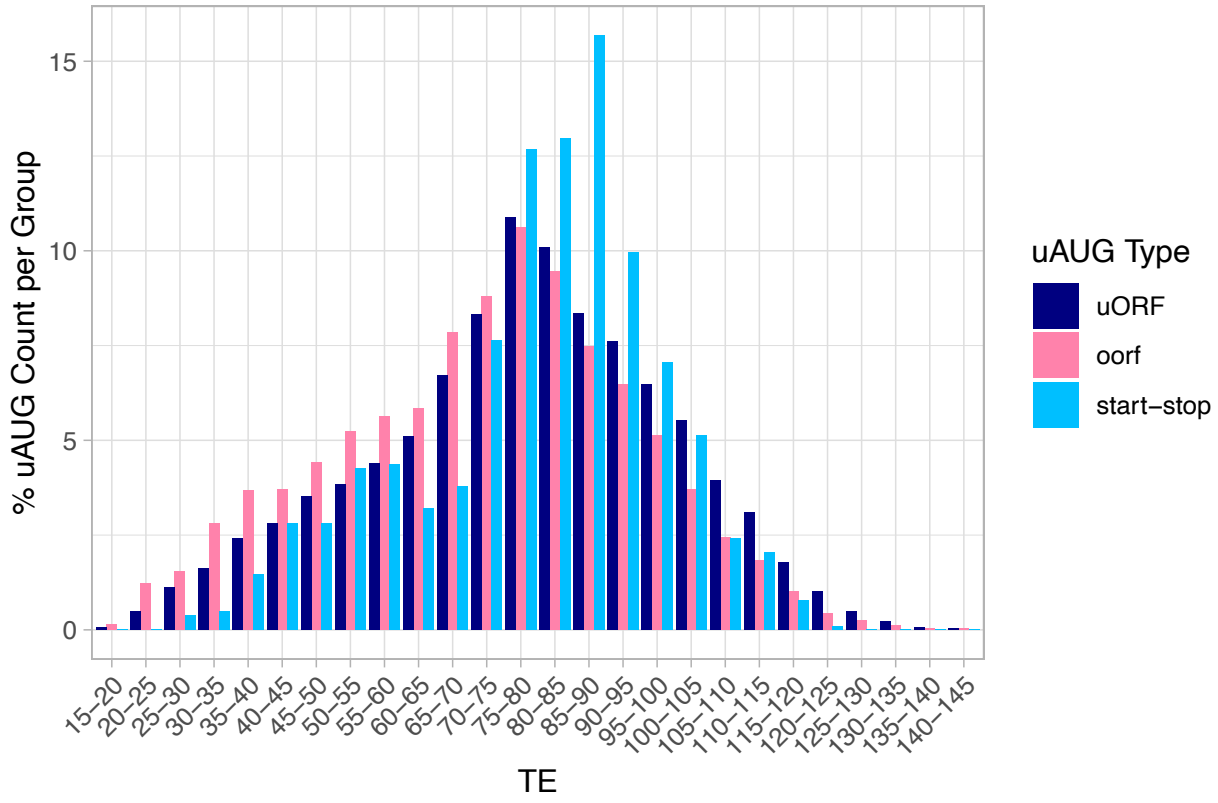


Figure 3.8: Translational efficiencies for predicted uAUGs
Using an experimentally validated set of uAUG translational efficiency scores (TE scores), plotted TE scores for the three groups of uAUGS: uORF, oORF and start-stops. Low TE scores are less efficiently translated, with high scores being more efficiently translated. oORFs had significantly lower TE scores than uORFs (Wilcoxon, $P<1x10^{-15}$).

I hypothesised that uAUGs with stronger Kozak consensus sequences are more likely to be translated efficiently, and thus have higher TE scores. I used the Kozak

consensus motifs, which I categorised as “weak”, “moderate” and “strong” as an indication as to their translational efficiency. To explore how the TE scores compare to the Kozak consensus sequence [42], I plotted TE scores grouped by Kozak consensus sequences for both the predicted uORF set (Figure 3.9A) and Ribo-seq uORFs that had canonical AUG start codons (Figure 3.9B). There is a clear delineation with “weak” Kozak sequence grouping towards the lower TE scores and “strong” Kozak tending to group at higher TE scores, albeit with variation.

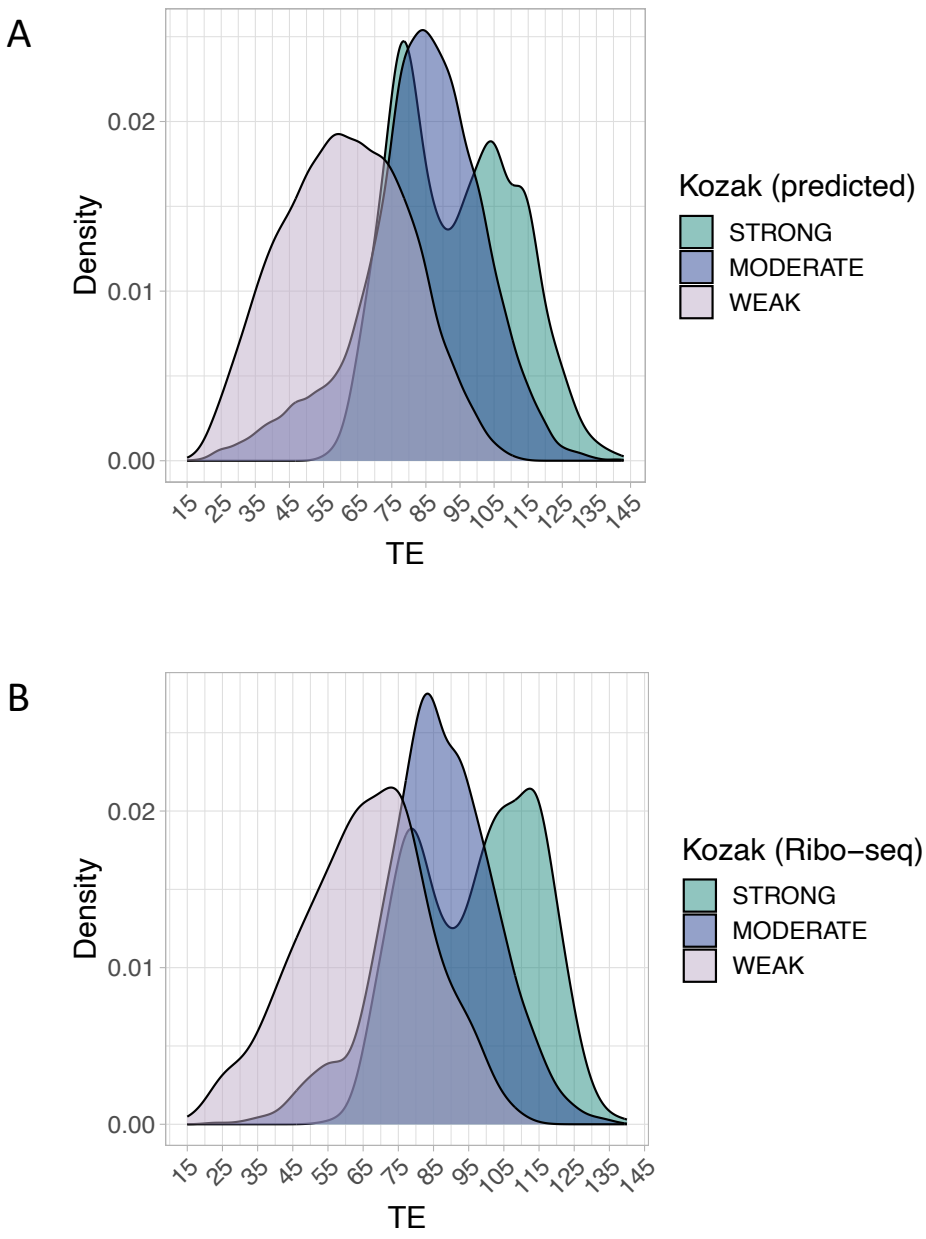


Figure 3.9: Comparing TE scores to Kozak consensus sequence strengths
A) *Predicted uORF scores. Mean TE score for each Kozak group: weak=61.4, moderate=83.8, strong=93.8.* ***B)*** *Ribo-seq uORF that begin with AUG scores. Mean TE score for each Kozak group: weak=65.4, moderate=86.1, strong=97.3.*

3.2.4 uAUGs are strongly depleted across 5'UTRs

To test whether uAUG codons ("ATG") occur more or less frequently than we would expect by chance in 5'UTRs, I shuffled all MANE 5'UTR sequences 1000 times while maintaining di-nucleotide composition. I calculated an observed/expected (o/e) score for each of the 64 codons. uAUGs were significantly more depleted than would be expected by chance (Figure 3.10) and appear as the most depleted codon across 5'UTRs.

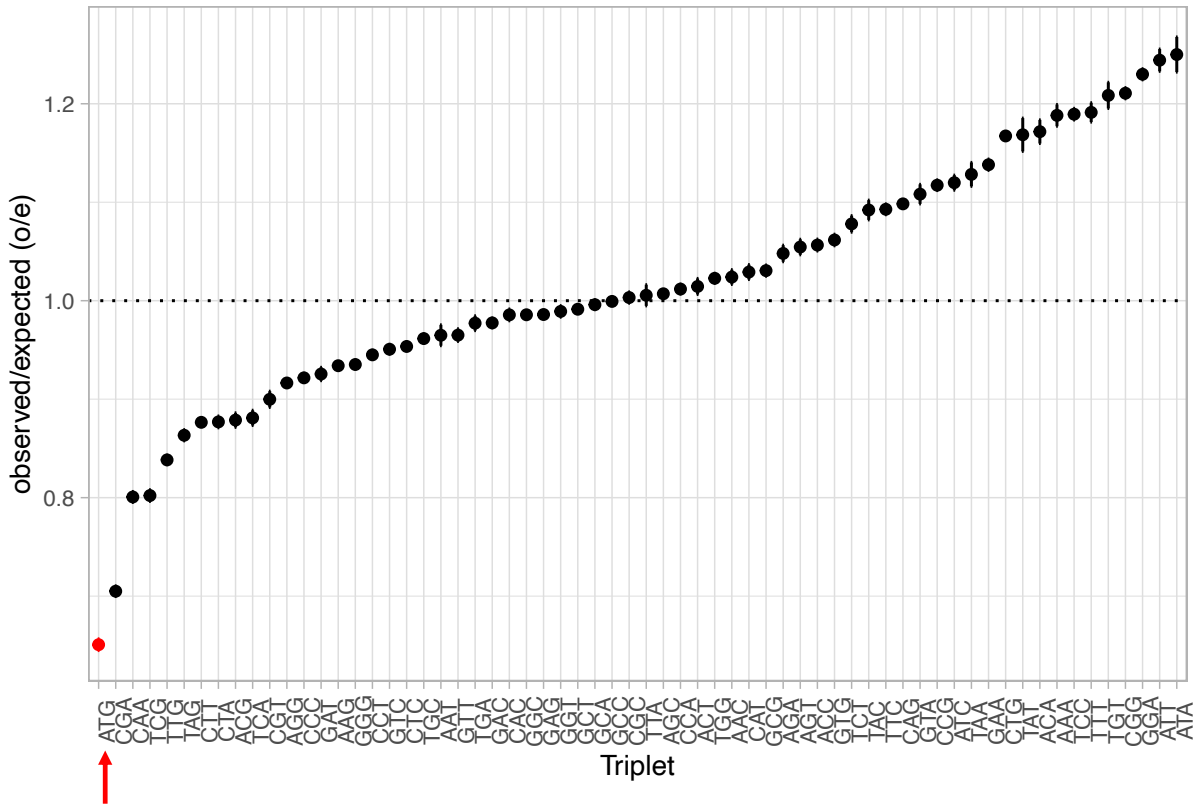


Figure 3.10: *MANE 5'UTR sequences were shuffled 1000 times to generate an observed/expected (o/e) ratio for all codons in 5'UTRs. o/e values below 1 indicate codon appearing at a lower frequency than expected by chance; exceeding 1 suggests higher than expected frequency. The dotted line denotes o/e=1. ATG codons were significantly more depleted than would be expected by chance. Values plotted with 95% confidence intervals (from the empirical distribution).*

3.2.5 Comparative Conservation and Variant Deleteriousness Analysis in 5'UTRs

Using PhyloP scores [145] to measure conservation between species, I calculated 5'UTR, uORF and start-stop conservation scores (Figure 3.11A). I also used Combined Annotation-Dependant Depletion (CADD) scores which measure variant deleteriousness for all possible single nucleotide substitutions at each position, integrating more than 60 genomic features [159]. I looked at CADD scores for

5’UTRs, uORFs and start-stops (Figure 3.11B). Compared to the mean 5’UTR exonic PhyloP scores, uORF start and stop codons and start-stops were conserved to a significantly greater degree (Wilcoxon, All $P<1x10^{-15}$). However, there was no significant difference between 5’UTR exonic CADD scores and uORF start and stop codons and start-stops CADD scores (Wilcoxon, All $P>5.9x10^{-4}$).

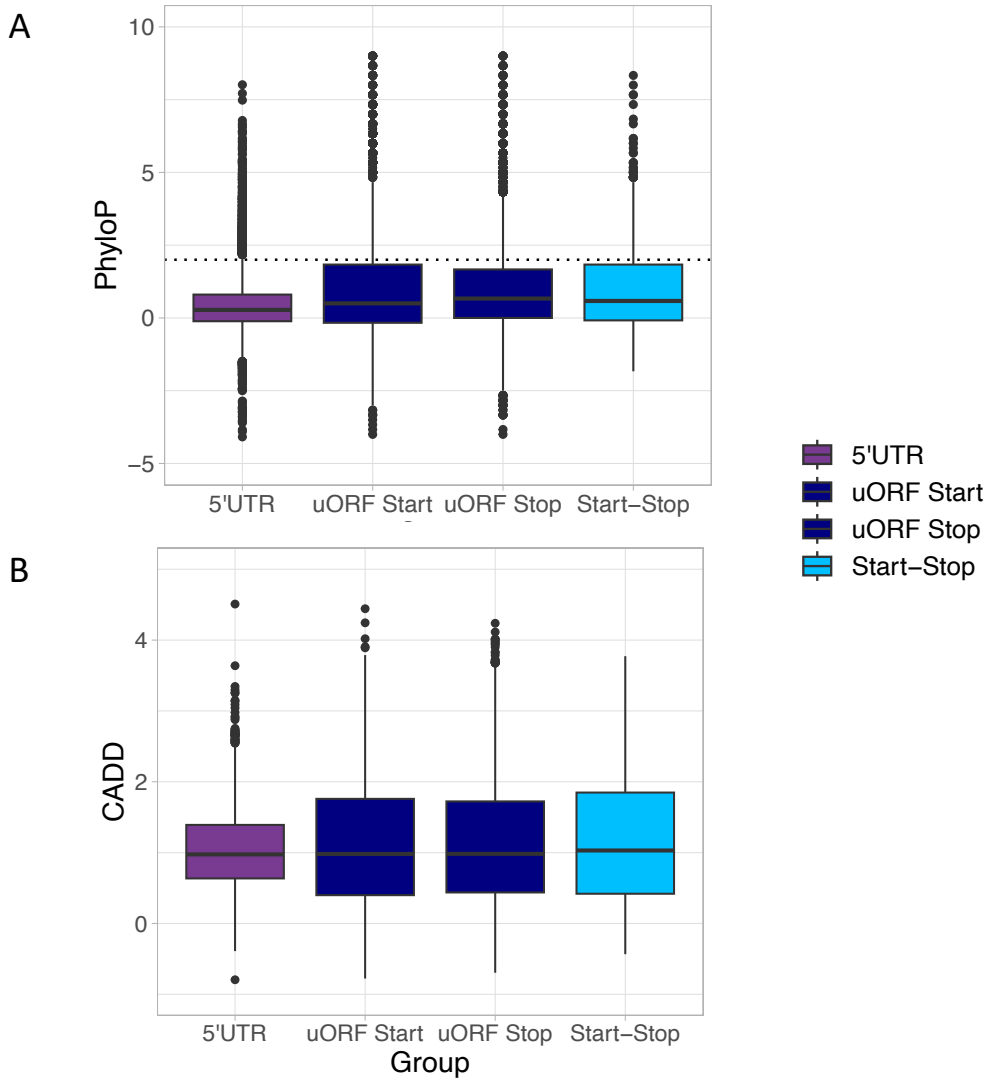


Figure 3.11: PhyloP and CADD

A) *PhyloP scores for 5'UTR, uORF start, uORF stop and start-stops. Compared to the mean 5'UTR exonic PhyloP scores, uORF start and stop codons and start-stops were conserved to a significantly greater degree (Wilcoxon, All $P<1x10^{-15}$). Dotted line is where Phylop=2.* ***B)*** *CADD scores for 5'UTR, uORF start, uORF stop and start-stops. There was no significant difference between 5'UTR exonic CADD scores and uORF start and stop codons and start-stops CADD scores (Wilcoxon, All $P>5.9x10^{-4}$).*

3.2.6 Longer 5'UTRs have greater propensity to form secondary structures

mRNA is capable of folding and forming secondary structures through complementary base-pairing within itself. The propensity of a sequence to form RNA secondary structures can be predicted from high GC content and low minimum free energy (MFE) [14,160]. I used RNAfold [161] to compute the MFE prediction per 5'UTR. Longer 5'UTRs had lower MFE (R^2=0.891) as expected; there is more chance for folding to occur (Figure 3.12), but there was no significant correlation with GC content (R^2=0.006) (Figure 3.13).

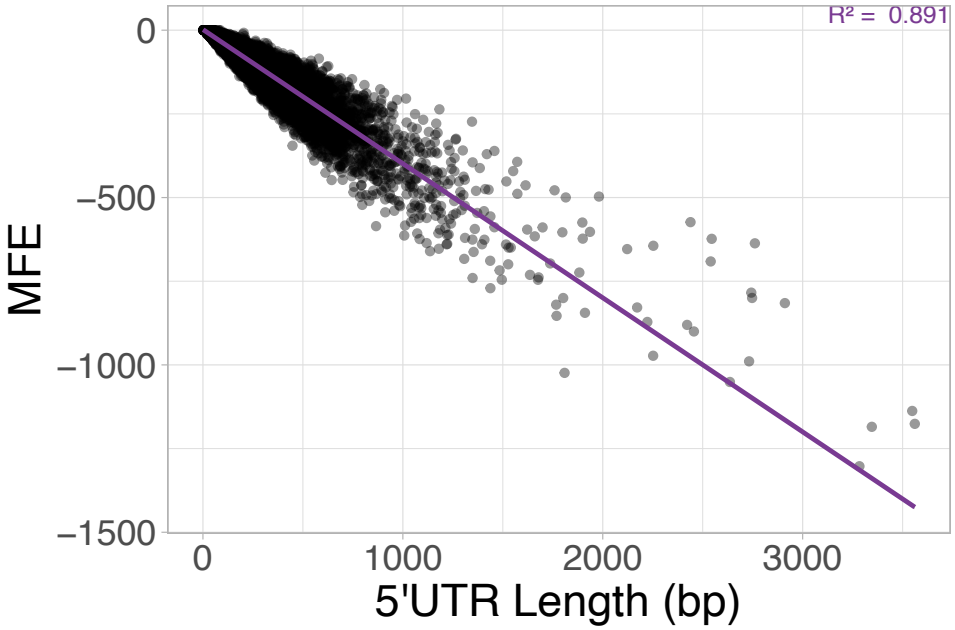


Figure 3.12: *Minimum Free Energy (MFE) scores for 5'UTRs. Longer 5'UTRs are correlated with lower MFE scores (R^2=0.891); longer 5'UTRs have a higher propensity to form secondary structures.*

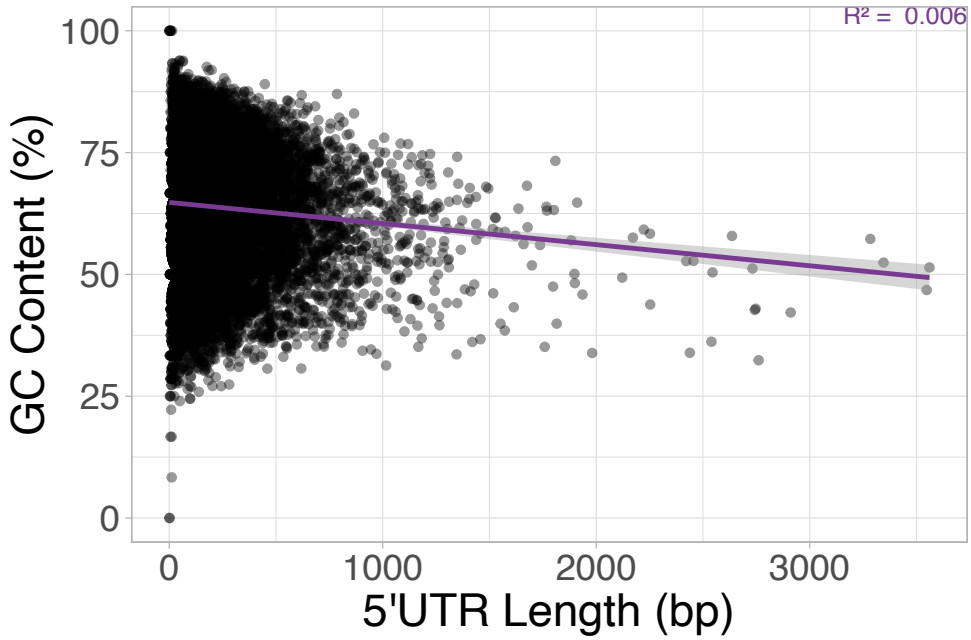


Figure 3.13: *Longer 5'UTRs are not correlated with higher GC content (R^2=0.006).*

3.3 Discussion

In this chapter I show that the length, frequency and types of uAUGs present vary widely across the 5'UTRs of human genes. Contrary to my initial expectation, the lengths of 5' and 3' UTRs do not show strong correlation. I had wondered whether genes requiring tight regulation would have longer and more complex 5' and 3'UTRs. However, this observation (that the 5' and 3' UTR lengths are not correlated) could underscore the distinctly different regulatory capabilities of 5' and 3' UTRs. This is likely to be dependent on the specific gene they are regulating with each UTR modulating different modes of control. For instance, some genes may need more control over protein quantity which is facilitated by the 5'UTR rather than control over protein localisation which is more characteristic of the 3'UTR.

uAUGs occur less frequently than we would expect by chance, possibly as they can be disruptive to protein production. The work here reflects earlier work by others [35,37]; and uAUGs of uORFs are the most conserved sites in 5'UTRs [37].

I used two different uORF sets throughout this work, a predicted set derived from every AUG within each 5'UTR, and an experimental set from Ribo-Seq [31]. The predicted uORF set likely contains many uORFs that are not translated - some 5'UTRs have over 40 uAUGs, with many of them overlapping frames. Not all the uAUGs in these scenarios would cause a ribosome to initiate and translate, but indicates high potential for at least some of the uAUGs to be translated. By comparison, the highest number of translated uORFs from the Ribo-seq data in one gene was 11. Conversely, due to necessary stringent filtering, tools used to detect uORFs and tissue and temporal specificity of uORFs, there are likely many uORFs that are translated, but that are not captured in the Ribo-seq data. Other work has also shown preferential uORF translation under stress conditions [27,162]. The predicted uORF set is also only based on canonical start sites, whereas 45.3% of the Ribo-seq uORFs use non-canonical start sites. Therefore, there are likely many more potentially translated uORFs which are excluded from the predicted uORF set. The predicted set has many more uORFs (18,064) which is useful for statistical power. The Ribo-seq uORF set is much smaller (5,052) but has the advantage of evidencing translation, and is transcript agnostic.

Due to the respective limitations of these uORF sets, in the following chapter I use both the predicted uORFs and Ribo-seq uORF sets, as they each have their own value.

3.4 Methods

3.4.1 5'UTR set, UTR length and introns

MANE Select transcripts from v1.0 of the MANE resource [144] were used to define a single 5'UTR per gene. Of 19,062 MANE Select transcripts, 18,764 had annotated 5'UTRs and 18,781 3'UTRs. Notably, 298 (1.6%) MANE Select transcripts do not have an annotated 5'UTR and were excluded.

UTR length was calculated as the total length of all exons for each UTR using genomic coordinates. Length = (end of exon - start of exon) +1.

The MANE select transcripts contain annotated 5'UTR exons. To determine the number of introns per 5'UTR, I calculated the total number of 5'UTR exons and subtracted one for each transcript. To account for potential introns that may occur between the last 5'UTR exon and the CDS start, I used the CDS start codon position (from the MANE select transcript). If there was a distance between the last 5'UTR exon and the CDS start, I considered it an additional intron and added it to the count.

3.4.2 Identifying and classifying uAUGs

I only considered the canonical AUG start codon. I identified all AUGs in the sequence of each 5'UTR as upstream AUGs (uAUGs). Each uAUG was then annotated as one of the following categories, and these annotations are used throughout this thesis:

1. As a start-stop, if the uAUG was immediately followed by a stop codon.
2. As a uORF if there was an in-frame stop codon (UAA, UAG, UGA) within the 5'UTR. Where there were multiple uAUGs to one stop codon, all were considered uORFs. Where there were multiple stop codons to one uAUG, only the first stop codon was considered and counted as one uORF.
3. As an oORF if there was no in-frame stop codon within the 5'UTR. These were further subdivided into out-of-frame oORFs if the uAUG was not in-frame with the CDS, or in-frame N-terminal extensions (NTEs) if the uAUG was in-frame to the CDS.

To determine where in the 5'UTR an AUG was positioned, I took its position within the exonic sequence and divided it by the 5'UTR exon length.

Extracting the MANE UTR sequences was done by Frederik Heymann Lassen.

3.4.3 uAUG consensus motifs

As the TE motifs from Noderer *et al.* [41] is an 11 bp context (XXXXXXAUGXX), where the uAUG TE sequence was not complete as too close to the start of the

5’UTR, these uAUGs were excluded from this analysis. Additionally, 5’UTRs <6bp long were excluded as a TE cannot be identified.

For the Kozak set, I used the original Kozak consensus sequence findings (GCCA/GCCAUGG) [42]. If the -3 base is an A or G and the +3 is a G relative to the first base of the uAUG, this is considered a “strong” Kozak sequence. A “moderate” Kozak strength requires only one of these conditions, and it is “weak” if neither are present. The -3 and +3 base is described as having the biggest effect on translation initiation [42]. Where the uAUG consensus sequence was not complete as too close to the start of the 5’UTR, these uAUGs were excluded from this analysis.

3.4.4 Ribo-seq Data

Ribo-seq data (detailed in section 2.2) was downloaded from https://smorfs.ddnetbio.com/ [31] and filtered to include only uORFs. The code to determine the overlap between Ribo-seq uORFs and the predicted uORFs in the MANE Select set was done by Maria Fernandes.

3.4.5 Codon Shuffle

To investigate codon composition and explore whether there are biases in some codons appearing more or less frequently in 5’UTRs, a statistical method which utilises shuffling of the sequences can be used. Each 5’UTR sequence is shuffled at random many times, whilst preserving the di-nucleotide composition; keeping every 2 bases together. This is important as the occurrence of nucleotide pairs is not random

and can reflect important underlying biology and ensures that whilst shuffling, the overall sequence composition remains somewhat representative of the original sequence. One can then compare the number of codons in the shuffled sequence ("expected") to the unshuffled original sequence ("observed") to determine whether a codon appears more or less frequently than expected by chance.

All MANE 5'UTR sequences were shuffled 1000 times while maintaining di-nucleotide composition using the uShuffle package [163]. Counting the occurrence of each codon, I calculated the average codon count per gene (codon count/1000) to generate an "expected" codon count per 5'UTR. In the unshuffled MANE 5'UTR sequences I counted the occurrence of each codon to generate the "observed" codon count per 5'UTR. Per gene, I generated an o/e by dividing the observed codon counts by the expected. Once I had an o/e per gene, I calculated the mean o/e for each codon across all 5'UTRs. The shuffling of the MANE 5'UTR sequences using uShuffle was done by Frederik Heymann Lassen. Elston D'Souza aided in code writing for calculating the o/e.

3.4.6 PhyloP and CADD

5'UTR bases were annotated with per-base vertebrate PhyloP scores (phyloP100way, alignments of 99 vertebrate genomes to the human genome) [145] retrieved in R using the GenomicScores package by Alex Martin-Geary. Separate means were calculated for each gene across (1) all 5'UTR bases, (2) all uORF start and stop codons within the 5'UTR, and (3) all bases of start-stops within the 5'UTR.

CADD (146) v1.6 scores were extracted using the CADD version 2.2.0 release files and tabix (HTSlib v1.9: foss/2018b) to filter MANE 5'UTR coordinates. This was done by Elston N. D'Souza. Separate means were calculated for each gene across (1) all 5'UTR bases, (2) all uORF start and stop codons within the 5'UTR, and (3) all bases of start-stops within the 5'UTR.

The PhyloP and CADD methods described here are the same ones used throughout the rest of the thesis.

3.4.7 Secondary Structure Predictions

Minimum free energy (MFE) predictions can be done using algorithms that calculate the free energy associated with folding using thermodynamic parameters which include stability of base pairs and other factors that can influence stability of RNA secondary structures. The Vienna RNA package is one such algorithm to calculate MFE. It was downloaded 28 July 2022 (https://www.tbi.univie.ac.at/RNA/index.html) and I used the RNAFold v2.5.1 program on MANE Select 5'UTR full exonic sequences to predict the MFE scores.

The GC % content of each 5'UTR was calculated by dividing the total number of G and C bases by the length of the 5'UTR.

$$GC\ content = \frac{G + C}{UTR\ Length} \times\ 100$$

3.4.8 Statistical tests for Chapters 3-6 (inclusive)

To account for multiple testing, I calculated a study-wide (Chapters 3-6 inclusive) *P*-value threshold of 5.9×10^{-4} using a Bonferroni correction based on 85 statistical tests. All *P*-values less than 1×10^{-15} are reported as $P < 1 \times 10^{-15}$.

3.4.9 Code availability for Chapters 3-6 (inclusive)

All code and datasets that were used to generate the results throughout this thesis is available at https://github.com/nechamawieder/utr_thesis.

Chapter 4

Investigating how 5'UTRs vary by gene tolerance to LoF

4.1 Introduction

As demonstrated in the previous chapter, 5'UTRs vary dramatically across human genes. Previous work has shown that 5'UTRs of genes where heterozygous loss-of-function (LoF) variants cause developmental disorders (DD) are longer and have more introns than all other genes [86]. Variants that create uAUGs or disrupt uORFs are under stronger negative selection in genes that are intolerant to LoF [88]. Based on these findings, I wanted to extend this work and investigate how 5'UTRs vary by gene tolerance to changes in dosage. Genes differ in their dosage sensitivity; there is a range of tolerance to increases and/or decreases in gene expression levels. In other words, too much or too little gene expression can be deleterious. I hypothesised that genes that are sensitive to changes in dosage and hence need to tightly control the levels of protein they produce may have different 5'UTRs to genes that are tolerant to dosage changes.

The Genome Aggregation Database (gnomAD) has scored protein-coding genes along a spectrum that represents tolerance to inactivation, termed the "loss-of-function observed/expected upper bound fraction" (LOEUF) [147]. This method compares the observed number of predicted loss-of-function (pLoF) variants to the

expected number per gene. Low LOEUF scores are indicative of genes intolerant to LoF, while high LOEUF scores indicate genes which are tolerant to LoF. Here, I used gene tolerance to LoF as a proxy for overall dosage sensitivity. I used LOEUF scores to bin genes into deciles of intolerance to LoF [147] and assessed 5'UTR features across LOEUF deciles. I found that genes most intolerant to LoF had distinctly different and more complex 5'UTRs that are enriched for cis-acting regulatory elements (including uAUGs).

4.2 Results

4.2.1 Genes that are intolerant to LoF have longer 5'UTRs

As 5'UTR length may be a proxy for containing more regulatory elements, I wondered whether there would be a difference in 5'UTR length between genes with varying tolerance to LoF. 5'UTR length increases with decreased tolerance to LoF (Figure 4.1), with the 5'UTRs of genes in the lowest LOEUF quintile being significantly longer than those in the highest LOEUF quintile (mean length 269 bp vs 162 bp; Wilcoxon $P<1x10^{-15}$). In other words, genes that are intolerant to LoF have significantly longer 5'UTRs. It is important to note that LOEUF is linked with CDS length as LOEUF is calculated on variants within CDS (Figure 4.2), with shorter genes having less confident LOEUF estimates. Longer genes have typically higher expected variants leading to tighter confidence intervals which is what LOEUF is derived from. Therefore, as shorter genes have less confident LOEUF estimates, I repeated this

analysis after removing the shortest genes - the bottom 10% of CDS length. The results remained significant; genes most intolerant to LoF still had longer 5'UTRs (Figure 4.3; Wilcoxon $P<1x10^{-15}$). Additionally, to mitigate a potential confounding effect of an artificial correlation between 5'UTR length and LOEUF, which could be due to a different mRNA length, I calculated the proportion of 5'UTRs relative to its corresponding full mRNA exonic length. This was done by showing the length of each 5'UTR as a percentage of the total mRNA length. Genes in the lower LOEUF deciles, the ones most intolerant to LoF, had a greater 5'UTR/mRNA length, indicating that 5'UTRs are longer over and above mRNA length (Figure 4.4, Wilcoxon $P<1x10^{-15}$).

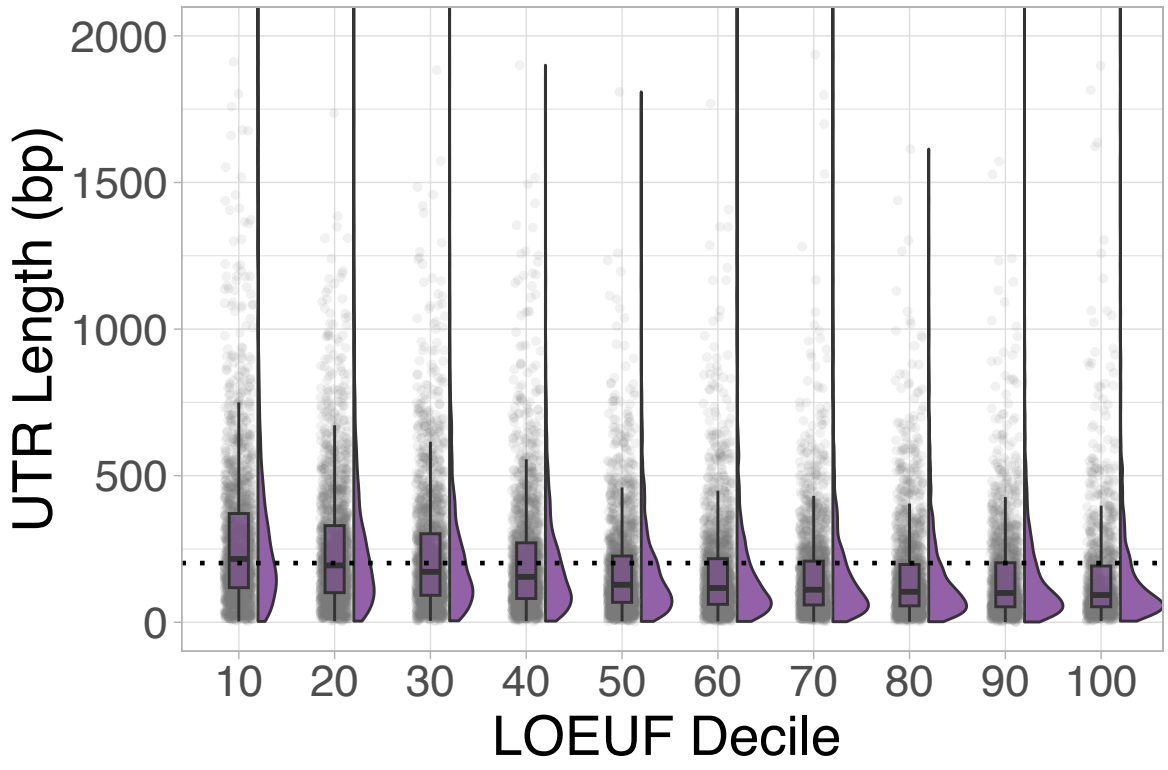


Figure 4.1: *5'UTRs increase in length with decreasing tolerance to LoF (Wilcoxon $P<1x10^{-15}$). The average 5'UTR length across all genes (202 bp) is shown by a dotted line. The y-axis was truncated at 2,000 bp (13 genes had 5'UTRs >2,000 bp). Data for this figure is available in Appendix Table 1.*

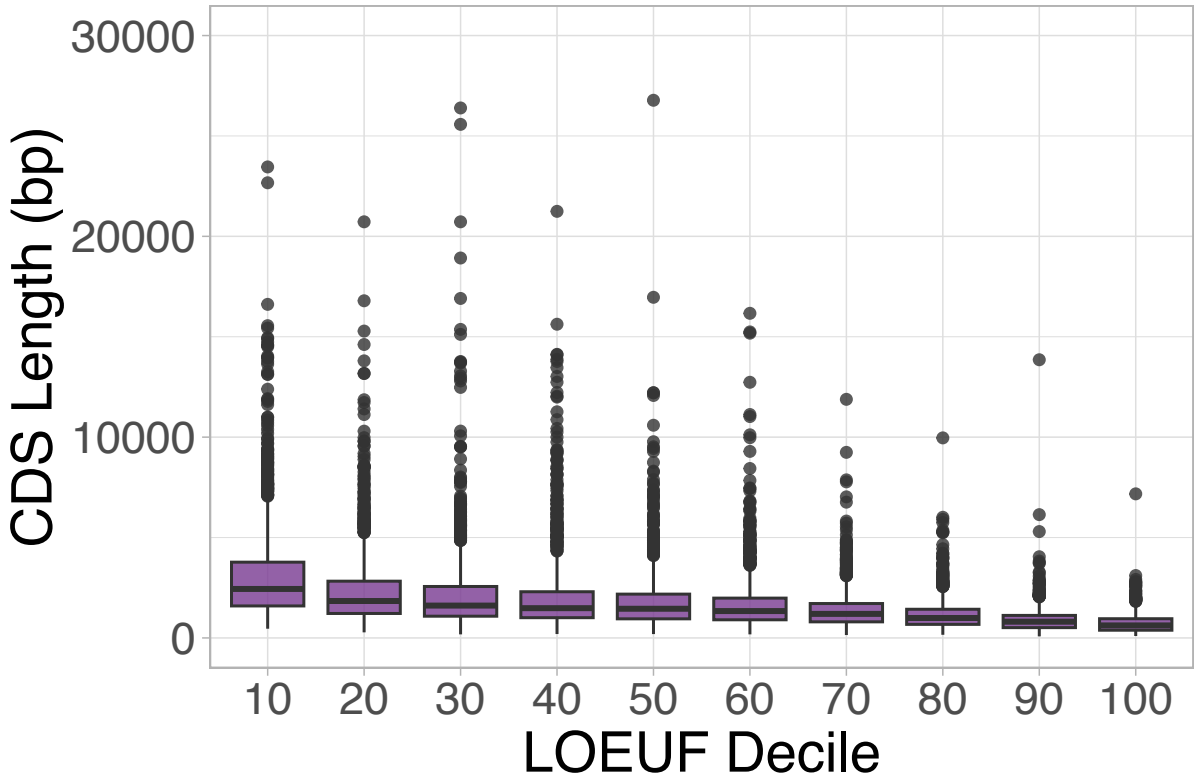


Figure 4.2: *CDS length across LOEUF deciles. Genes with longer CDS have lower LOEUF scores. The y-axis was truncated at 30,000 bp (1 gene had CDS >30,000 bp).*

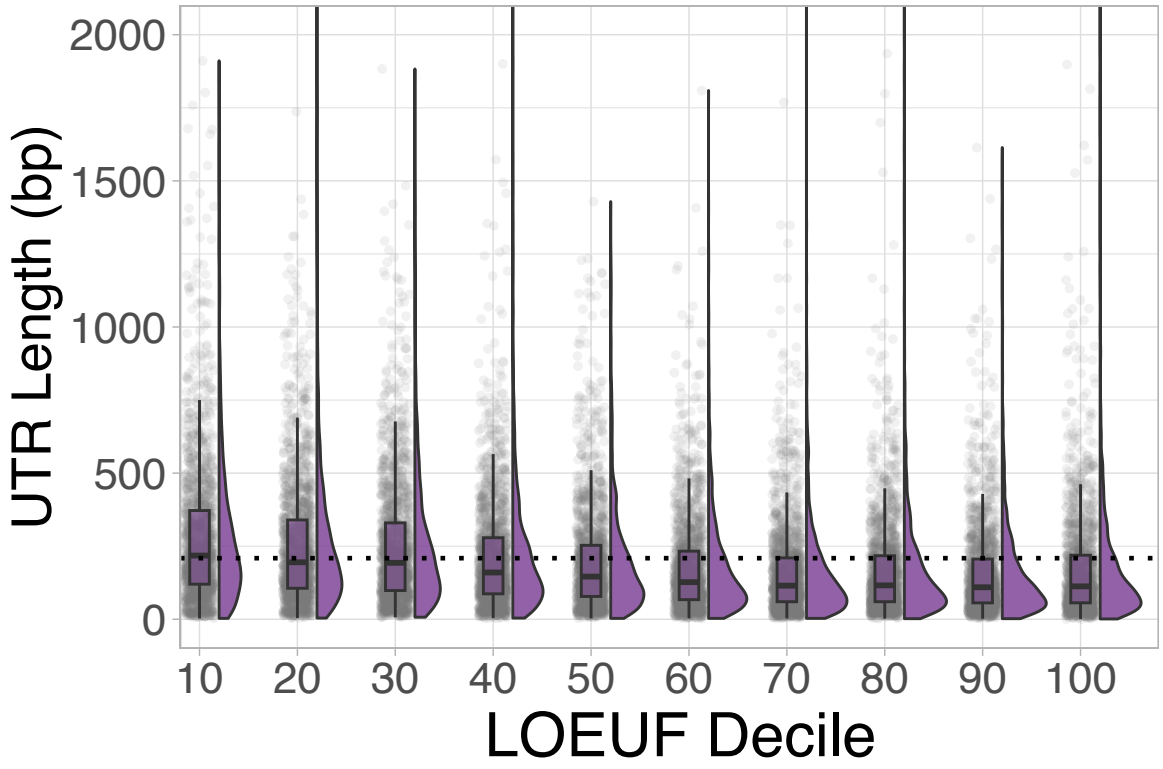


Figure 4.3: *5'UTR length by LOEUF decile with genes within bottom 10% of coding sequence length distribution removed. The average 5'UTR length across all genes (209 bp) is shown by a dotted line. The y-axis was truncated at 2,000 bp (11 genes had 5'UTRs >2,000 bp).*

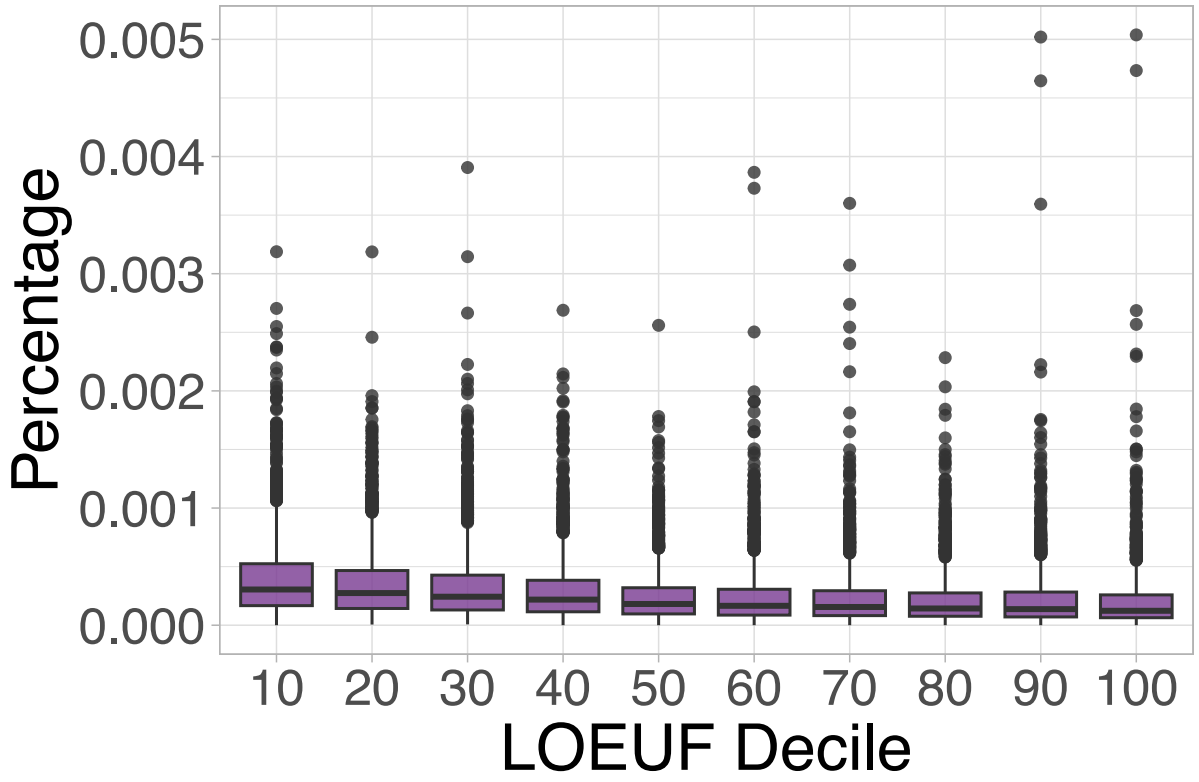


Figure 4.4: *5'UTR as a proportion of total mRNA length; 5'UTRs are a significantly larger proportion of the total mRNA in genes most intolerant to LoF (Wilcoxon P<1x10^{-15}).*

4.2.2 Genes most intolerant to LoF contain more uORFs

Next, I assessed the proportion of genes in each LOEUF decile with different categories of uAUGs. Genes most intolerant to LoF more frequently contain uORFs and start-stops than LoF tolerant genes (Figure 4.5; 46.2% vs 27.8%; P<1x10^{-15}, and 6.8% vs 4.5%; P=8.5x10^{-05} for uORFs and start-stops respectively). This is true both using predicted uORFs (Figure 4.5) and the uORFs detected by Ribo-Seq (Figure 4.6, 30.0% vs 7.4%, P<1x10^{-15}).

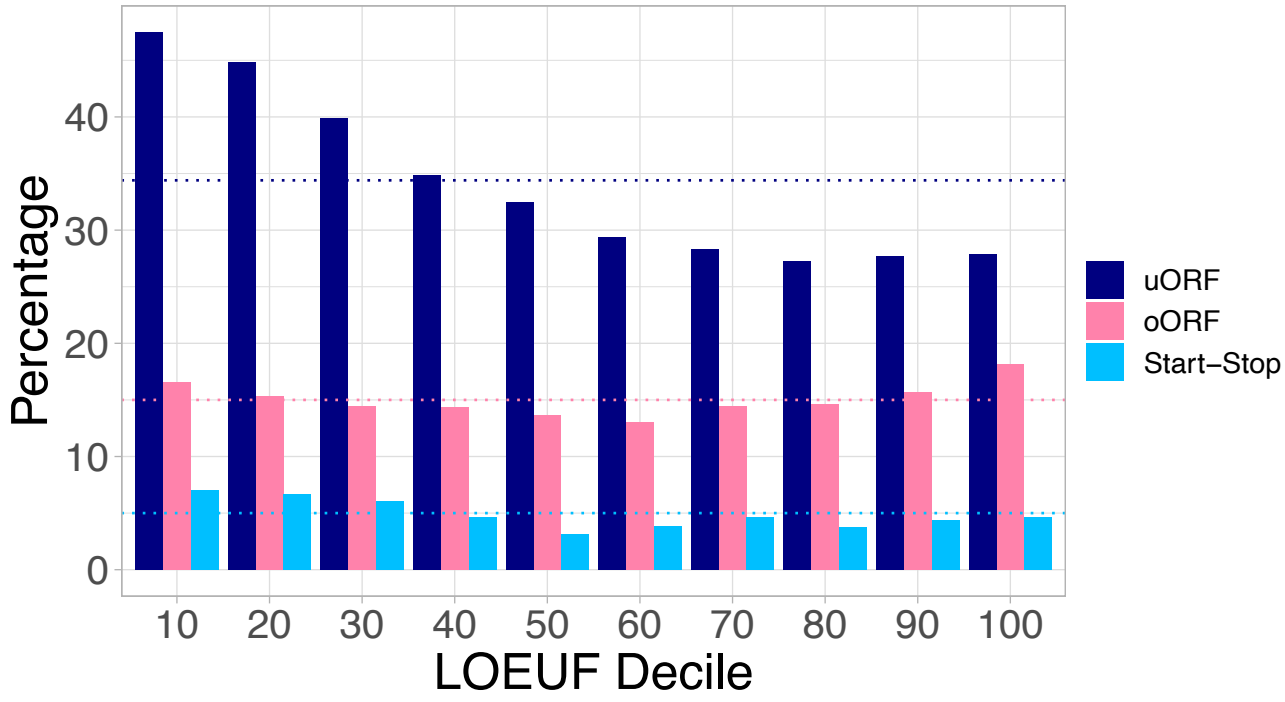


Figure 4.5: *Genes most intolerant to LoF more often have uORFs (46.2% vs 27.8%;Chi-square P<1x10^{-15}) and start-stops (6.8% vs 4.5%; Chi-square P=8.5x10^{-05}) than genes most tolerant to LoF. oORFs were not enriched in genes most intolerant to LoF (16.0% vs 16.7%; Chi-square P=0.42). The average numbers of each uAUG type across all 5'UTRs are shown by dotted lines. uORF: upstream open reading frame; oORF; overlapping open reading frame. Data for this figure is available in Appendix Table 2.*

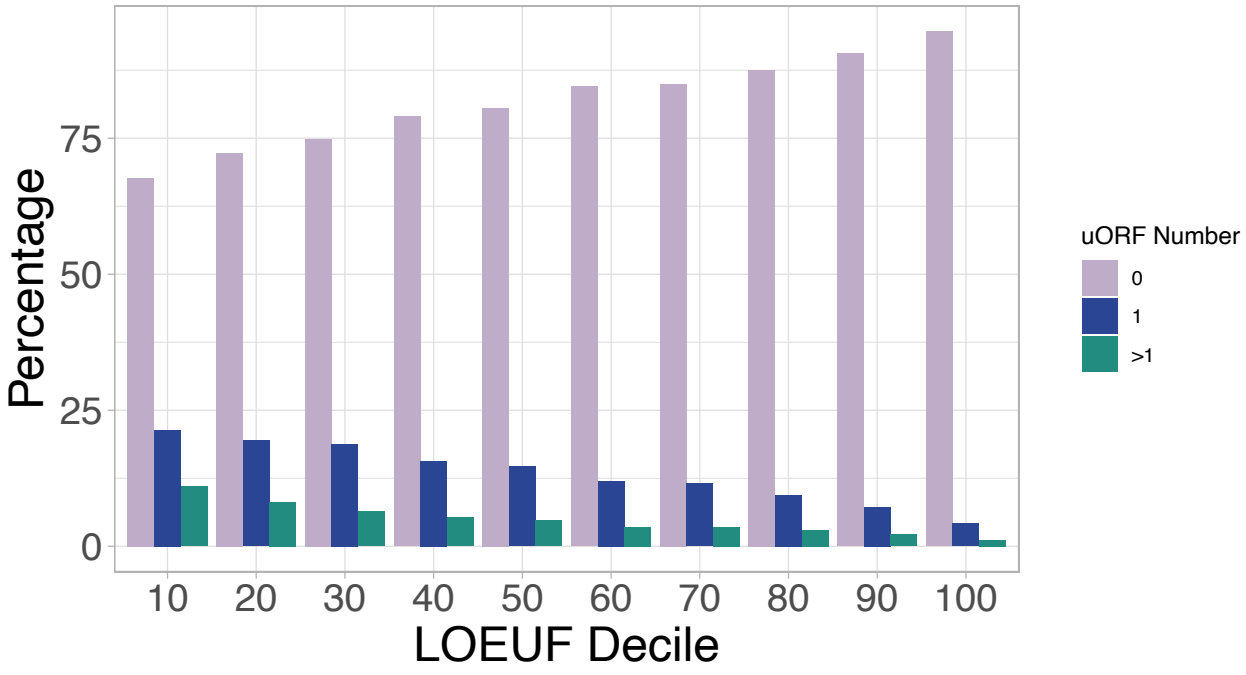


Figure 4.6: *The 5'UTRs of genes that are more intolerant to LoF more often contain Ribo-seq uORFs than genes that are tolerant to LoF (30.0% vs 7.4%; Chi-square $P<1x10^{-15}$).*

However, a major factor influencing detection of Ribo-seq uORFs is gene expression; if a uORF is highly expressed there is more chance of detecting it via Ribo-seq. To account for this, I stratified genes based on their expression using data from the Genotype-Tissue Expression (GTEx) project. GTEx analyses gene expression levels from a wide range of tissue samples [149]. I split genes into four quartiles of mean expression across tissues in GTEx, and for each quartile plotted Ribo-seq uORFs across LOEUF deciles (Figure 4.7). Genes most intolerant to LoF more frequently contain Ribo-seq uORFs across all gene expression quartiles (Chi-square, Q1 $P<1x10^{-15}$; Q2 $P=3.8x10^{-12}$; Q3 $P=6.5x10^{-14}$; Q4 $P=8x10^{-11}$).

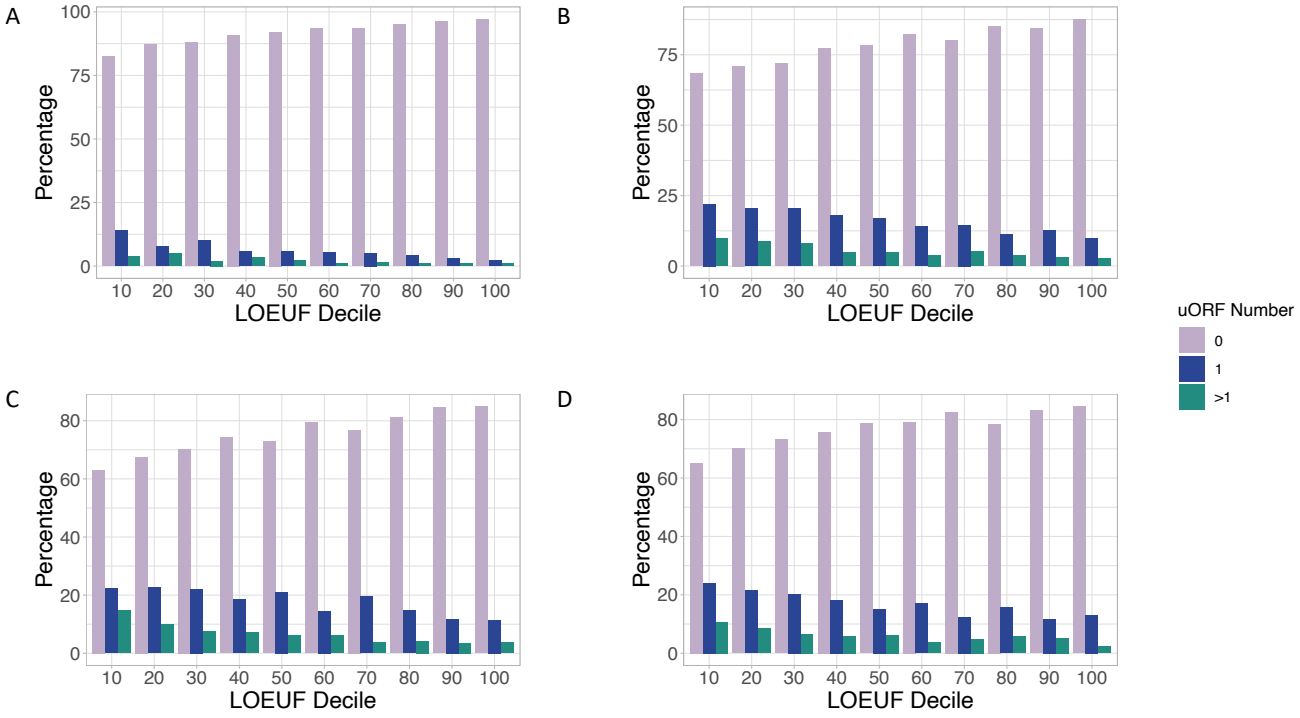


Figure 4.7: Ribo-seq uORFs with Gene Expression Data
The 5'UTRs of genes that are more intolerant to LoF more often contain Ribo-seq uORFs than genes that are tolerant to LoF. This remains true after splitting genes into four quartiles based on their expression levels in GTEx, with quartile 1 (Q1) being the lowest gene expression level to quartile 4 (Q4) being the highest gene expression level (Chi-square, ***A****: Q1 P<1x10^{-15};* ***B****: Q2 P=3.8x10^{-12};* ***C****: Q3 P=6.5x10^{-14};* ***D****: Q4 P=8x10^{-11}).*

I would expect there to be more uAUGs in LoF intolerant genes as they have longer 5'UTRs. To account for this difference in 5'UTR length across deciles, I computed the number of uAUGs per base pair (bp) in each 5'UTR. The 5'UTRs of the most LoF intolerant genes have significantly fewer uAUGs per bp compared to the most tolerant genes (Figure 4.8; mean=0.009 uAUG per bp vs 0.013 uAUG per bp; Chi-square P<1x10^{-15}), suggesting that uAUGs are selectively depleted from these genes. Despite this overall depletion, 52.1% of LoF intolerant genes (bottom quintile of

LOEUF) contain at least one uAUG, suggesting that uAUGs may play an important role in translational regulation of these genes.

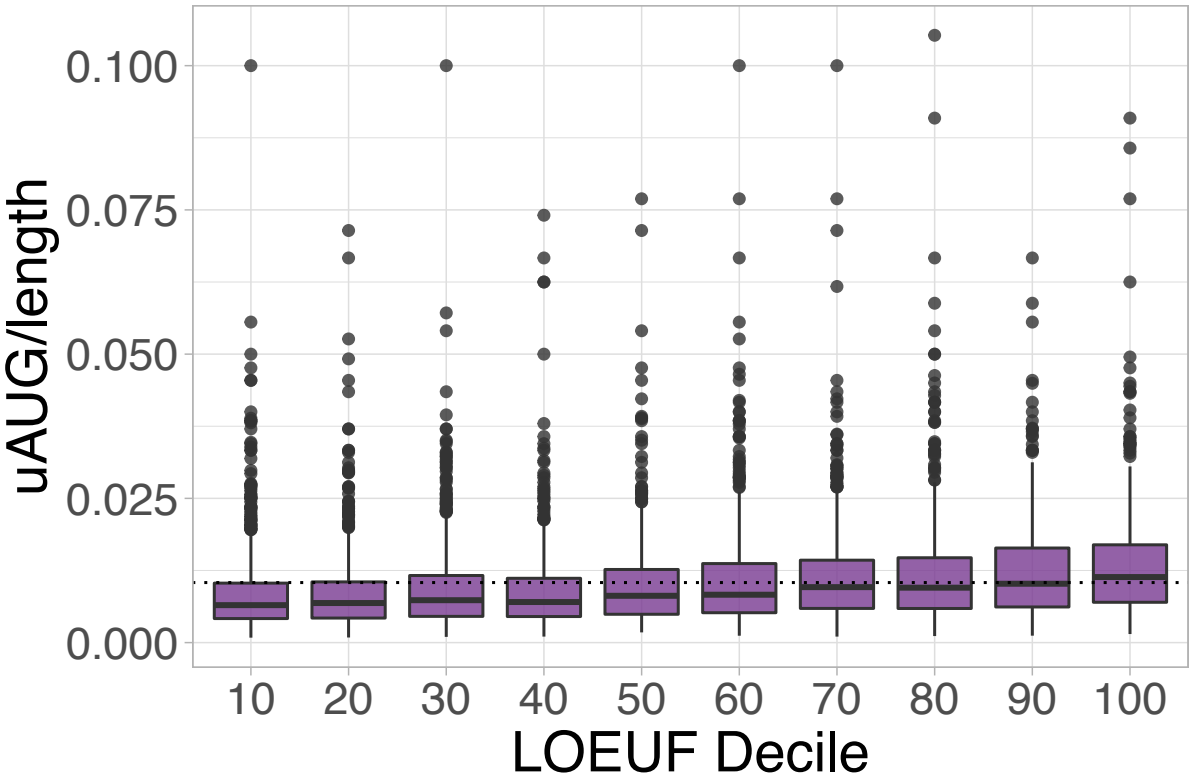


Figure 4.8: *Genes most intolerant to LoF had fewer uAUGs per base pair (mean 0.009 uAUG per bp vs 0.013 uAUG per bp; Chi-squareP<1x10^{-15}). Average uAUG/length is 0.01 (dotted line).*

4.2.3 uORF start codon consensus sequence analysis

I wanted to explore whether uORF start codons had different consensus sequences between LOEUF deciles. I wondered whether uORFs of genes most intolerant to LoF were more or less likely to be translated; were they required for stronger or weaker

repression of the CDS? To determine whether the likelihood of uORF translation, and hence strength of repression of downstream CDS translation, differed between LOEUF deciles, I compared the start contexts of predicted uORFs. I used the original Kozak consensus sequence [42] (methods in Chapter 3.4.3), and also a newer dataset of experimentally measured translational efficiencies (TE), quantified across a range of cell lines [41]. Using the Kozak consensus sequence, I categorised uORF start codon contexts into three categories (strong, moderate and weak) based on likelihood of being translated, with “strong” being the most likely to be translated to “weak” being the least favourable consensus sequence. I found that the intolerant deciles had more uORFs with “strong” Kozak consensus sequences than the tolerant deciles (Figure 4.9, 21.2% vs 17.3%; Chi-square P=9x10^{-5}). However, I saw no significant difference in TE of predicted uORFs across deciles (Figure 4.10A, Wilcoxon P=0.6).

Additionally, as canonical AUG start codons are the most efficiently translated start codons compared to non-canonical start codons [164], I assessed the Ribo-seq uORF start codons across LOEUF deciles. There was no significant enrichment of canonical over non-canonical start site usage of the Ribo-Seq uORFs (Figure 4.10B; Chi-square, P=0.18). Together, these results show that predicted uORFs in genes most intolerant to LoF have enhanced translation potential as they have stronger Kozak consensus sequences. However, this was not reflected using the TE scores for predicted uORFs or in canonical site usage in Ribo-seq uORFs.

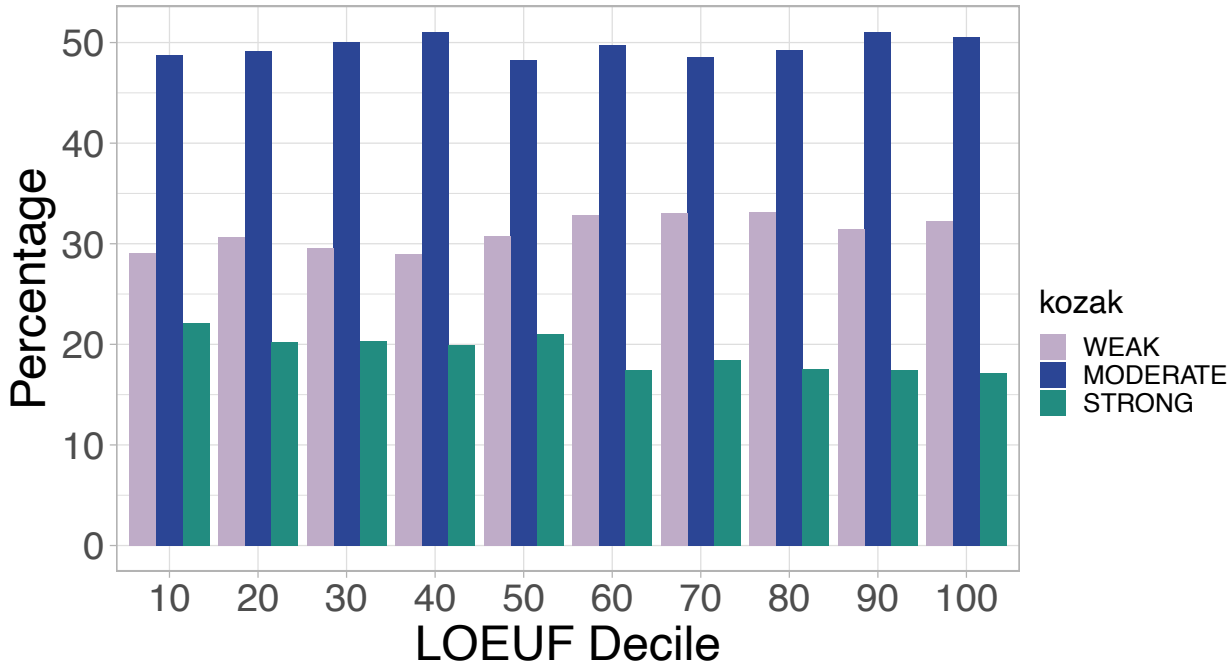


Figure 4.9: *Kozak consensus sequences for uORFs across LOEUF deciles. Intolerant deciles had more uORFs with "strong" Kozak consensus sequences than the tolerant deciles (Chi-square $P=9x10^{-5}$).*

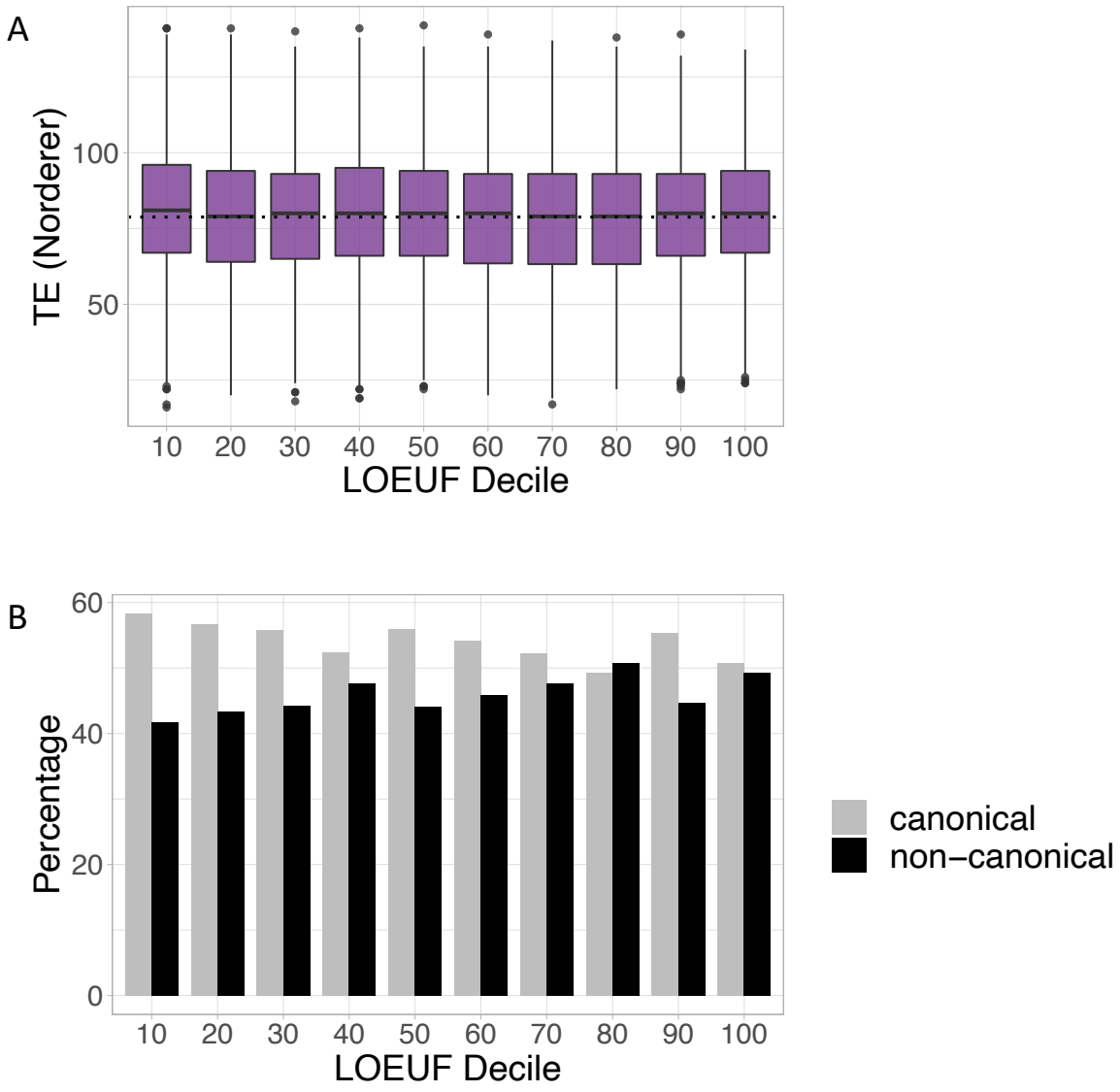


Figure 4.10: uAUG Translation

A) *There was no difference between translational efficiencies (TE) for predicted uAUGs across LOEUF deciles (Wilcoxon P=0.6).* ***B****) Whilst Ribo-seq uORFs in genes intolerant to LoF appear to more frequently have canonical start-codons, this difference is not statistically significant (Chi-square P=0.18). Data for this figure is available in Appendix Table 3.*

4.2.4. uORF features that impact reinitiation

I hypothesised that the uORFs in LoF intolerant genes might be optimised to be less disruptive to CDS translation. After translating a uORF, the ribosome can reassemble and translate the CDS. The efficiency of this ribosome reinitiation has been observed to be dependent on the length of the uORF, its sequence, and the distance between the end of a uORF and start of the CDS [165–167]. uORFs have been found to be depleted in the 100 bp region immediately upstream of the CDS, suggesting that uORFs close to the CDS are selected against as they are more repressive of CDS translation [168].

Some amino acid codons are translated more efficiently (“optimal” codons) than others (“non-optimal”). This distinction is thought to be driven by the balance of tRNA availability versus codon demand [169]. To define this, codon-specific tRNA adaptive indices (tAIs) can be calculated, which is a measure of codon optimality. tAIs are calculated based on the abundance of intracellular tRNA and binding strength between a codon and tRNA. I used tAIs generated from mammalian cells to calculate a codon optimality score per Ribo-seq uORF [170]. I hypothesised that uORFs consisting of overall optimal codons may be translated more efficiently and quickly, and be less disruptive to protein production. This could hold true whether the ribosome dissociates or reinitiates after; there would be less chance of ribosomal collisions and quicker reinitiation. However, I found no significant differences of uORF codon optimality between deciles (Figure 4.11; Wilcoxon P=0.17). It is important to

note that defining codon optimality is complex; it has been shown that alterations in cellular tRNA availability is tightly coordinated with changes in mRNA expression [205]. These changes impact translational profiles and differ between differentiating and proliferating cells. Therefore, assessing codon optimality in this static gene set does not consider changes in mRNA expression and other factors; in which cells these uORFs may be translated and the tRNA availability at specific time points.

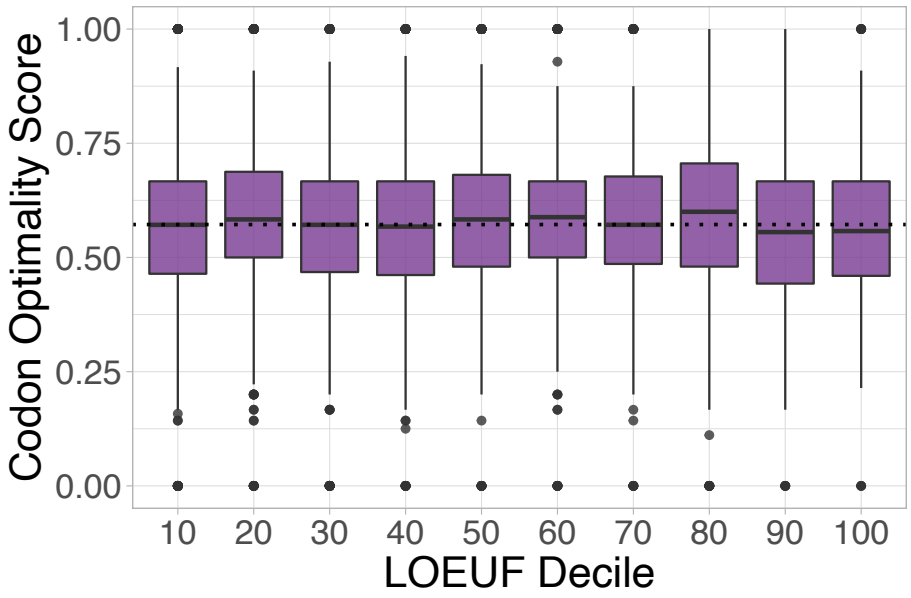


Figure 4.11: *There was no significant difference in codon optimality score for Ribo-seq uORFs between lowest and highest two deciles (Wilcoxon P=0.17), average for all genes is dotted line.*

Shorter uORFs have been demonstrated to allow for more efficient reinitiation than longer uORFs [166,171]. I observed a small, but significant difference in average uORF length across deciles (means 52.5 bp vs 59.1 bp, Wilcoxon P=4.9x10^{-06}), but only when considering the predicted uORF and not the Ribo-Seq set (Figure 4.12A, 4.12B; Wilcoxon P=0.9).

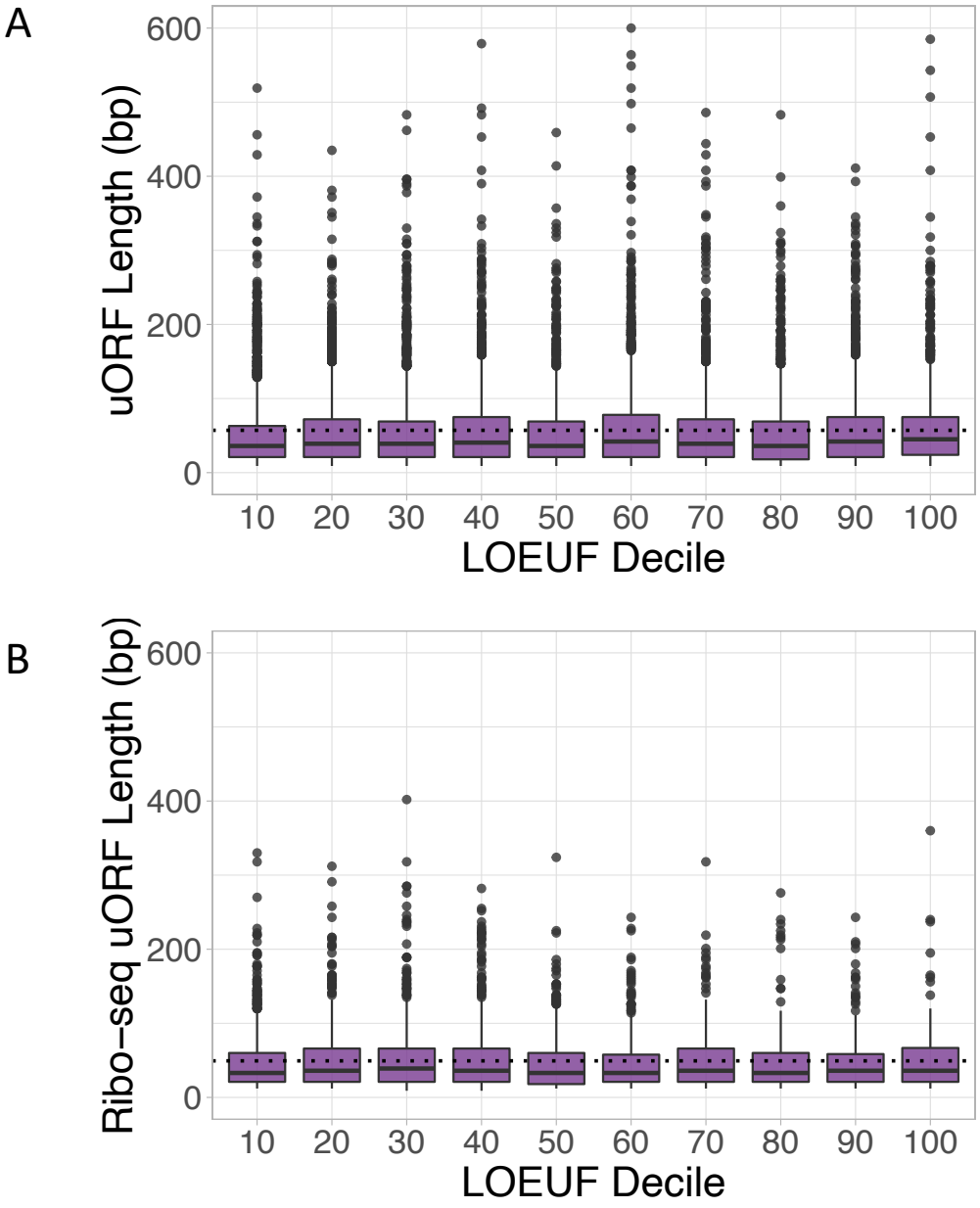


Figure 4.12: uORF Lengths

***A**) Predicted uORFs in the lower two deciles were slightly shorter compared to the highest two deciles (Wilcoxon P=4.9x10^{-06}). Mean uORF length is shown as a dotted line. The y-axis is truncated at 600bp (7 uORFs were >600bps). **B**) There is no significant difference in Ribo-seq uORF lengths between lowest and highest two LOEUF deciles(Wilcoxon, P=0.9). Mean uORF length is shown as a dotted line. The y-axis is truncated at 600bps (1 uORF was >600bps).*

I wanted to observe the uORF to CDS distance and see if this differed between LoF intolerant genes compared to LoF tolerant genes. I calculated the distance from the stop codon of the last uORF in each 5'UTR to the CDS start. In LoF intolerant genes, uORFs are significantly further upstream of the CDS start (Figure 4.13; means 99 bp vs 77 bp, Wilcoxon P=1.3x10^{-04}). In other words, uORFs in genes most intolerant to LoF are further away from the CDS; these genes have a greater potential reinitiation distance and are possibly less disruptive to CDS translation.

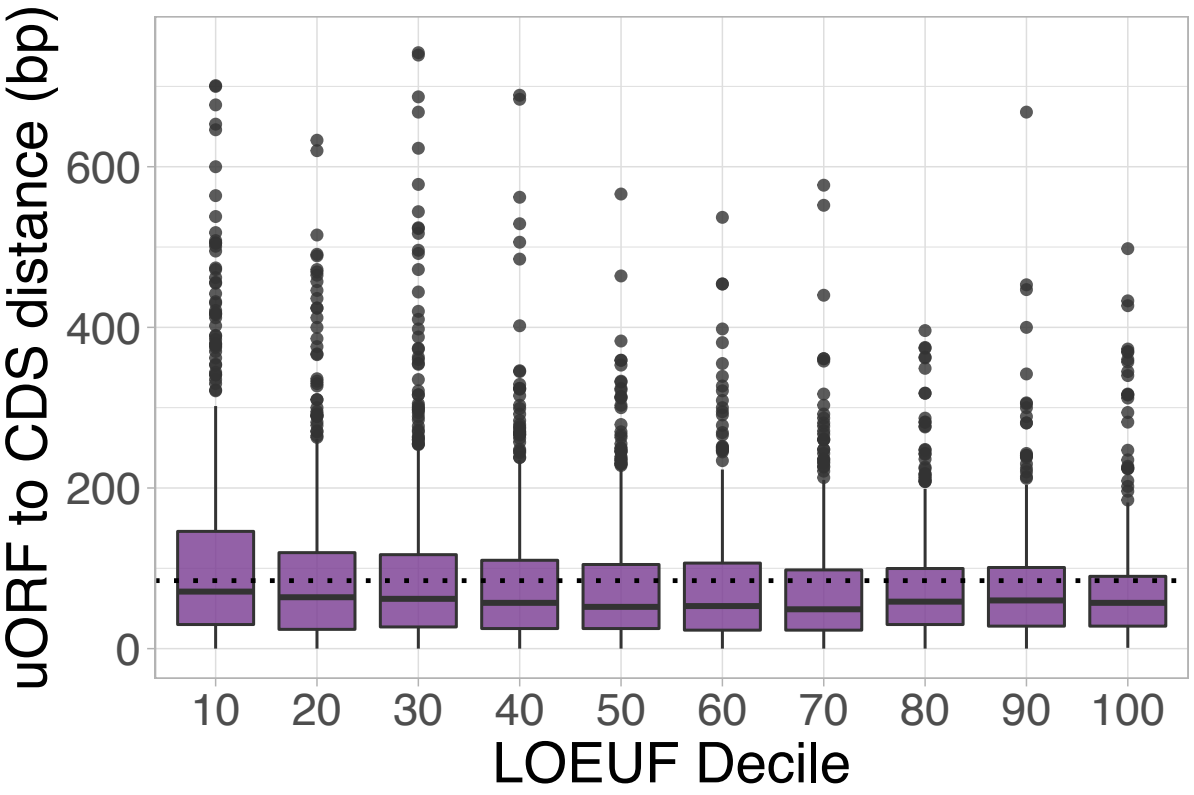


Figure 4.13: *The distance between the stop codon of the last uORF to the CDS start (of the predicted uORF set) was greater in the genes most intolerant to LoF (Wilcoxon P=1.3x10^{-04}). Average distance is 84.7bps (dotted line). The y-axis is truncated at 750bp (3 genes had distance >750bps).*

4.2.5 5'UTRs of LoF intolerant genes are more conserved

I hypothesised that if the 5'UTRs of dosage sensitive genes play a significant role in regulating gene product dosage then they would be more highly conserved across species and genetic variants across them more likely to be deleterious. To test this, I used two different metrics, PhyloP and CADD (background in sections 2.4 and 2.5). I used PhyloP scores [145] to assess whether 5'UTRs across the LOEUF spectrum have differences in their degree of conservation between 100 species. The 5'UTRs of LoF intolerant genes are more highly conserved than LoF tolerant genes, shown by significantly higher PhyloP scores (Figure 4.14A; T-test $P<1x10^{-15}$). This is even more pronounced when looking specifically at start and stop codons of predicted uORFs and start-stops (Figure 4.14A; T-test all $P<1x10^{-15}$). Combined Annotation-Dependant Depletion (CADD) scores measure variant deleteriousness for all possible single nucleotide substitutions at each position [146]. Higher CADD scores indicate higher likelihood that a variant at that position is deleterious. I assessed whether 5'UTRs as a whole, uORF start and stops and start-stops had significantly different CADD scores across the LOEUF spectrum. I found that there is a similar pattern with CADD scores as PhyloP, with CADD scores increasing with decreased LoF tolerance (Figure 4.14B). As with PhyloP, start-stops showed the most marked difference in CADD scores across deciles.

To demonstrate that the higher PhyloP and CADD scores is greater than what would be expected given the increased length of LoF intolerant 5'UTRs (given that they

have longer sequences), I repeated the analysis only on 5'UTRs between 100-300 bp in length. The results remained significant; (both T-test $P<1x10^{-15}$)

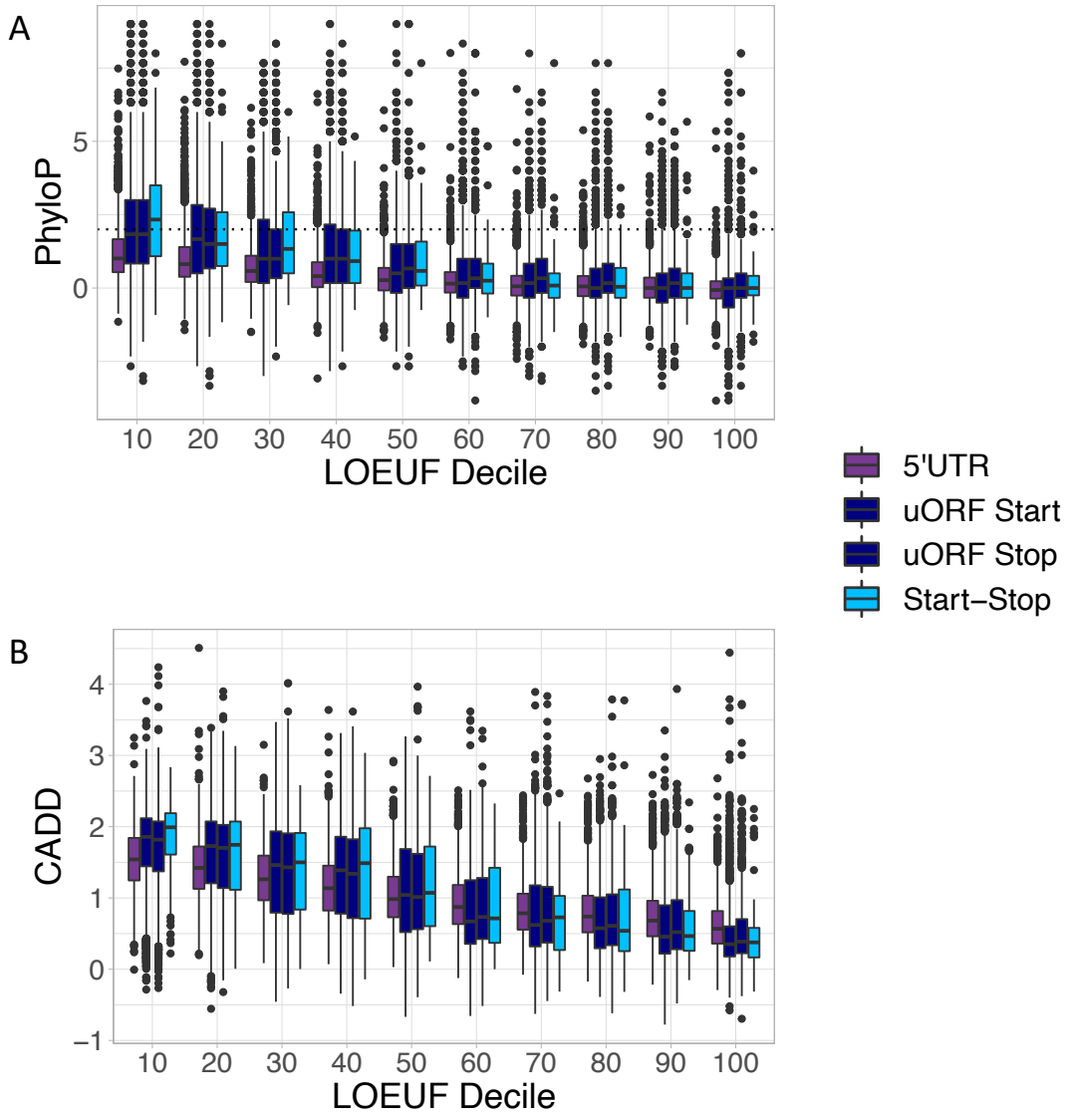


Figure 4.14: PhyloP and CADD
A) *The 5'UTRs of genes most intolerant to LoF are more conserved. Average PhyloP scores are plotted for 5'UTRs, uORF start codons, uORF stop codons and start-stops. The dotted line denotes PhyloP=2. Data for this figure is available in Appendix Table 4.* ***B****) CADD scores for 5'UTRs, uORFs and start-stops across LOEUF deciles were higher in lower deciles; in genes more intolerant to LoF (T-test 5'UTRs $P<1x10^{-15}$, uORF start $P<1x10^{-15}$, uORF stop $P<1x10^{-15}$, start-stop $P<1x10^{-15}$).*

4.2.6 Increased Transcription Start Site Usage in Genes Intolerant to LoF

Whilst I have used the MANE Select transcript set to limit the above analyses to a single, representative transcript per gene, many genes have multiple 5'UTR isoforms. Alternative transcription start site (TSS) usage and alternative splicing are major contributors to transcript isoform diversity and gene regulation [172].

Assessing 5'UTR alternative splicing possibilities, there was no significant difference in the proportion of genes that have 5'UTR introns across LOEUF deciles (Figure 4.15, Chi-square *P*=0.19).

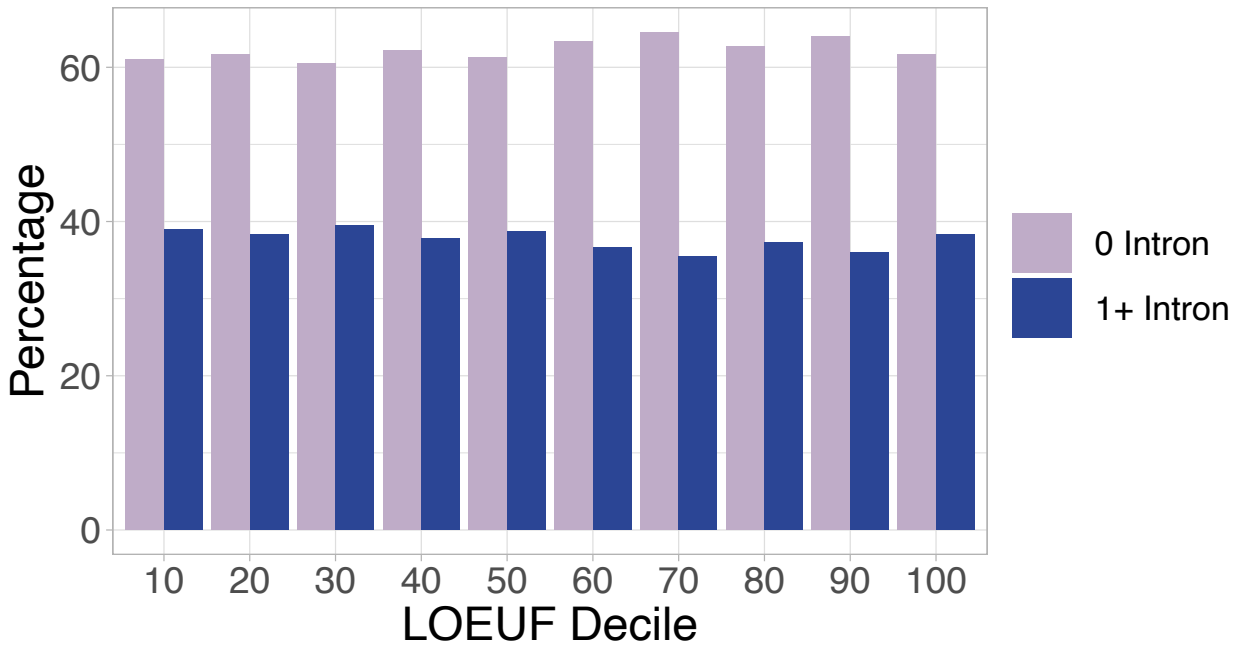


Figure 4.15: *There was no significant difference between 5'UTR intron proportions across LOEUF deciles (Chi-square P=0.19).*

Cap Analysis of Gene Expression (CAGE) is a technique used to identify and quantify 5' ends of mRNA transcripts, allowing us to analyse TSS usage. To observe the TSS diversity of 5'UTRs across the LOEUF spectrum, I used CAGE data from the FANTOM5 consortium [150]. Genes most intolerant to LoF were significantly more likely to have multiple associated CAGE peaks, i.e., greater transcript isoform diversity, when compared to genes most tolerant to LoF (Figure 4.16; CAGE peak >1, 91.9% vs 72.4%, Chi-square $P<1x10^{-15}$; CAGE peak ≥6, 44.6% vs 16.3%, Chi-square $P<1x10^{-15}$).

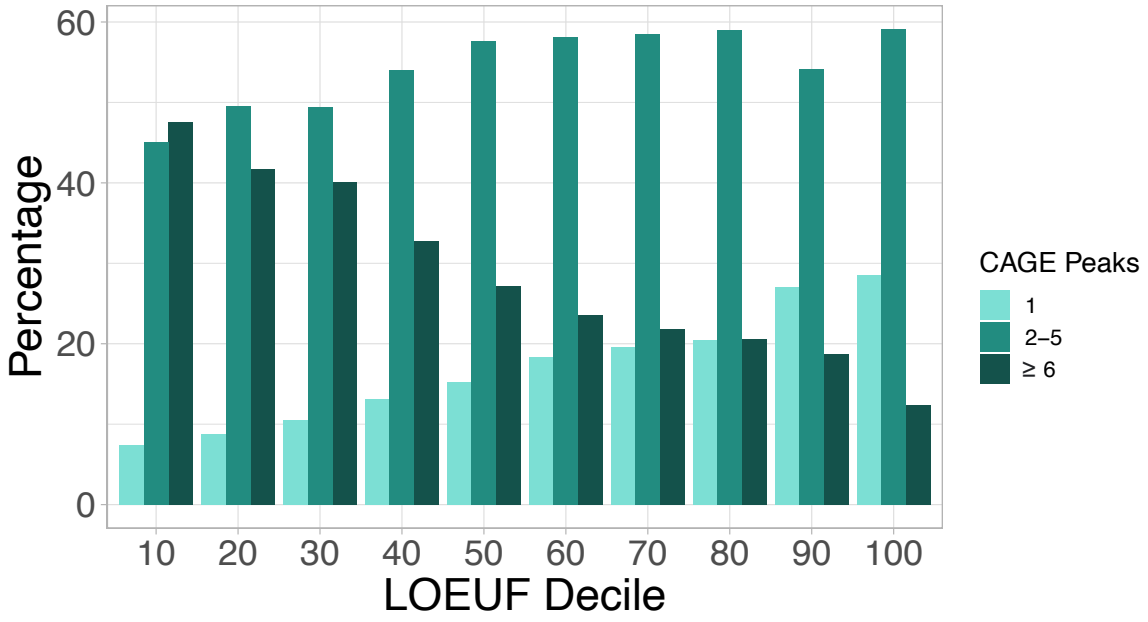


Figure 4.16: *Genes most intolerant to LoF were significantly more likely to have multiple associated CAGE peaks when compared to genes most tolerant to LoF (Figure 5; CAGE peak >1, 91.9% vs 72.4%, Chi-square $P<1x10^{-15}$; CAGE peak ≥6, 44.6% vs 16.3%, Chi-square $P<1x10^{-15}$). Data for this figure is available in Appendix Table 5.*

As this analysis may be confounded by gene expression levels, with more highly expressed genes having more associated CAGE peaks, I used GTEx data as detailed above and the result remained significant in all four quartiles (Figure 4.17, all Chi-square A: Q1 >1 CAGE peak $P<1x10^{-15}$ ≥6 CAGE peak $P<1x10^{-15}$; B: Q2 >1 CAGE peak $P=3.8x10^{-12}$ ≥6 CAGE peak $P<1x10^{-15}$; C: >1 CAGE peak Q3 $P=6.5x10^{-14}$ ≥6 CAGE peak $P<1x10^{-15}$; D: Q4 >1 CAGE peak $P=8x10^{-11}$ ≥6 CAGE peak $P<1x10^{-15}$).

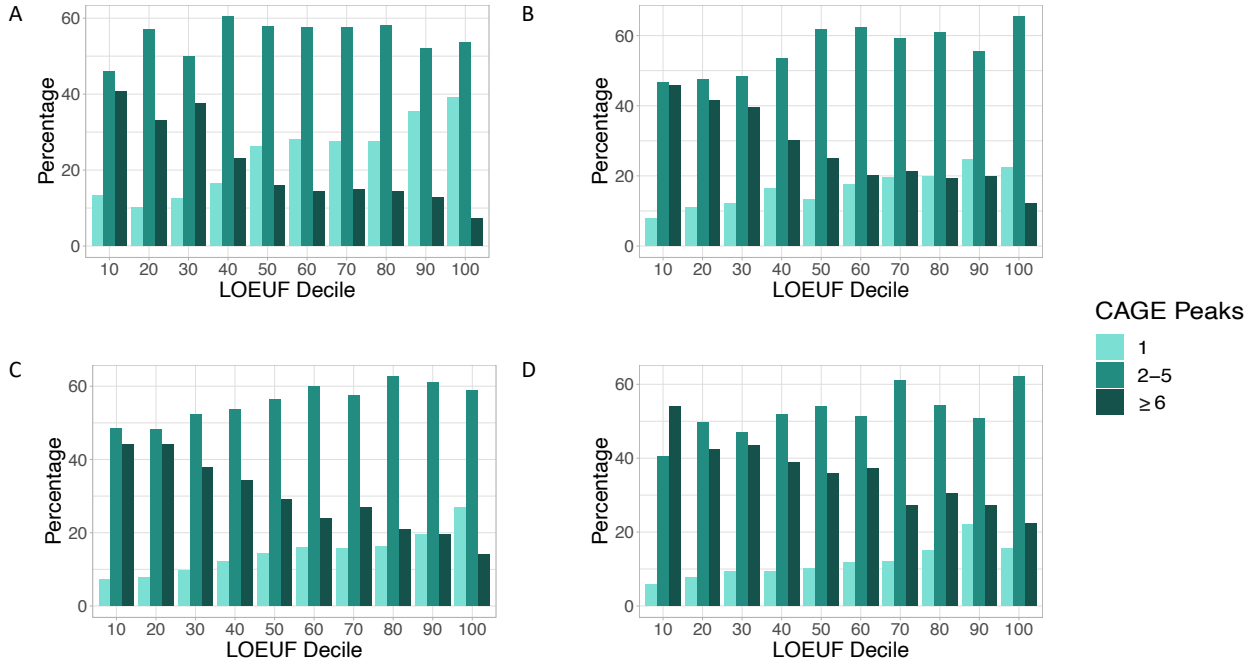


Figure 4.17: CAGE Gene Expression
*CAGE Gene Expression Genes most intolerant to LoF were significantly more likely to have multiple associated CAGE peaks when compared to genes most tolerant to LoF. This remains true after splitting genes into four quartiles based on their expression levels in GTEx, with quartile 1 (Q1) being the lowest gene expression to quartile 4 (Q4) being the highest gene expression level (All Chi-square, **A**: Q1 >1 cage peak $P<1x10^{-15}$ ≥6 cage peak $P<1x10^{-15}$; **B**: Q2 >1 cage peak $P=3.8x10^{-12}$ ≥6 cage peak $P<1x10^{-15}$; **C**: >1 cage peak Q3 $P=6.5x10^{-14}$ ≥6 cage peak $P<1x10^{-15}$; **D**: Q4 >1 cage peak $P=8x10^{-11}$ ≥6 cage peak $P<1x10^{-15}$).*

4.2.7 Genes Intolerant to LoF Have Increased Propensity to Form Secondary Structures

Secondary structures within 5'UTRs are thought to cause inefficient ribosomal scanning impacting translation [52] and are negatively correlated with mRNA stability [20]. The propensity of a sequence to form RNA secondary structures can be predicted from high GC content and low minimum free energy (MFE) of predicted secondary RNA structures [20,160]. I wondered whether there were significant differences of GC content and MFE scores of 5'UTRs between LOEUF deciles; in other words, are certain genes 5'UTRs more or less likely to be structured? I calculated the MFE score and GC content per 5'UTR [161] (methods in Chapter 3, 3.4.6). The more negative the MFE score the greater the propensity for secondary structures to form. The most LoF intolerant genes had lower MFE scores (Figure 4.18A: mean MFE=-115 vs -55, Wilcoxon $P<1x10^{-15}$) and a higher GC content (Figure 4.18B; mean=67.3% vs 59.9%; Wilcoxon $P<1x10^{-15}$) than LoF tolerant genes, indicating a higher likelihood for the 5'UTRs of LoF intolerant genes to be structured. To demonstrate that this greater propensity to create secondary structures is over and above what would be expected given the increased length of LoF intolerant 5'UTRs (given that longer sequences have a greater propensity to create secondary structures), I repeated the analysis only on 5'UTRs between 100-300 bp in length. The results for both MFE and GC content remained significant (both Wilcoxon $P<1x10^{-15}$). However, 100-300bp is still a wide range especially as MFE is so dependent on length; the more bases the more likely for adjacent base pairing. These

results suggest that genes that are intolerant to LoF are more likely to have stable secondary structures within their 5'UTRs.

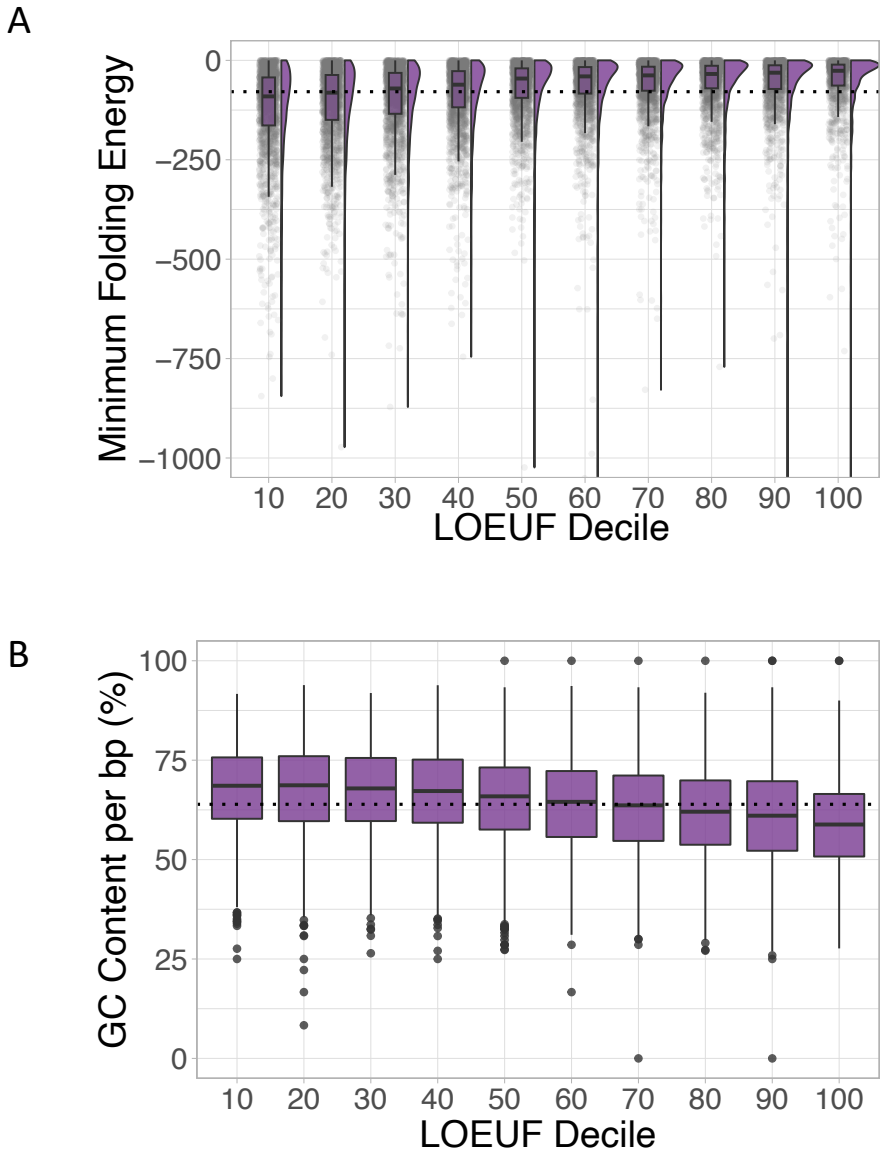


Figure 4.18: 5'UTR Secondary Structure Propensity
A) *The 5'UTRs of genes most intolerant to LoF have lower minimum free energy (MFE) scores, representing a higher propensity to fold and create structured mRNAs (Wilcoxon $P<1x10^{-15}$). The average MFE across all 5'UTRs is shown as a dotted line (-78.8). The y-axis was truncated at -1,000 (6 genes had MFE <-1000). Data for this figure is available in Appendix Table 6.* ***B)*** *Genes most intolerant to LoF had higher GC content than LoF tolerant genes (Wilcoxon $P<1x10^{-15}$). Average GC content for all genes is 63.9% and is shown by a dotted line.*

4.3 Discussion

These findings show that genes sensitive to LoF have significantly different 5'UTRs; they are longer, more conserved, have greater transcript isoform diversity, a higher propensity to be structured, and contain more uORFs, than genes that are tolerant to LoF. The increase in length and complexity of the 5'UTRs of LoF sensitive genes points to the importance of post-transcriptional/translational regulation in controlling the levels of encoded proteins.

There are some seemingly conflicting findings in this chapter; it is unclear if 5'UTRs in LoF intolerant genes are optimised for efficient ribosomal scanning and translation of the CDS. LoF intolerant genes contain more uORFs and the uORFs are further away from the CDS implying they may be less repressive but these genes 5'UTRs are more structured which can impede ribosomes. One might expect the 5'UTRs of these LoF intolerant genes to be less complex to allow for efficient translation of the CDS. If they are genes that are essential to life and required, why would they have longer, more complex 5'UTRs which would reduce or slow gene expression and therefore be inefficient? I would expect genes essential to cell function ("cell essential genes") to have short, uncomplex 5'UTRs to allow for efficient gene expression of proteins needed for basic cell function. Although, if they are essential to cell function, I would assume that they would also be LoF intolerant; as reducing or removing essential genes required for basic cell survival would surely be harmful. However, this might point to them being recessive. Indeed, the flagship gnomAD paper [147] plots

autosomal recessive genes across LOEUF deciles and they are not concentrated in the low LOEUF deciles (Figure 4.19A). However, cell essential genes do occur more frequently in the low LOEUF deciles (Figure 4.19B, 6.6% vs 1%, Chi-square $P<1x10^{-15}$).

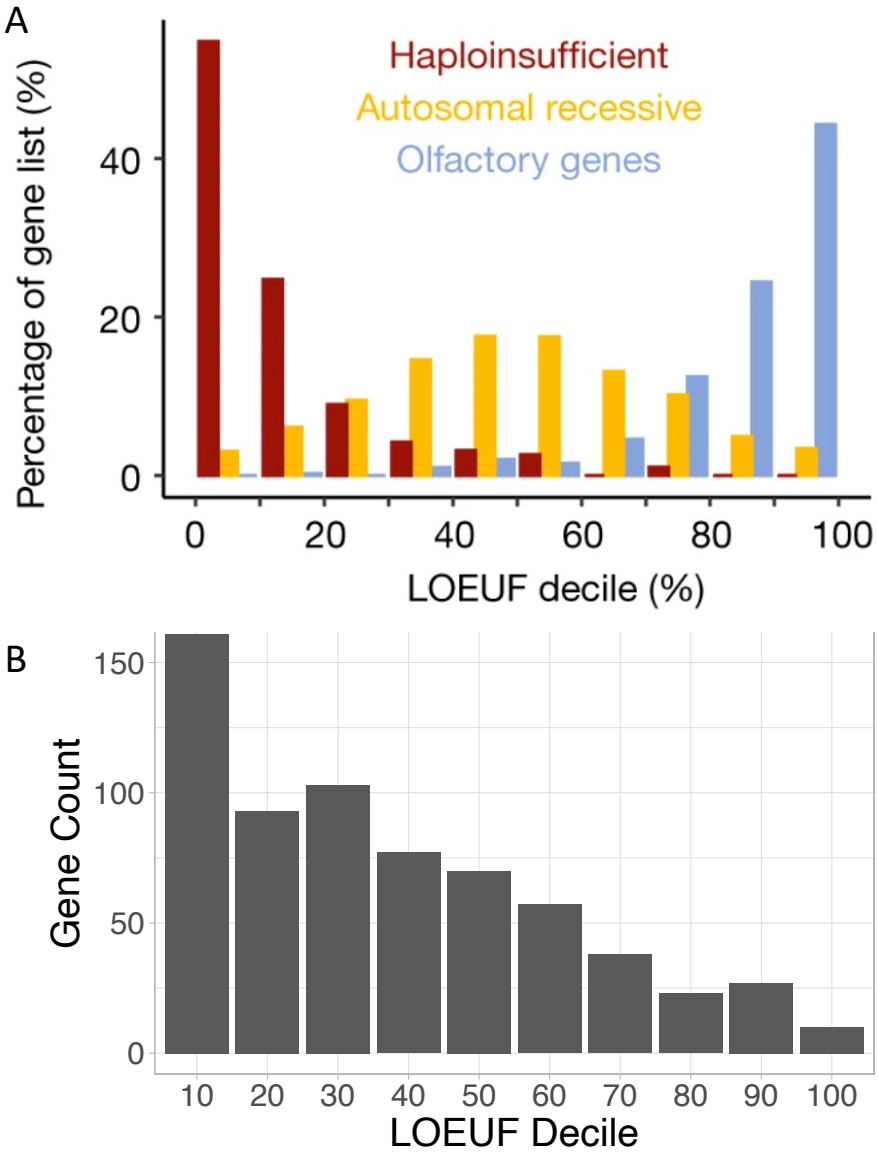


Figure 4.19: Autosomal recessive and cell essential genes across LOEUF deciles
A) *Autosomal recessive genes are not concentrated in the low LOEUF deciles. Image obtained from (147) Creative Commons Attribution 4.0 International License.* ***B)*** *Cell essential genes occur more frequently in genes most intolerant to LoF (6.6% vs 1%, Chi-square* $P<1x10^{-15}$*).*

To begin investigating cell essential genes 5'UTRs, I used a curated list of cell essential genes used in the gnomAD paper [147]. Comparing 5'UTRs of cell essential genes to all genes 5'UTRs, cell essential genes 5'UTRs are significantly shorter overall (mean length 115 bp vs 205 bp; Wilcoxon $P<1x10^{-15}$), indicating these genes are probably more efficiently translated, with less chance of uAUGs occurring. This is only preliminary work; there is probably much to uncover here. However, I hypothesise that LoF intolerant genes contain a range of genes that go from requiring high levels of gene expression which serve essential cell function, with shorter, less complex 5'UTRs, to genes which tend to be highly specialised; only expressed in certain times or places, requiring very specific amounts of protein to be produced (which might be controlled by uORFs) and other mechanisms to slow ribosomes to allow for proteins to fold correctly. The work in this chapter analysed broad trends in 5'UTRs, but also demonstrated that there is a range within each LOEUF decile within each category; this may point at gene function as well as tolerance to LoF as influencing 5'UTR composition. I believe that these findings, along with the range of variability observed in 5'UTRs, suggests that gene regulation mediated by 5'UTRs is complex, nuanced and dependent upon each individual gene. Additionally, some 5'UTR features may not be elucidated yet or may have multiple functions. Features like increased structure and uORFs may be a by-product that the gene has evolved to tolerate. Keeping a broad and holistic view of 5'UTR features and function is important.

This work aimed to provide a general picture of the variation in 5'UTR complexity across the LOEUF spectrum, but it has several limitations. I only analysed a single transcript per gene; I used the highly curated MANE Select transcript set, which likely reflects the most clinically relevant transcript per gene. I acknowledge that there are other relevant transcripts that I have not included. To mitigate not accounting for complexity at the level of alternative 5'UTR isoforms I used CAGE data to determine the number of TSS's per gene, however, this only assesses differences in TSS usage and not alternative splicing within 5'UTRs derived from the same TSS. With emerging long read RNA sequencing data it would be interesting to investigate 5'UTR alternative splicing across different tissues. I would also be curious to see the potential of 5'UTRs preferentially splicing in or out exons with translated upstream start codons; this could give insight into gene expression control, as well as determine whether alternative splicing is more or less common between LOEUF deciles. Additionally, the work using MFE as a proxy of secondary structures is difficult to draw a conclusion from as MFE is tightly linked to 5'UTR length; the longer a sequence is the more likely for adjacent base pairing to occur. As shown, LEOUF intolerant genes have longer 5'UTRs so it follows that they would also have lower MFE values. However, these 5'UTRs also had higher GC content which is an indication of propensity for secondary structures to form, so I think there may be a signal here; regardless of length these 5'UTRs may have a true propensity to form secondary structures. Repeating the MFE analysis with explicit linear modelling or restricting to only 5'UTRs of a specific length e.g. 200 bp may help remove some of

the inherent relationship between MFE and 5'UTR length, and detect if there is a true difference in MFE scores of 5'UTRs across LOEUF deciles.

Here I have focussed on uORFs as *cis* regulators of translation, i.e. uORFs affecting translation of the genes they immediately precede. However, there is evidence from mass spectrometry that some uORFs encode a detectable peptide product (SEPs; smORF encoded peptides) [34,173]. Some SEPs may have a biological function [173]. Further work needs to be done to find and curate these and to understand their role. But it is important to take into account that uORFs may be producing functional peptides, adding more complexity and which may make us redefine the uORF, and indeed the 5'UTR as a concept.

I have assessed 5'UTRs and observed lots of variability of features within them but would like to explore whether this is also true of 3'UTRs. For example, LoF intolerant genes that have short 5'UTRs, do they have longer and more complex 3'UTRs to compensate?

In conclusion, this work showed how LoF intolerant genes have a distinct 5'UTR profile, but no doubt there is much more to uncover that will teach us more about 5'UTR features and function. 5'UTR variability may limit attempts to use the 5'UTR

features to predict gene dosage sensitivity and points to a much more complex regulatory landscape.

4.4 Methods

4.4.1 Categorising 5'UTRs into deciles of LoF tolerance

LOEUF scores were downloaded from gnomAD (v2.1.1). I filtered to Ensembl canonical transcript and where genes had multiple LOEUF scores I kept the transcript with the higher score. They were then binned into deciles. I matched each gene to the MANE set based on Ensembl stable gene id's (17,337 genes). LOEUF has been calculated for all transcripts in GRCh37. MANE is on GRCh38 and hence many MANE transcripts are not represented in the LOEUF file. To get around this, I took a single LOEUF score per gene, taking the value assigned to the Ensembl canonical transcript for every gene. I used the same UTR definition as given in Chapter 3, 3.4.1.

4.4.2 Accounting of differences in gene expression

I used gene expression data from the GTEx project RNAseq based analysis file called "Median gene-level TPM by tissue" (23 October 2023) [149]. I matched each gene to the MANE set based on Ensembl stable gene id's. I took a mean gene expression per gene across all tissues (measured in TPM - transcript per million). I split the data into 4 quartiles (Q1-Q4), ranging from low expression to high expression.

4.4.3 Codon optimality

Some amino acid codons are translated more efficiently by the ribosome and therefore codons can range from being optimal to non-optimal. This distinction is thought to be driven by the balance of tRNA availability versus codon demand [169]. To define this, codon-specific tRNA adaptive indices (tAIs) can be calculated, which is a measure of codon optimality. tAIs are calculated based on the abundance of intracellular tRNA and binding strength between a codon and tRNA. I used tAI scores determined in HeLa cells by Forrest *et al.* [170], to determine overall codon optimality of uORFs. This data contains each codon scored as either “optimal” or “not-optimal”.

Each codon in a Ribo-seq uORF was translated into whether it was optimal (noted as 1) or not (noted as 0). Adding these numeric codons, I then divided by the total number of codons for each uORF to get a total optimality score; with higher scores being more optimal.

4.4.4 Assessing transcription start site (TSS) diversity

CAGE data [150] was downloaded from FANTOM5 “CAGE peak based annotation table of robust CAGE peaks for human samples” (30 November 2022). I used CAGE peaks which uniquely associate to a gene. CAGE data only included HGNC id’s so these were used to match with MANE genes.

4.4.5 Cell Essential Gene Set

Cell essential gene set (“CEGv2_subset_universe.tsv”) was downloaded (1st July 2021) from https://github.com/macarthur-lab/gene_lists/tree/master/lists. The list was derived from CRISPR/SpCas9 studies by Hart *et al.* [151] and was used in the gnomAD flagship paper [147].

4.4.6 Statistics

For all statistical tests in this chapter, I compared the lowest and highest LOEUF quintiles. For Figure 4.5, I compared genes having 0, or 1 or more uORFs/oORFs/start-stops between the lowest and highest LOEUF quintiles.

Chapter 5

Determining how 5'UTRs differ across disease genes

5.1 Introduction

I have analysed 5'UTRs features and have observed patterns and trends in these regions across a spectrum of genes tolerance to LoF, but as shown, there is considerable variety within each category. Gene role and function varies and this may be reflected in 5'UTR differences.

I wanted to extend these findings and previous work that has shown that 5'UTRs of genes where heterozygous loss-of-function (LoF) variants cause developmental disorder (DD) are longer and have more introns than all other genes [86]. I was interested to see whether the distinct 5'UTR features observed in LoF intolerant genes persist across gene sets associated with disease. I investigated 5'UTR features of three disease gene-sets and compared them to the average across all genes. I looked at genes within which predicted LoF variants have been reported to cause DD and cancer, as well as a wider set of dosage sensitive (DS) genes. For DD genes, I further curated them by filtering them to dominant or recessive genes and comparing 5'UTR features between these groups, given the former are more likely to be highly dosage sensitive. I analysed oncogenes (Onc) and tumour suppressor

genes (TSGs) separately due to their inherent differences in how they impact cell proliferation. Finally, for DS genes I compared haploinsufficient (HS) and triplosensitive (TS) genes.

5.2 Results

5.2.1 Disease genes have longer 5'UTRs

I first looked at 5'UTR length across the gene sets. Whilst 5'UTRs average 202 bp in length, disease gene 5'UTRs are significantly longer, except DD recessive genes which are significantly shorter (Figure 5.1; All Wilcoxon; DD dominant: 369 bp, $P<1x10^{-15}$; DD recessive: 169 bp, $P=2.7x10^{-08}$; Onc: 260 bp, $P=1.5x10^{-05}$; TSG: 254 bp, $P=2.9x10^{-04}$; HS: 279 bp, $P<1x10^{-15}$; TS: 253 bp, $P<1x10^{-15}$).

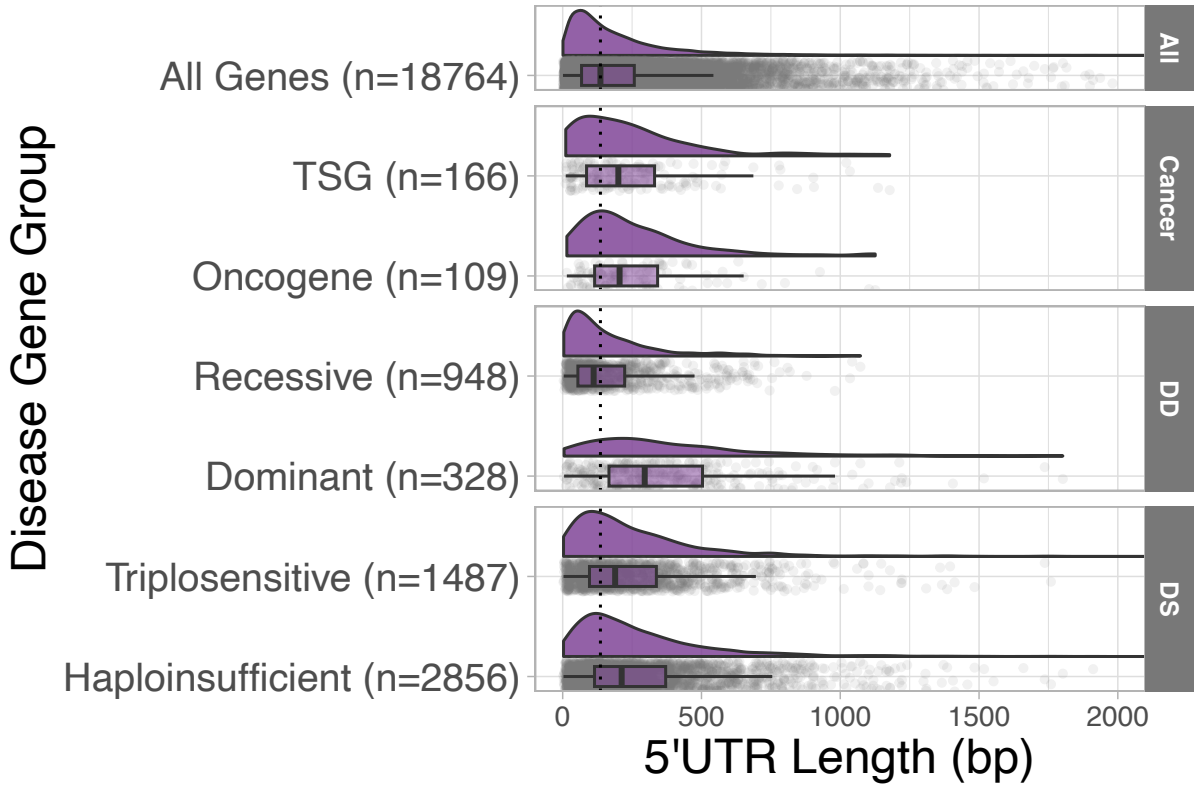


Figure 5.1: *The 5'UTRs of disease genes are significantly longer (Wilcoxon: DD dominant: $P<1x10^{-15}$; Onc: $P=1.5x10^{-05}$; TSG:$P=2.9x10^{-04}$; HS: $P<1x10^{-15}$; TS: $P<1x10^{-15}$) with the exception of DD recessive genes which are significantly shorter (Wilcoxon $P=2.7x10^{-08}$), when compared to the average across all genes. The median 5'UTR length for all genes (136 bp) is shown by the dotted black line. The x-axis was truncated at 2,000 bp (22 genes had 5'UTRs >2,000 bp).*

		5'UTR Length			
Gene Group	**Sub-group**	**mean**	**median**	**range**	**sd**
All Genes	All Genes (n=18764)	201.61	136	1-3561	217.82
Cancer	Oncogene (n=109)	259.45	205	15-1126	211.02
Cancer	TSG (n=166)	254.02	200.5	11-1178	230.68
Developmental Disorder	Dominant (n=328)	368.97	295.5	5-1802	295.91
Developmental Disorder	Recessive (n=948)	168.68	111.5	4-1073	167.08
Dosage Sensitive	Haploinsufficient (n=2856)	278.51	212.5	3-2253	242.36
Dosage Sensitive	Triplosensitive (n=1487)	252.69	189	3-2252	229.04

Table 5.1: *Disease genes 5'UTR lengths, data for Figure 5.1.*

5.2.2 5'UTRs of disease genes contain more uORFs

Assessing occurrence of uAUG type in different disease gene sets, a significantly higher number of disease gene 5'UTRs contain uORFs than the average of 34.4% across all genes (Figure 5.2; All Chi-square; DD dominant: 57.9%, $P<1x10^{-15}$; TSG=49.4%, $P=6.5x10^{-05}$; HS=45.7%, $P<1x10^{-15}$; TS=40.7%, $P=1.4x10^{-07}$), although the difference is not significant for the oncogene cancer gene set (43.1%, Chi-square $P=0.07$). Start-stop elements are only significantly enriched in HS genes (7.1% vs 5.0%; Chi-square $P=3.1x10^{-08}$), however, given the small number of genes that contain start-stops, it is likely underpowered to detect a significant enrichment in these smaller gene sets. In contrast, DD recessive genes have significantly fewer uORFs and start-stops (Figure 2; Chi-square $P=2.7x10^{-06}$ (uORFs); Chi-square $P=1.3x10^{-04}$ (start-stops)).

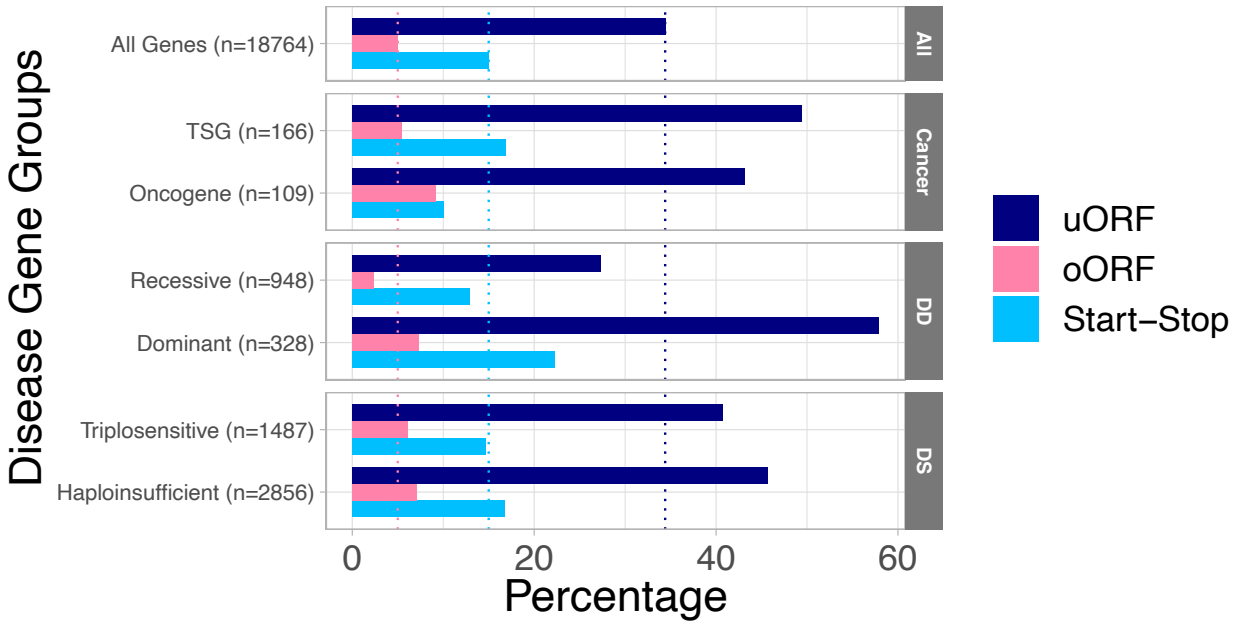


Figure 5.2: *Disease genes significantly more often contain uORFs (Chi-square: DD dominant: 57.9%, $P<1x10^{-15}$; TSG=49.4%, $P=6.5x10^{-05}$; HS=45.7%, $P<1x10^{-15}$; TS=40.7%, $P=1.4x10^{-07}$), when compared to all 5'UTRs. Start-stops are only significantly enriched in HS genes ($P=3.1x10^{-08}$). DD recessive genes have significantly fewer uORFs and start-stops (Chi-square $P=2.7x10^{-06}$ (uORFs); Chi-square $P=1.3x10^{-04}$ (start-stops)). The dotted lines mark the percentage of all genes set with each uAUG type.*

		oORF			uORF			Start-Stop		
Gene Group	**Sub-group**	**count**	**total**	**%**	**count**	**total**	**%**	**count**	**total**	**%**
All Genes	All Genes (n=18764)	2806	18764	14.95	6461	18764	34.43	940	18764	5.01
Cancer	Oncogene (n=109)	11	109	10.09	47	109	43.12	10	109	9.17
Cancer	TSG (n=166)	28	166	16.87	82	166	49.40	9	166	5.42
Developmental Disorder	Dominant (n=328)	73	328	22.26	190	328	57.93	24	328	7.32
Developmental Disorder	Recessive (n=948)	123	948	12.97	259	948	27.32	22	948	2.32
Dosage Sensitive	Haploinsufficient (n=2856)	478	2856	16.74	1306	2856	45.73	203	2856	7.11
Dosage Sensitive	Triplosensitive (n=1487)	219	1487	14.73	605	1487	40.69	91	1487	6.12

Table 5.2: *uAUG type occurrence in disease gene sets. Data for Figure 5.2.*

5.2.3 5’UTRs of disease genes do not contain more introns

To see if there is increased alternative splicing possibilities of disease gene 5’UTRs, I assessed intron occurrence. I tested whether each of the disease gene groups 5’UTRs contained zero or one or more introns compared to all genes. There is no significant difference in the number of 5’UTR introns between disease gene sets and all genes (Figure 5.3; all Chi-square; DD dominant: P=0.07; Onc: P=0.09, TSG: P=0.15; HS: P=0.18; TS P=0.39). This holds true except for DD recessive genes which show a marked difference and have proportionally significantly fewer 5’UTR introns (Figure 5.3; Chi-square P=4.7×10^{-06}).

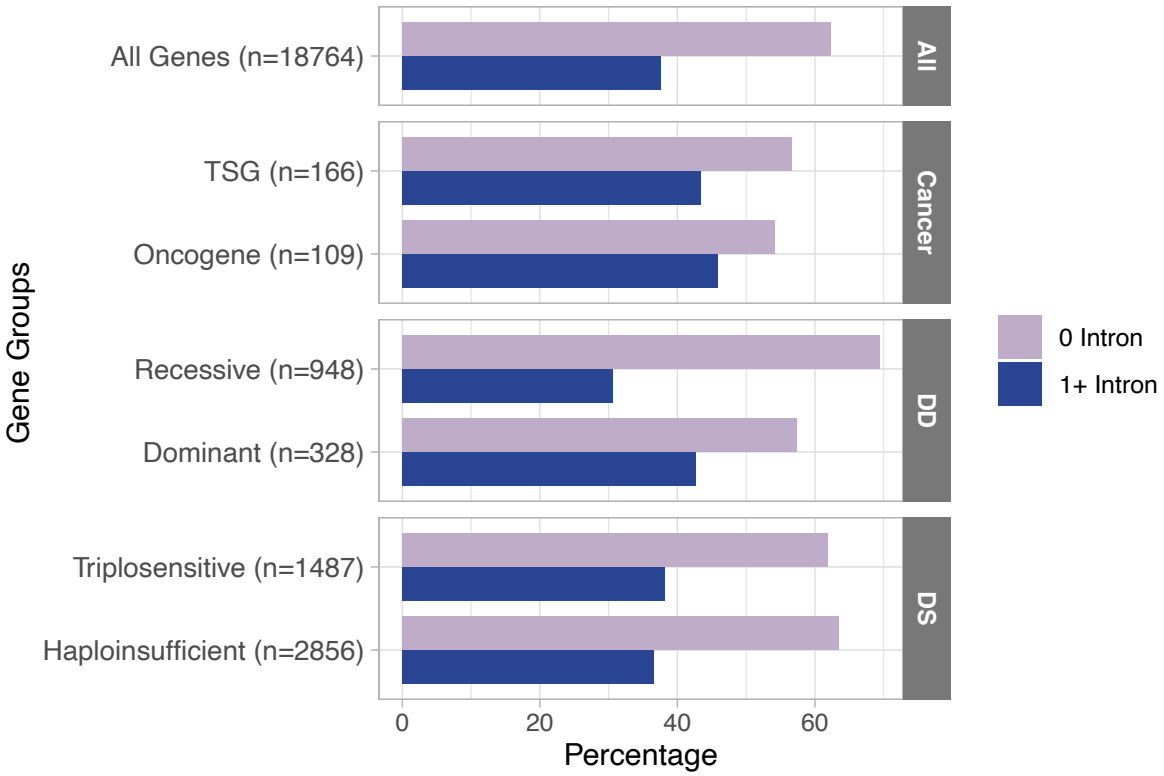


Figure 5.3: *There is no observable difference in the proportion of 5UTRs that have introns between disease gene sets and all genes (DD dominant: Chi-square P=0.07; Onc Chi-square P=0.09; TSG Chi-square P=0.15; HS Chi-square P=0.18; TS Chi-square P=0.39), except for DD recessive genes, where significantly fewer had introns (Chi-square P=4.7x10^{-06}).*

Gene Group	Sub-group	intron	count	total	%
All	All Genes (n=18764)	0 Intron	11698	18764	62.34
All	All Genes (n=18764)	1+ Intron	7066	18764	37.66
DD	Dominant (n=328)	0 Intron	188	328	57.32
DD	Dominant (n=328)	1+ Intron	140	328	42.68
DS	Haploinsufficient (n=2856)	0 Intron	1813	2856	63.48
DS	Haploinsufficient (n=2856)	1+ Intron	1043	2856	36.52
Cancer	Oncogene (n=109)	0 Intron	59	109	54.13
Cancer	Oncogene (n=109)	1+ Intron	50	109	45.87
DD	Recessive (n=948)	0 Intron	658	948	69.41
DD	Recessive (n=948)	1+ Intron	290	948	30.59
DS	Triplosensitive (n=1487)	0 Intron	920	1487	61.87
DS	Triplosensitive (n=1487)	1+ Intron	567	1487	38.13
Cancer	TSG (n=166)	0 Intron	94	166	56.63
Cancer	TSG (n=166)	1+ Intron	72	166	43.37

Table 5.3: *Disease gene 5'UTR intron occurrence. Data for Figure 5.3.*

5.2.4 Disease gene 5'UTRs are more highly conserved

Disease gene 5'UTRs are also significantly more conserved with higher PhyloP scores when compared to all genes (Figure 5.4; all T-test; DD dominant: $P<1x10^{-15}$; Onc: $P=9x10^{-06}$; TSG: $P=4.3x10^{-08}$; HS: $P<1x10^{-15}$; TS: $P<1x10^{-15}$). Again, contrastingly, DD recessive genes are significantly less conserved (Figure 5.4; PhyloP, T-test $P=4.9x10^{-08}$).

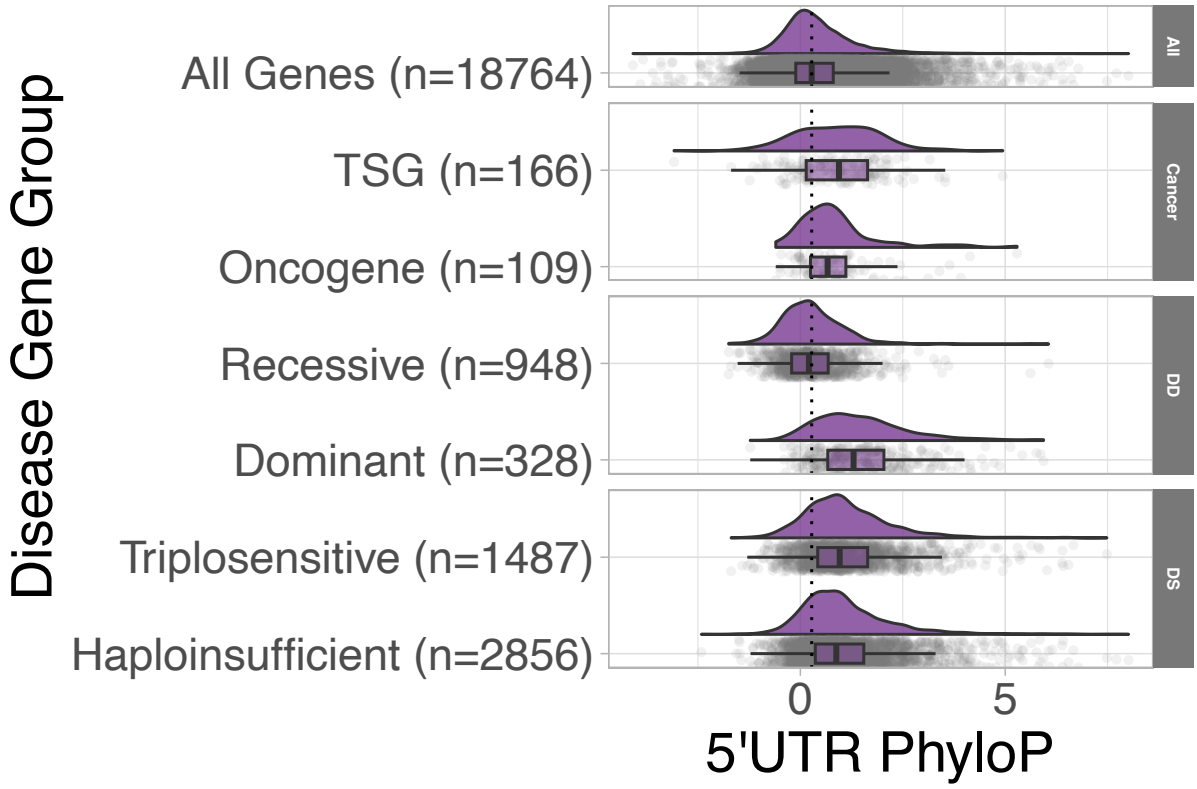


Figure 5.4: *Disease gene 5'UTRs are significantly more conserved (T-test: DD dominant: $P<1x10^{-15}$; Onc: $P=9x10^{-06}$; TSG: $P=4.3x10^{-08}$; HS: $P<1x10^{-15}$; TS: $P<1x10^{-15}$) except DD recessive genes which are significantly less conserved (T-test: $P=4.9x10^{-08}$), compared to all genes. The dotted black line is the median PhyloP score for all genes (0.28).*

		PhyloP				
Gene Group	**Sub-group**	**mean**	**median**	**sd**	**min**	**max**
All Genes	All Genes (n=18764)	0.418	0.277	0.869	-4.086	8.012
Cancer	Oncogene (n=109)	0.855	0.666	0.983	-0.600	5.290
Cancer	TSG (n=166)	0.895	0.943	1.077	-3.083	4.939
Developmental Disorder	Dominant (n=328)	1.487	1.299	1.164	-1.224	5.936
Developmental Disorder	Recessive (n=948)	0.284	0.209	0.764	-1.750	6.060
Dosage Sensitive	Haploinsufficient (n=2856)	1.073	0.877	1.093	-2.417	8.012
Dosage Sensitive	Triplosensitive (n=1487)	1.133	0.960	1.085	-1.691	7.483

Table 5.4: *Disease gene 5'UTR PhyloP scores. Data for Figure 5.4.*

5.2.5 DD Recessive genes 5'UTRs show a marked difference

I observed a significant difference in the 5'UTRs of DD recessive genes. When compared to the average across all genes, the 5'UTRs of DD recessive genes were significantly shorter (Figure 5.1; mean=169 bp, Wilcoxon P=2.7x10^{-08}), have fewer uORFs and start-stops (Figure 5.2; Chi-square P=2.7x10^{-06} (uORFs); Chi-square P=1.3x10^{-04} (start-stops)), are significantly less conserved (Figure 5.3; PhyloP, T-test P=4.9x10^{-08}) and also have significantly fewer 5'UTR introns (Figure 5.4; Chi-square P=4.7x10^{-06}). The lower complexity of the 5'UTRs of recessive genes likely reflects their insensitivity to changes in dosage; loss of one allele does not cause disease and they may not require tight gene regulation. The observation that these 5'UTRs are significantly different to the all gene set average suggests that the all gene set may contain many genes that are sensitive to dosage changes. To account for this, I tested DD recessive genes against genes in the middle two LOEUF deciles; these middle LOEUF genes are not extremely LoF tolerant or intolerant. There was no significant difference in 5'UTR length (mean 169 vs 177 bp, Wilcoxon P=0.04), the

number of uORFs (Chi-square *P*=0.66), or mean PhyloP scores (T-test *P*=0.1). There are still significantly fewer introns in DD recessive genes (Chi-square $P=8.2x10^{-05}$). This points towards the fact that indeed, DD recessive genes 5'UTRs are not significantly different from an all gene set. However, they are markedly different to the other disease gene sets in this work; they do not have longer and more complex 5'UTRs. In conclusion, recessive DD genes have 5'UTRs reflective of genes that are tolerant to changes in dosage and are in stark contrast to dominant DD genes.

5.3 Discussion

These disease genes appear to have similar 5'UTR profiles to the genes intolerant to LoF, as demonstrated in the previous chapter (Chapter 4). The increase in length and complexity of the 5'UTRs of these disease gene groups, points to the importance of post-transcriptional/translational regulation in controlling the levels of encoded proteins. As discussed, the variety between 5'UTRs of LoF intolerant genes and disease genes likely reflects the range of gene expression requirements of these genes. Some genes are expressed only during development and in precise quantities and some are expressed often throughout a lifetime. Whilst the 5'UTRs of LoF intolerant genes tend to be much longer than average, some LoF intolerant and known dosage sensitive genes have very short 5'UTRs. For example, the 5'UTR of *FOXF1*, a haploinsufficient developmental disorder (DD) gene which is in the 2nd LOEUF decile (intolerant to LoF), is only 43 bp long. LoF variants in *FOXF1* are a

known cause of alveolar capillary dysplasia with misalignment of pulmonary veins [174]. This is a lethal disorder and almost all infants with this condition die within the first month of life [175].

I observed a stark difference between DD dominant and DD recessive genes. Recessive genes are not sensitive to changes in dosage, have shorter 5'UTRs with less complexity. Shorter 5'UTRs may point to gene function; their protein products may be required at a higher rate. The lower complexity of the 5'UTRs of recessive genes likely reflects their insensitivity to changes in dosage; loss of one allele does not cause disease and they may not require tight gene regulation. However, this could actually reflect a protective mechanism; loss of one allele does not cause disease as there is a backup allele to make up for it. Interestingly, there are some genes that can act via both a dominant and recessive mode of inheritance (18 in this dataset). An example of one is *ITPR1* which causes Spinocerebellar Ataxia Type 15 when dominantly inherited or Gillespie syndrome when inherited in a recessive manner, both via a LoF mechanism. The 5'UTR is 350 bp, and it is in the 10th decile of LOEUF scores; in the group of genes most tolerant to LoF. This is another example of how there is a variety of gene roles and kind of protein it encodes, and therefore possibly why 5'UTRs vary quite drastically.

I observe increased length and complexity across both haploinsufficient and triplosensitive gene sets, although I acknowledge that there is considerable overlap between these sets.

Generally, disease genes and LoF intolerant genes that I have assessed appear to have distinctly more complex 5'UTRs. To further unpick the 5'UTR variety observed, I would be curious by the prospect of stratifying gene groups based on 5'UTR features in conjunction with other genomic characteristics to discern potential distinct modes of action. For example, considering a more complex 5'UTR (longer, more structured and more uORFs) together with 3'UTR length and gene expression levels could give insight into gene roles and functions. It would also be interesting to further explore disease gene 5'UTRs across genes that have incomplete penetrance and variable expressivity and see if they contain a distinct 5'UTR which plays a part in these interesting genetic phenomena. Exploring genomic combinations, including 5'UTRs, has the potential to enrich understanding of gene behaviour and function.

5.4 Methods

5.4.1 Filtering disease-gene sets

The different gene sets used are detailed in section 2.10.

DD genes were downloaded (18 February 2021) from the Development Disorder Genotype-Phenotype Database (DDG2P). DD genes were restricted to genes with ‘confirmed’ or ‘probable’ roles in developmental disorders (i.e. removing any genes with limited evidence of disease association) and that are reported to cause disease via a loss-of-function disease mechanism. They were split into whether they were biallelic (recessive) or monoallelic (dominant). There were 18 genes that were recorded as both biallelic and monoallelic.

The COSMIC Cancer Gene Census (extracted 22 February 2021) was filtered to only include genes where nonsense, frameshift and missense mutation types are known to be involved in cancer (i.e. removing genes only associated with large structural changes) and then further filtered to oncogene or TSG only as cancer gene type.

5.4.2 Statistical tests

For all statistical tests I compared the disease gene group against all MANE Select 5’UTRs with that specific disease group removed.

Chapter 6

Investigating start-stop variants in rare disease

6.1 Introduction

Start-stops are start codons that are immediately followed by a stop codon, with no codons in between. Very little has been mentioned about start-stops in the literature, and they have also been referred to as the “minimum ORF” [38–40]. Work by Rendleman *et al.* has shown that start-stops are thought to cause ribosome stalling and appear to repress downstream protein expression [39]. The small ribosomal subunit recognises and binds to the start codon, thus attracting the large ribosomal subunit to initiate translation. Usually, when the initial methionine-charged tRNA is at the P-site of the ribosome, the next tRNA would enter at the A-site. In the case of a start-stop however, as the next codon is a stop codon, no tRNA enters, leaving the ribosome in an unusual simultaneous initiating and terminating state. Translation is then terminated, and the large ribosomal subunit detaches, but the smaller subunit can remain bound. This creates several mechanisms through which start-stops may alter downstream CDS expression. The small ribosomal subunit can attract another large ribosomal subunit, which can then go through the same process, thus stalling the ribosome, and cause a physical block to other scanning ribosomes, causing protein production to be reduced or produced at a slower rate. Alternatively, the small subunit can also continue traversing until it reaches and recognises the next AUG

codon and recruits another large ribosomal subunit. However, if the start-stop is too close to the CDS there may not be sufficient space for continued traversing and reinitiation at the CDS AUG. There is also the possibility that through impaired termination, the ribosome is released and does not continue traversing, and the CDS is not translated.

Whether start-stops repress protein expression more or less than uORFs or oORFs is not clear - the foundational experimental work showed that start-stops affect protein expression as much, or more than, a uORF in one gene, regardless of its position in the 5'UTR [39], but this has not been repeated by others in a systematic way. Additionally, the uORF used in the experiment was nine nucleotides long; containing only one codon between the start and stop codon. A longer uORF would be more suitable; as shown in Chapter 3, the average uORF with Ribo-seq evidence of translation is 49 bp long.

As shown in Chapter 4, start-stops showed the most marked difference in PhyloP and CADD scores across LOEUF deciles. As a largely uncharacterised feature, I wanted to investigate them further. Additionally, to the best of my knowledge, there are no reported pathogenic examples of start-stop variants causing human disease in the literature. Therefore, I wanted to explore large scale genomic variant databases for start-stop creating or removing variants and assess whether they could be a plausible cause of disease, in a similar manner as to what is seen with uORFs and oORFs.

6.2 Results

6.2.1 Start-stops are a rare 5'UTR feature

Start-stops are a rare event within 5'UTRs; looking in the MANE 5'UTR set to determine occurrence of start-stops, only 5% of genes have at least 1 start-stop within the 5'UTR, with up to a maximum of 8 start-stops in one gene. Of genes with start-stops, 89% only have 1 start-stop within their 5'UTR (Figure 6.1). For this work I only considered start-stops with the canonical AUG start codon.

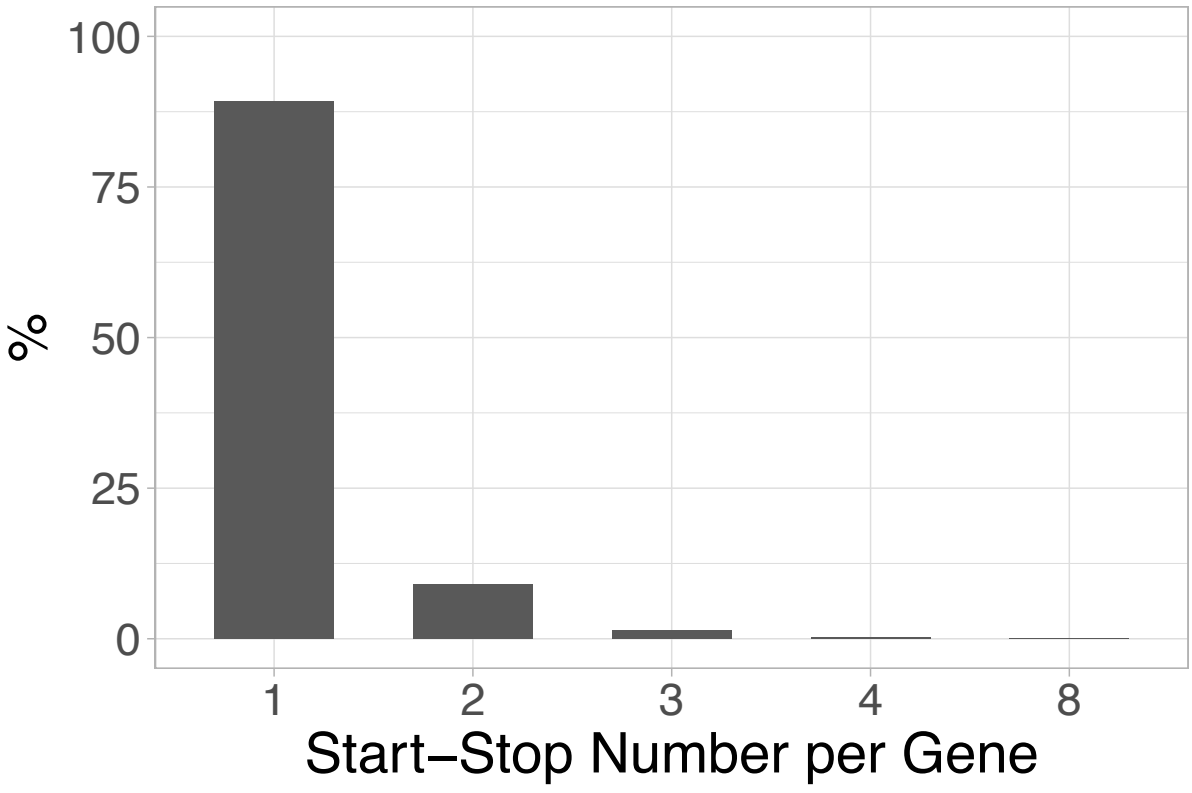


Figure 6.1*: In genes with at least one start-stop, 89% of them have only one start-stop in the 5'UTR.*

6.2.2 Start-stops appear more frequently than expected by chance

To assess whether the occurrence of a stop codon immediately after a uAUG occurs more or less frequently than expected by chance, compared to any other codon, and as a test of whether they may be selected for or against, I shuffled MANE Select 5'UTR sequences 1000 times each whilst preserving di-nucleotide frequency. I looked at the observed (O) occurrence (in MANE 5'UTRs) of any possible codon after an ATG ("hexamers") compared to the expected (E) (in the shuffled sequences) to generate an observed/expected (O/E) score. If the O/E score is equal to one, this hexamer appears as frequently as expected by chance, <1 occurs less frequently than expected and >1 the hexamer appears more frequently than expected by chance. Interestingly, I found that start-stops occur more frequently than expected than by chance, with the three hexamers representing start-stops all amongst the five most enriched hexamers (Figure 6.2). Intriguingly, similar codons TGT and TAT also occur more frequently than expected by chance immediately after an ATG.

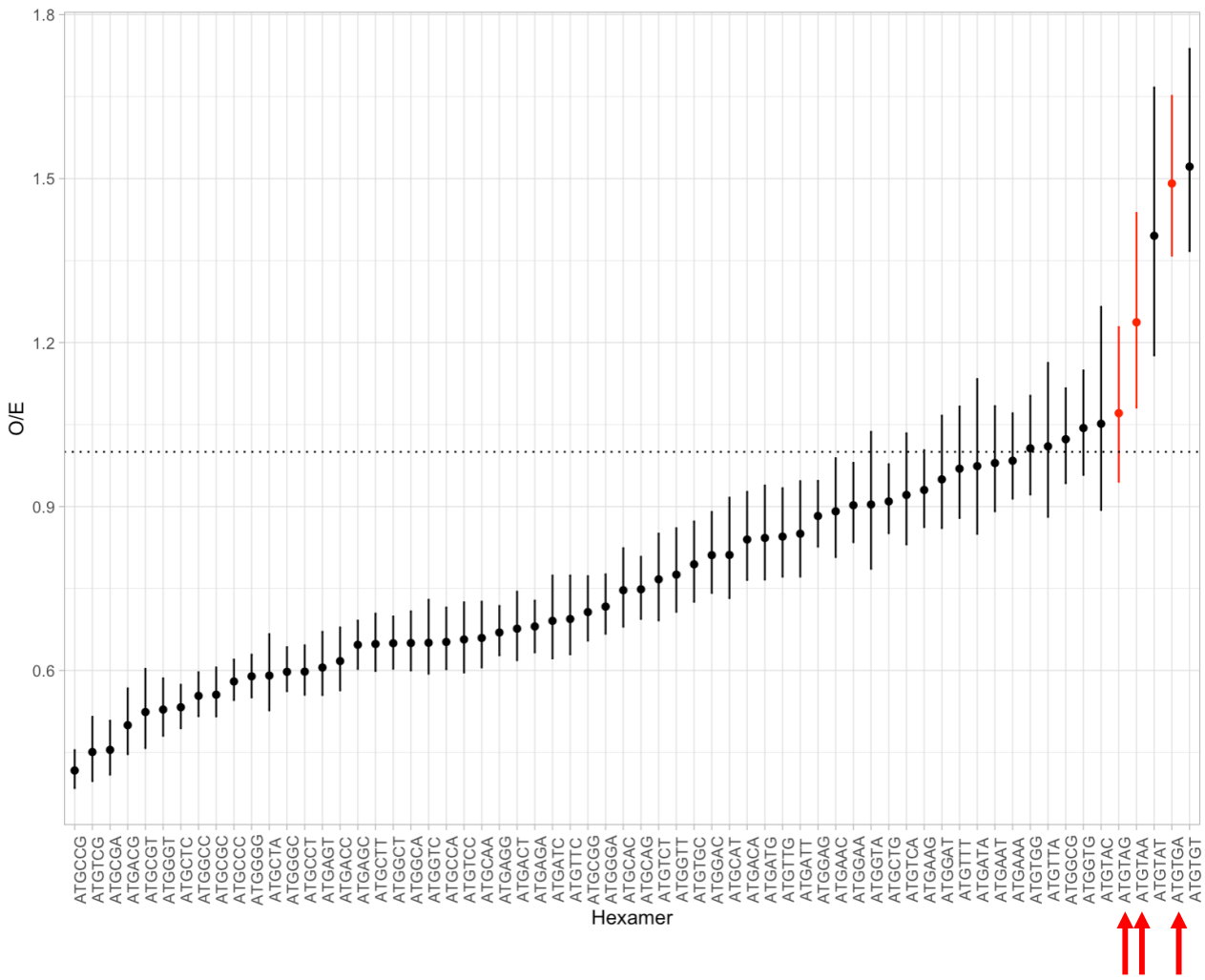


Figure 6.2: Start-stop occurrence
MANE 5'UTR sequences were shuffled 1000 times to generate an observed/expected (O/E) ratio for all ATG hexamers in 5'UTRs. O/E values below 1 indicate that a hexamer appears at a lower frequency than expected by chance; exceeding 1 suggests higher than expected frequency. The dotted line denotes O/E=1. Start-stops appear significantly more frequently than would be expected by chance. The red data points indicate the three types of start-stops. Values plotted with 95% confidence intervals (from the empirical distribution).

As a control, I used the complementary strand of 5'UTRs to repeat this analysis. I wanted to see if start-stops would appear more or less frequently on the "untranslated" strand, than seen on the translated strand. Here I searched 5'UTRs for

the complement sequence and the three bases immediately preceding it. I found that start-stops on the complementary strand also occur more often as would be expected than by chance (Figure 6.3; O/E values; ATGTGA=1.5 vs TCACAT=1.2; ATGTAG=1.1 vs CTACAT=1.2; ATGTAA=1.2 vs TTACAT=1.5).

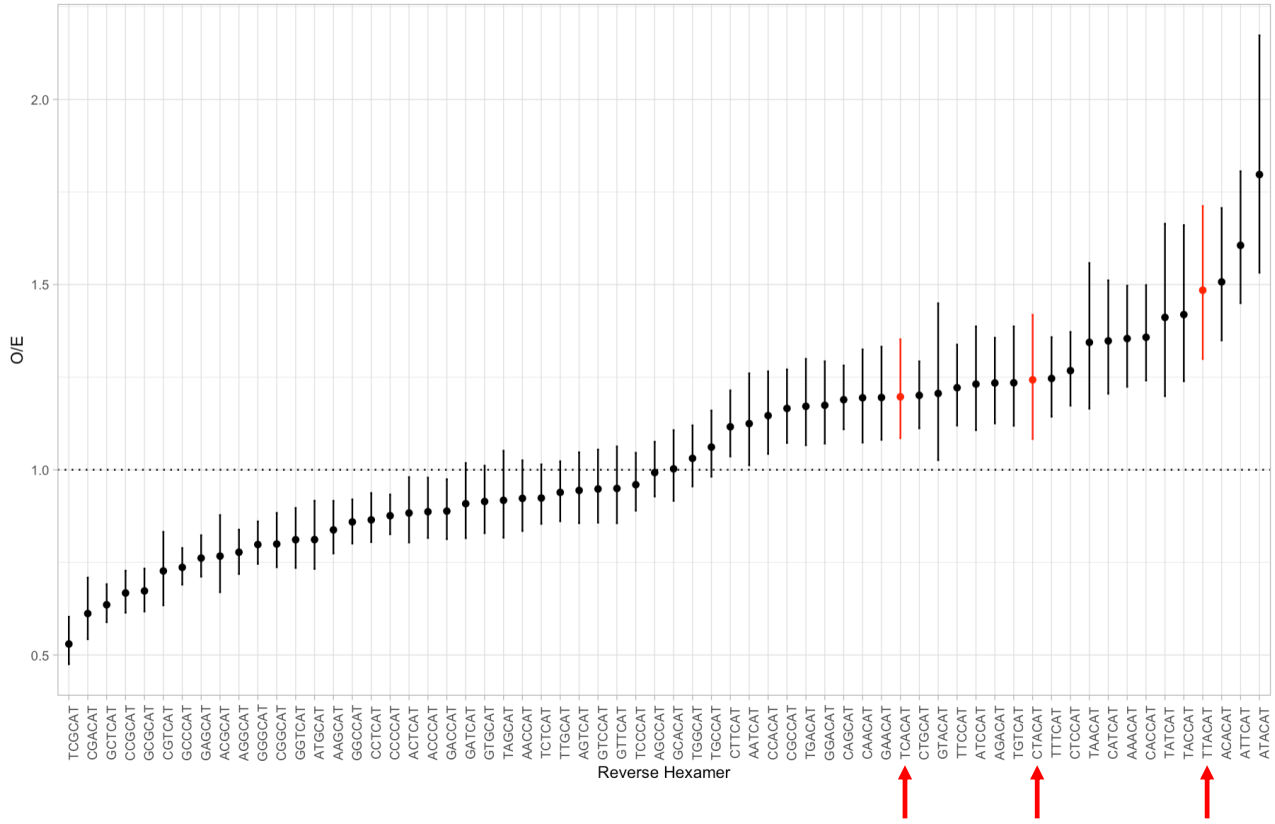


Figure 6.3: **Start-stops on the reverse strand**
Start-stops occur more frequently than expected by chance on the reverse untranslated strand too. The red data points indicate the three types of start-stops. The dotted line denotes O/E=1. Values plotted with 95% confidence intervals (from the empirical distribution).

I then wanted to see whether start-stops occur more or less frequently than by chance in genes intolerant to LoF compared to tolerant to LoF genes. I looked at O/E scores in genes in the lowest 3 LOEUF deciles (most intolerant to LoF) compared to the genes in the highest 3 LOEUF deciles (most tolerant to LoF) and found that start-

stops occurred more frequently than expected by chance only in genes most intolerant to LoF (Figure 6.4). In summary, only genes most intolerant to LoF have more start-stops than would be expected by chance.

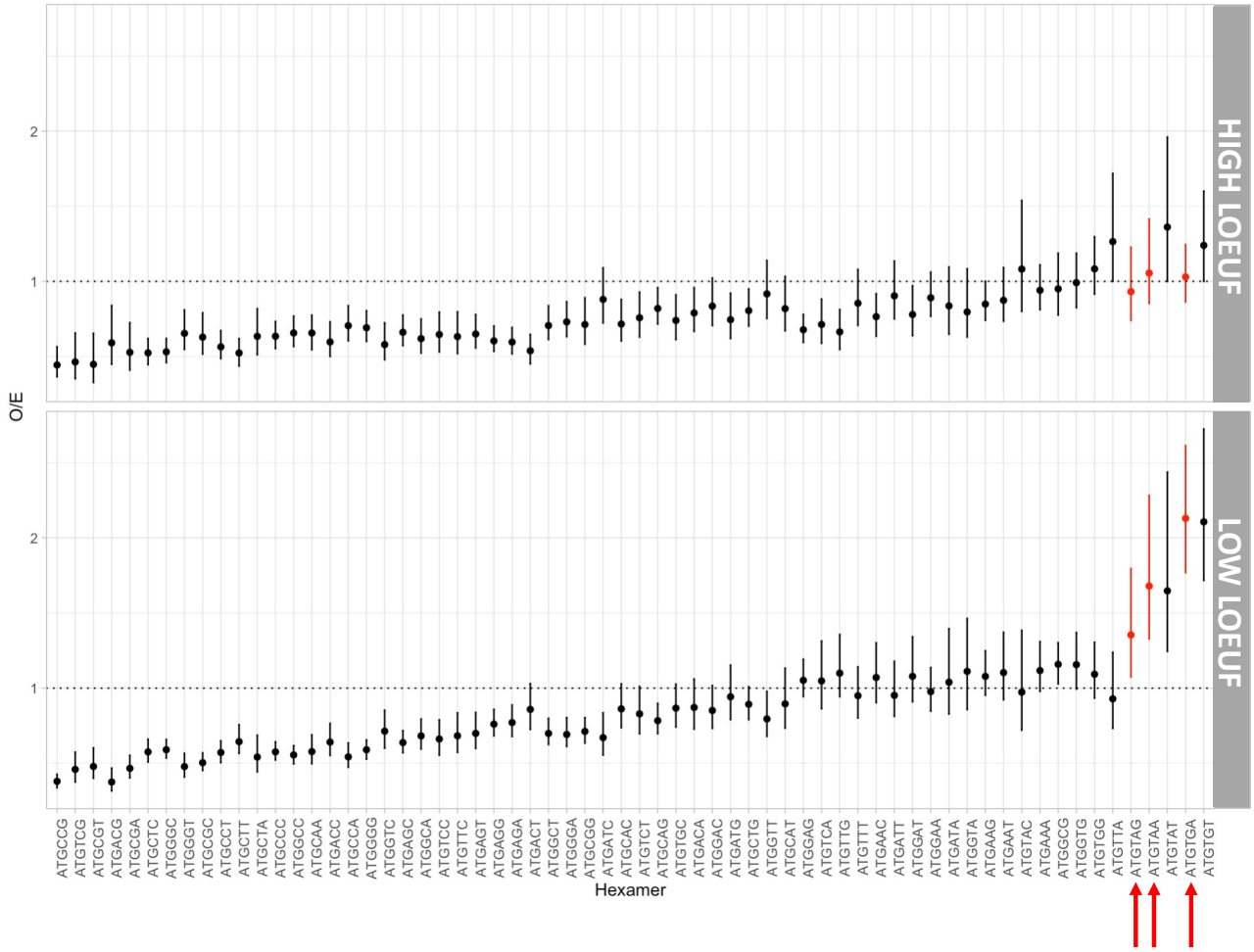


Figure 6.4: **ATG-hexamers in LOEUF deciles**
O/E for ATG-hexamers for low LOEUF deciles (bottom 3 deciles) and high LOEUF deciles (top 3 deciles). Start-stops only occur more frequently than expected by chance in low LOEUF deciles; genes most intolerant to LoF. Red data points indicate start-stops. Values plotted with 95% confidence intervals (from the empirical distribution).

6.2.3 Start-stop variants in rare disease

I wanted to assess whether creating or removing a start-stop could be a cause of disease, in a similar manner to what is seen with uORF/oORF variants, where they reduce or increase CDS expression leading to disease. I analysed two datasets - ClinVar and the Genomics England (GEL) 100,000 genomes project dataset. ClinVar is a public database of curated clinical variants linked to phenotype [156]. GEL is a dataset containing 100,000 sequenced whole genomes split into two arms; participants with rare disease, including where possible sequencing of parents and siblings, and both germline and somatic sequencing for participants with rare or unusual cancer [157,158].

I ran the Ensembl variant effect predictor (VEP) plugin UTRannotator [176] on all 5'UTR variants in ClinVar. This annotates 5'UTR variants with their predicted effect with reference to uAUGs. I found six variants that created or removed a start-stop, shown in Table 6.1. However, three variants had no reported disease association, meaning they were listed in ClinVar without being linked to any specific disease. As I only considered variants associated with disease, these could not be evaluated. The three other variants were as follows. *FANCC*:NM_000136.3:c.-76T>G is a uAUG-creating variant classified for Fanconi anaemia (FA) but classed as likely benign, so not thought to be a cause of disease. It is also a difficult variant to predict the impact of as the variant creates a start-stop where the last base of the start-stop is also the first base of the existing oORF (AU**U**UA**AUG**U → AU**G**UA**A****UG**U) (Figure 6.5A). If the start-stop were to cause binding of a ribosome, it is unlikely that the oORF would also

subsequently be “seen” and translated. This variant could therefore remove the function of the existing oORF and replace it with a start-stop. This would possibly lead to increase of gene expression, depending on whether the start-stop is translated more than the oORF and whether start-stops are more or less repressive than oORFs. However, as both are in theory usable, it is difficult to know which one is more likely to be utilised by a ribosome. Additionally, as the variant is within 6 bp of the oORF start codon, it could influence the oORF start codons consensus sequence and therefore increase or decrease the likelihood of it being translated. To assess the uAUG consensus sequences of the oORF and start-stop, I used previously described Kozak score groups and TE scores, where lower scores have a weaker consensus sequence (Chapter 3.4.3). The existing oORF start codon has a weak Kozak consensus sequence and TE score of 58, and the variant would change this to a moderate Kozak consensus and TE score of 83. The start-stop uAUG codon created by this variant would have a weak Kozak consensus sequence and 85 TE score. Therefore, the variant would cause two uAUGs in close proximity with similar TE scores, with the start-stop having a weak Kozak consensus sequence and the existing oORF Kozak changed to a moderate one. It is important to note that start-stops by definition cannot have a “strong” Kozak consensus sequence as the base immediately following the start codon is a “T”.

The second ClinVar variant, *FANCI*:NM_001113378.2:c.-10A>G is also a uAUG-creating variant linked to FA, however this variant has been classified as a variant of unknown significance (VUS). This variant also replaces an existing out-of-frame

oORF, with Ribo-seq evidence of translation [177–179], with a start-stop (Figure 6.5B). The variant creates a new start codon that terminates at the first position of the start codon of the oORF, so the sequence changes from “AU***A***UGA” to “AU***G***UGA”, which creates a different uAUG, changes the reading frame and therefore removes the oORF. FA is caused by a LoF mechanism. If the start-stop is more repressive than the existing oORF this could be a viable disease mechanism but predicting the impact of this variant is difficult without functional studies. However, it is more likely that the opposite is true; that the native oORF is more repressive. Assessing possible change to the consensus sequence, the existing oORF uAUG has a moderate Kozak consensus sequence and TE score of 70, and the variant uAUG has moderate Kozak and a TE score of 79; both uAUGs having similar consensus strengths.

The third and final ClinVar variant, *BTD*:NM_001370658.1:c.-148C>T is a stop-gain variant. This turns a predicted, short (15 bp), weak Kozak uORF into a weak Kozak start-stop. There are also 2 oORFs downstream in this 5’UTR (one in- and one out-of-frame), Figure 6.5C. The classification of this variant is classed as conflicting - one report of pathogenic and one of VUS. Biotinidase deficiency, the disorder linked to the gene containing this variant, is also a LoF disease. Again, it is difficult to ascertain the impact of this variant on downstream CDS translation without functional studies.

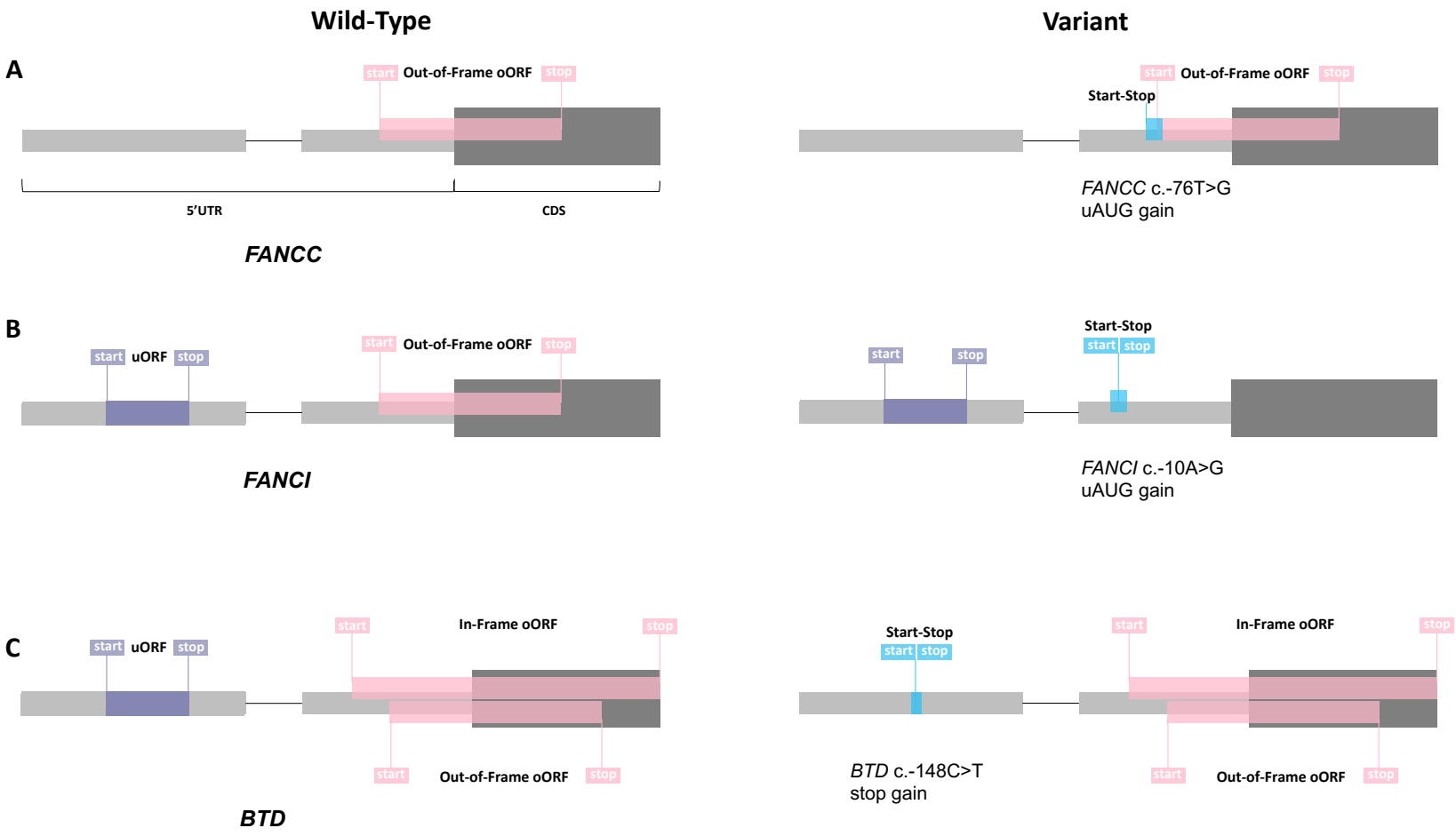


Figure 6.5: ClinVar start-stop creating variants
A) *FANCC 5'UTR contains one out-of-frame oORF with Ribo-seq evidence of translation [177–179]. The FANCC uAUG creating variant c.-76T>G, associated with Fanconi anaemia (FA), creates a start-stop where the last base of the start-stop is also the first base of the existing oORF. This variant is classed in ClinVar as benign.* ***B)*** *The FANCI 5'UTR has one out-of-frame oORF and one uORF with Ribo-seq evidence of translation [177–180]. The FANCI c.-10A>G uAUG-creating variant, associated with FA, creates a uAUG overlapping the existing oORF uAUG, changing its frame and rendering it to a start-stop. This variant is classed in ClinVar as*

a variant of unknown significance (VUS). **C)** *BTD 5'UTR contains a uORF and 2 oORFs. The BTD c.-148C>T variant creates a stop codon next to the existing uORF's start codon, changing it into a start-stop. This variant is associated with Biotinidase deficiency and is classed as conflicting - one report of pathogenic and one of VUS.*

gnomAD [147] is an aggregated database of >700,000 exomes and >70,000 genomes of presumed healthy individuals. Allele frequency (AF) is the frequency of an allele at a particular locus in a population; if it is very high in gnomAD this indicates that this allele cannot cause a rare disease. I assessed the number of individuals in gnomAD with each variant and interpreted it according to the disorder it is associated with [181]. If a phenotype is extremely severe, such as a rare developmental disorder, we would not expect to observe it in gnomAD, whereas a variant linked to a hearing loss disorder we may expect to find at greater numbers in gnomAD. The three ClinVar variants had allele frequencies (AF) of $<2.2x10^{-4}$ in gnomAD (Table 6.1). Four individuals in gnomAD had the *FANCC*:NM_000136.3:c.-76T>G, and 262 individuals had the *FANCI*:NM_001113378.2:c.-10A>G variant, none of whom are homozygous for these variants. FA is usually inherited in an autosomal recessive manner; we can expect to see the variant in the population. However, it is difficult to ascertain in gnomAD if the individuals have a compound heterozygous variant which would likely mean they are affected with FA. There are 535 individuals in gnomAD with the *BTD*:NM_001370658.1:c.-148C>T variant, one of whom is homozygous. There can be severe to mild forms of Biotinidase deficiency and it is also usually inherited in an autosomal recessive manner, so it is difficult to ascertain whether this individual is likely affected and to what severity.

Annotating variants with PhyloP and CADD scores can also be useful in determining whether they are particularly conserved, indicating their essentiality, or predicted to be deleterious. Using the suggested cut-off scores (Phylop = 7.367, CADD = 25.3)

from Pejaver *et al.* [182] for supporting evidence thresholds in evaluating variants pathogenicity as per the American College of Medical Genetics (ACMG) guidelines, none of these variants (and other variants analysed below) scores were higher than these cut-off scores, suggesting that they are not particularly conserved or predicted to be deleterious. It is also important to note that these metrics are calibrated on protein-coding regions of the genome so can be difficult to interpret for non-coding regions.

Variant	ClinVar ID	Mutational class	Pathogenicity rating	Disease	gnomAD allele frequency (AF)/individuals in gnomAD	PhyloP/CADD
FANCC c.-76T>G chr9:95249367A>C	702917	uAUG gain	Likely benign	Fanconi anaemia	$6.6x10^{-6}$/4	0.648/7.354
FANCI c.-10A>G chr15:89247638A>G	887850	uAUG gain	VUS	Fanconi anaemia complementation group I	$1.6x10^{-4}$/262	0.143/4.712
BTD c.-148C>T chr3:15601763C>T	343901	Stop gain	Conflicting interpretations of pathogenicity: Pathogenic(1); VUS(1)	Biotinidase Deficiency	$2.2x10^{-4}$/535	1.36/14.74
TARS2 c.-22G>A chr1:150487429G>A	383907	uAUG loss	Likely benign	n/a	$6.4x10^{-4}$/1001	2.019/11.72
SCN5A c.-37C>T chr3:38633344G>A	516637	uAUG gain	Likely benign	n/a	$2.9x10^{-5}$/45	0.177/10.0
PIGM c.-137A>G chr1:160031876T>C	1223274	Stop loss	Benign	n/a	$1.2x10^{-2}$/11783	-2.194/0.331

Table 6.1: *Six variants in ClinVar are predicted to create or remove a start-stop. Three are not associated with a disease. The allele frequencies of variants and the number of individuals in gnomAD with the variant are noted.*

In summary, the ClinVar start-stop creating variants do not represent convincing causes of disease in these cases given the current evidence. A definitive ruling would only be possible with additional evidence such as segregation or functional studies.

Assessing the GEL dataset, I applied the VEP UTRannotator plugin [176] to the rare disease *de novo* trio probands variants, (35,753 5'UTR variants from 10,797 probands), to look for start-stop creating or removing variants. There were four variants in four participants that create or remove a start stop (Table 6.2). Two of the variants are in participants for whom a causal variant has already been identified in GEL; the start-stop creating variant *SMARCA2:*chr9:2016073C>T in a patient with Osteogenesis imperfecta has been solved with an alternative variant (*COL1A1:*17:50199462:T>C). The start-stop removing variant *NEK10*:chr3:27363744T>C has been solved but no causative variant details provided by GEL.

Gene	Variant	Disease	Variant class	MANE transcript	gnomAD frequency (AF)/individuals in gnomAD	PhyloP/ CADD	Case Status
RNF144A	chr2:6941069C>G	Undiagnosed monogenic disorder – congenital skeletal malformations	uAUG creating	yes	Variant absent from gnomAD	0.372/ 9.7	unsolved
SMARCA2	chr9:2016073C>T	Osteogenesis imperfecta	uAUG creating	yes	$7.4x10^{-3}$/1103	0.347/ 15.81	solved: Alternate variant *COL1A1*:17:50199462:T>C
LRRC53	chr1:74512563T>C	Skeletal muscle channelopathies	uAUG removing	yes	Variant absent in gnomAD	0.123/ 10.18	unsolved
NEK10	chr3:27363744T>C	Undiagnosed metabolic disorders	uAUG removing	no	$6.3x10^{-5}$/6	-0.225/ 7.035	solved: Variant not provided, note - consistent with a diagnosis of combined malonic and methylmalonic aciduria

Table 6.2: *Start-stop creating or removing variants in GEL rare disease arm.*

The two remaining unsolved variants are as follows. A *RNF144A:*chr2:6941069C>G start-stop creating *de novo* variant is only present in the proband and neither parent, however this gene has no links to the participants phenotype and is therefore not assumed pathogenic for this disorder. This gene has a low LOEUF score of 0.35, placing it in the 2nd LOEUF decile (intolerant to LoF), and has several uORFs. This variant would introduce a start-stop 17 bp downstream from a uORF which has Ribo-seq evidence of translation [178,179]. This variant is also absent in gnomAD. Without functional studies it is difficult to know if a ribosome could initiate at this start-stop after the uORF, and whether this variant would have an impact on translation or mRNA structure.

The final variant, *LRRC53*:chr1:74512563T>C removes an existing start-stop. The father of the proband is affected but does not have the variant. Due to the fact that the gene is not known to be associated with disease and this variant does not segregate with the disorder, it is not considered a causal variant. This gene also has an out-of-frame oORF downstream to the start-stop. The LOEUF score for this gene is 1.39, placing it in the 8th LOEUF decile, indicating tolerance to LoF. This variant is absent in gnomAD.

In conclusion, neither of the GEL rare disease start-stop creating or removing variants are thought to be the cause of disease as they either do not segregate and/or the genes are not linked to the disease.

6.2.4 Somatic cancer start-stop variants

To enhance the statistical power of this start-stop analysis, I also assessed the somatic cancer arm of GEL, expanding the pool of variants for analysis. The somatic cancer arm comprises whole genome sequences of cancerous tumours and blood sequencing. The variants I assessed are ones that are novel in the tumour and not present in the germline. I only considered variants in known oncogenes or tumour suppressor genes (TSGs), curated from the COSMIC cancer gene set [154]. Running UTRannotator on the 5'UTR tumour variants I found that there were 37 participants with 30 unique start-stop creating or removing variants; 17 in TSGs and 13 in oncogenes. The majority of cancers in this group were colorectal cancer (16 out of 37 cases).

Next, I split start-stop variants into two categories of mutation class, those most likely to cause a decrease in protein expression (uAUG-gain and uStop-loss) (Figure 6.6), which I hypothesised would be more likely to occur somatically in TSGs where reduced expression would lead to tumour growth, and those more likely to result in increased protein levels (uAUG-loss and uStop-gain), which I hypothesised would be deleterious in oncogenes; as they could cause increased gene expression which would lead to cell proliferation. For this hypothesis, I assume that start-stops are less repressive than uORFs or oORFs as the smaller subunit remains bound with the option of continuing scanning and initiating at a subsequent AUG, whereas uORFs and oORFs have distinct stop codons which generally promote dissociation of the

ribosome. Additionally, oORFs terminate within, or at the end, of the CDS in which case reinitiation would not yield functional peptides.

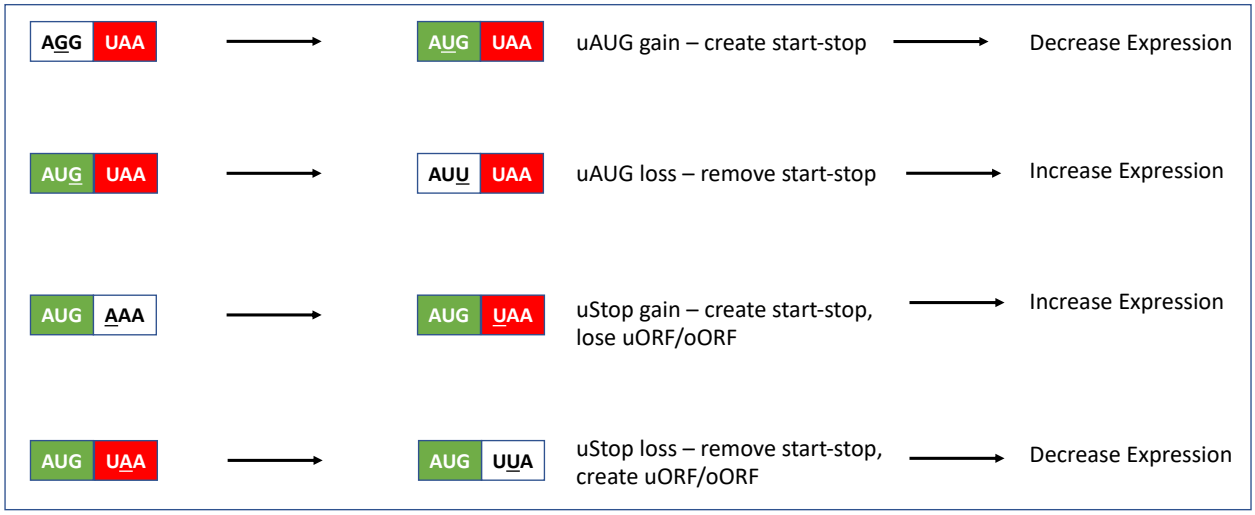


Figure 6.6: *Start-stop mutation classes. The four ways a start-stop variant could create or remove a start-stop. In the uStopgain variant scenario, the uAUG is pre-existing, likely forming a uORF or oORF, so in addition to creating a start-stop, this variant would also alter an existing uORF or oORF. With uStop-loss variants, as the uAUG remains, this would create a uORF or oORF. Green indicates start codons and red indicates stop codons.*

Figure 6.7 shows the number of patients in this cohort with the kind of start-stop variants they had. As predicted, the majority of uAUG-creating variants were in TSGs (12/16 patients), and majority of stop-gain variants were in oncogenes (10/11 patients).

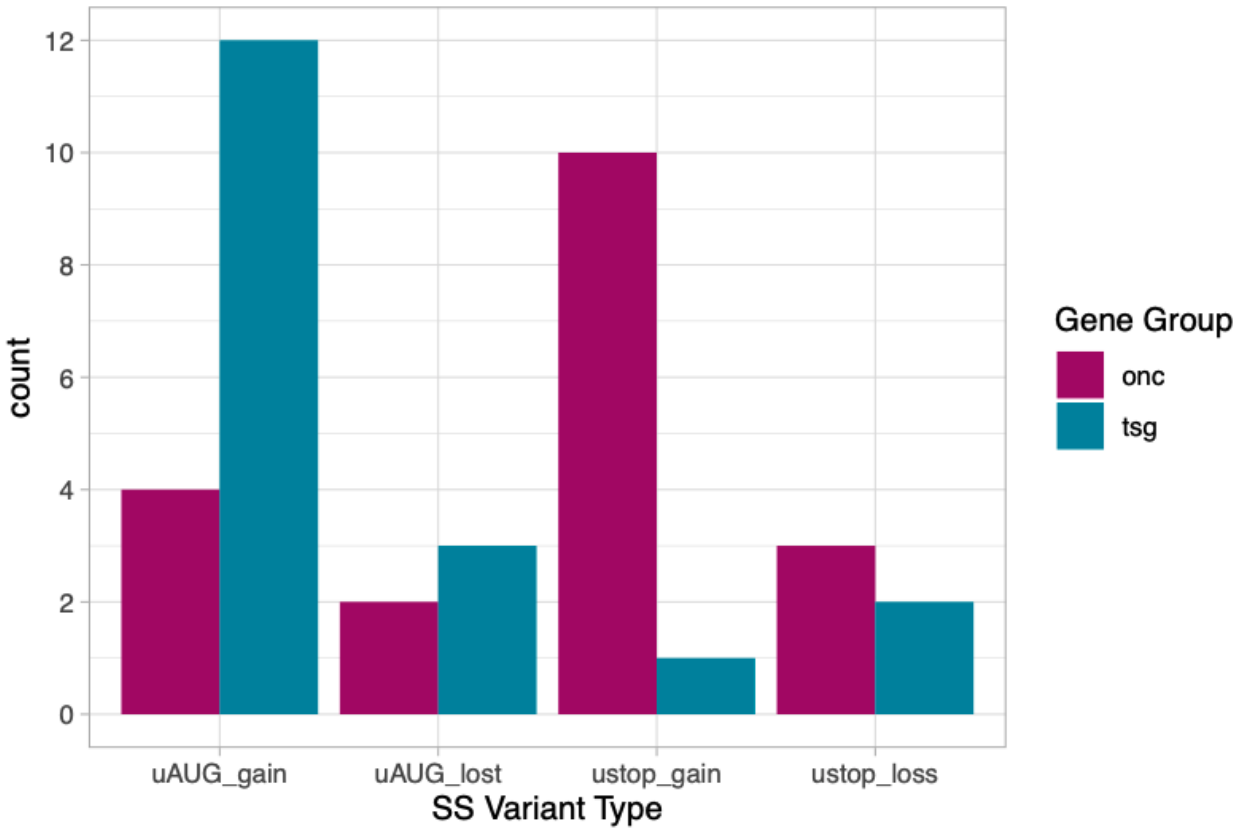


Figure 6.7: *Number of patients within each variant group in GEL cancer arm. 37 patients with 30 variants. Most start-stop variants were uAUG gain in TSGs (teal) and uStop-gain in oncogenes (dark pink).*

To test the significance of the amount of variants in a mutation class that was present in one gene group over another (TSG or oncogenes), I conducted T-tests. The only significant result was the uStop-gain mutation class, where oncogenes had significantly more start-stops created (10) via a stop-gain variant than TSGs (1) (T-test: uAUG-gain, P=0.05; uAUG-loss, P=1; stop-gain, P=0.009; stop-loss, P=1). A stop-gain variant is where an existing uORF or oORF is turned into a start-stop. If start-stops are less repressive than uORFs or oORFs then this could be a plausible cause of an increase in protein levels and hence increased cell proliferation. I looked at these 10 patients with stop-gain variants in more detail (Table 6.3).

Gene	Variant	Number of patients with variant	Cancer Type	gnomAD AF/ individuals in gnomAD	PhyloP/ CADD	Variant Impact
GRM3	chr7:86644399C>T	7	Colorectal (5) Endometrial carcinoma (2)	$6.6x10^{-7}$/3	0.065/ 6.078	13 possible uORFs in 5'UTR, 3 of which overlap this locus. Turns a weak Kozak uORF into a start-stop.
SGK1	chr6:134317680C>A	1	Endometrial carcinoma	Variant absent in gnomAD	0.92/ 14.43	There are 7 predicted uORFs in this UTR. Turns a strong Kozak uORF into a start-stop.
DGCR8	chr22:20085825G>A	1	Colorectal	Variant absent in gnomAD	4.494/ 22.1	There are 7 predicted uORFs in this UTR. Turns a moderate Kozak into a start-stop.
ALK	chr2:29921581G>A	1	Colorectal	$6.6x10^{-6}$/1	1.421/ 19.33	There are 3 uORFs, this variant turns a moderate Kozak uORF into a start-stop.

Table 6.3: *Stop-gain variant in oncogenes in the cancer arm of GEL. 7 patients had the same variant GRM3:chr7:86644399:C>T. All variants were MANE transcripts and changed a uORF into a start-stop.*

Seven of the patients (five colorectal, two endometrial) had the same variant (GRM3:chr7:86644399C>T). The variant has an AF in gnomAD of $9.4x10^{-6}$, with only three individuals having this variant, indicating it is rare in the general healthy population. The *GRM3* gene has 12 potential uORFs (uAUG start) and one experimentally validated uORF (AAG start) [179] (Figure 6.8). There are two uORFs specifically which overlap slightly upstream to the start-stop created, meaning that if a ribosome translated either of them it is very unlikely that the new start-stop would be "seen" by the ribosome. It is difficult to determine whether this loss of uORF, which is replaced with a start-stop, would lead to a difference in protein levels without any functional studies, but based on the number of existing potential uORFs and the ones overlapping the variant it is unlikely. As this variant did appear in several participants with the same cancer, I checked if it had been highlighted in any previous genome wide association studies (GWAS) for colorectal cancer by searching for it in the GWAS catalogue (https://www.ebi.ac.uk/gwas/) but it was absent. *GRM3* has been implicated in colon cancer [183] and melanoma [184]. *GRM3* (glutamate receptor, metabotropic, 3) is mainly expressed in the central nervous system but is significantly upregulated in the majority of human colonic adenocarcinomas and colon cancer cell lines [183]. In summary, this specific variant has not previously been implicated in cancer, and it is difficult to predict whether the start-stop would be recognised by a scanning ribosome without functional studies as it is in close proximity to other uORFs. Further investigation into this variant would include interrogating the literature and other cancer databases for other 5'UTR variants in *GRM3* that may be associated with cancer. Additionally, it is entirely possible that this variant is not

acting in a start-stop altering manner but via a different mechanism such as altering splicing or secondary structure.

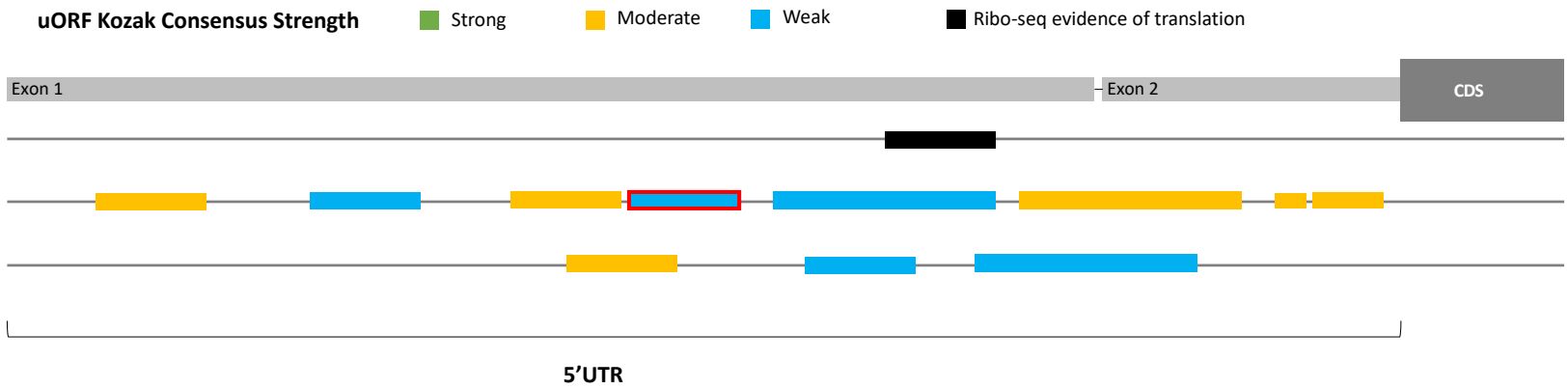


Figure 6.8: *Illustration of GRM3. There are 12 predicted uORFs and 1 uORF for which there is translational evidence [179]. The red uORF is the one that due to the variant chr7:86644399:C>T is turned into a start-stop.*

6.3 Discussion

In this chapter I show that start-stops are a rare phenomenon in 5'UTRs, and the chance of a stop codon occurring immediately after an AUG occurs more frequently than expected than by chance, but only in genes intolerant to LoF. Intriguingly, the hexamers ATGTGT and ATGTAT also appear more frequently than expected by chance, for reasons unknown.

Assessing start-stop creating or removing variants both in ClinVar and in GEL, I do not find any that are obviously pathogenic. Functional studies need to be done in order to confirm or refute the predictions. Additionally, this work has demonstrated how complex interpreting some 5'UTR variants can be given that any single variant may impact multiple elements in a number of different ways such as creating a start-stop within the frame of existing uORFs or altering Kozak consensus sequences of nearby uAUGs. Although the variants considered here are with regards to a start-stop mechanism, it is important to remember that there are other possible pathogenic mechanisms within 5'UTRs such as disrupting splicing or secondary structures that these variants could also be impacting.

More studies need to be done to confirm and expand upon the findings in the original start-stop study [39] and to illuminate more of the function and mechanism of start-stops, which will lead to more information useful in interpreting start-stop variants and their role in disease. I also think that possible start-stop impact on gene expression,

as demonstrated by other 5'UTR features, is specific to the genes they occur in, with dosage sensitivity and LoF tolerance, for example, affecting the extent of their impact. It is also possible that start-stop variants have a milder effect on protein expression and hence it would be interesting to explore start-stop creating or removing variants in a common disease context such as using the UK Biobank, to see if this increases power and helps us to learn more about start-stop biology.

6.4 Methods

6.4.1 Hexamer Shuffle

Shuffling of the MANE 5'UTR sequences was done as described in Chapter 3 (3.4.3). The same method was done on the forward strand and the reverse. The reverse was done by taking the forward 5'UTR strand and searched for CAT and the three bases preceding. I counted the occurrence of each hexamer across all genes in the unshuffled MANE 5'UTR set (observed), and in all genes in each simulation in the shuffled 5'UTRs (expected). I then calculated a O/E score per hexamer per simulation by dividing the observed hexamer counts by the expected. I then calculated a mean O/E per hexamer across all 5'UTRs.

6.4.2 ClinVar

5'UTR variants from ClinVar were downloaded from the FTP server (https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/weekly/) (27 May 2023). Mitochondrial variants were removed, and canonical chromosomes constrained to

5’UTR regions according to MANE v1.0 5’UTR regions. The VEP UTRannotator was run on the BMRC using Tabix. Downloading ClinVar variants and running UTRannotator was done by Elston N. D'Souza.

6.4.3 GEL

The rare disease *de novo* variants from GEL were accessed within GEL [185] using the RLabKey API [186]. All *de novo* variants were obtained from the *de novo* flagged variants LabKey table (v16). This table comprises a set (799,026) of confidently called *de novo* variants across 13,949 trios [185]. Data was limited to GRCh38 genome alignment, consenting participants and only those that had been identified as stringently filtered by GEL (QC filtering). The *de novo* set was also limited to probands only. Using MANE 5’UTR start and end coordinates, only variants within these ranges were extracted (35,753 5’UTR variants from 10,797 probands). To annotate whether the gene could be a reasonable match for the phenotype, any that had a strong association with the individual's phenotype were identified in PanelApp. All HPO terms for each individual were extracted and cross referenced with human phenotype ontology terms linked to the gene (https://hpo.jax.org/app/). The phenotype matching was done by Alexandra Martin-Geary.

Using MANE 5’UTR start and end coordinates as ranges of interest, and only TSG and oncogenes from COSMIC project [154] (downloaded 27 April 2023), somatic aggregated cancer variants (v13) were accessed within GEL [187] using RLabKey API [186]. The data was restricted to consenting participants only. Extracting the

somatic cancer variants and running UTRannotator on them was done by Maria Fernandes.

6.4.4 Annotating 5'UTR variants with UTRannotator

The ClinVar and GEL variants were annotated using Ensembl's variant effect predictor (VEP) v99.1 [188] with UTRannotator [176]. I considered variants as start-stop creating or lost by extracting annotations as:

1. uAUG gain/loss: UTRannotator annotates the uAUG position with reference to the stop codon. If the distance to stop codon = 3, this indicates this is a start-stop (Figure 6.9A).
2. uStop gain/loss: UTRannotator annotates uStop gain or loss with reference to frame of uAUGs in 5'UTR. If it was denoted as a uStop gain or loss I extracted the 3 bases immediately preceding it and filtered to only those that were an AUG codon (Figure 6.9B).

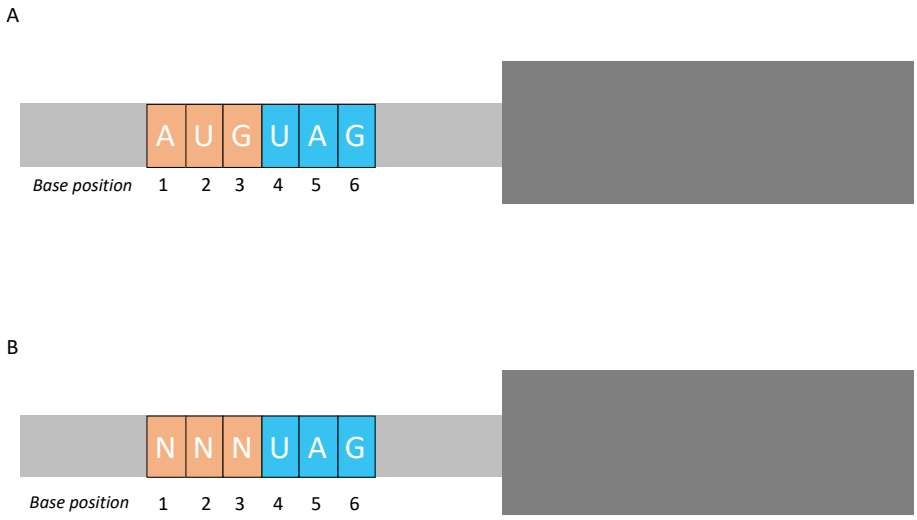


Figure 6.9: UTRannotator to detect start-stop creating or removing variants
A) *UTRannotator denotes uAUG gain or loss and distance the uAUG is to the stop codon. If the distance to the stop was three bases, this is considered a start-stop.* ***B)*** *UTRannotator denotes uStop gain or loss, and to determine whether the stop gained or lost is from a start-stop, the immediate three bases preceding the stop codon were extracted, and filtered to AUG codons.*

6.4.5 Annotating variants with gnomAD allele frequencies

Allele frequency scores were obtained using the gnomAD web browser (https://gnomad.broadinstitute.org/) v4.0.0. I inferred from the allele count and number of homozygotes the number of individuals in gnomAD with a variant. For example, if the allele count is 1127, and 11 are homozygous, that is 1127 - 22 = number of individuals.

6.4.6 Annotating variants with PhyloP and CADD scores

Variants were annotated with PhyloP and CADD scores obtained from the CADD single nucleotide variant webpage (https://cadd.gs.washington.edu/snv). Mammalian PhyloP scores were used.

Chapter 7

Conclusion and final remarks

In this thesis I have analysed 5'UTR features to gauge how they vary between genes and between gene sets. I present a review on the literature of UTR-perturbing variants and how they can lead to Mendelian disease. In addition, I have investigated start-stops, a lesser known 5'UTR feature, and explored variants that create or remove them in a disease context.

7.1 The challenges of defining UTR features and the complexity of interpreting UTR variants

I systematically assessed 5'UTR features and presented the literature to demonstrate where alterations in these features have been linked to disease. This work highlighted that whilst some things are known about UTRs, some things are not as well defined. For example, Internal Ribosome Entry Sites (IRES) have limited evidence supporting their widespread existence in humans. Downstream ORFs (dORFs) within the 3'UTR are enigmatic; how and why translation occurs within the 3'UTR is not well understood. This adds challenges to interpreting variants in these regions.

It is difficult to predict the proportion of non-coding rare disease diagnoses that are residing in the UTRs, but a recent study looking for UTR and promoter variants in 8,040 undiagnosed rare disease trios in GEL found 10 putative diagnoses, 9 of these

in UTRs [189]. A different study assessing 5'UTR variants in only inherited retinal conditions discovered 11 pathogenic variants in around 4,000 patients (a combination of their in-house data and GEL), explaining 0.28% of patients [87]. Even if UTRs contain a low number of pathogenic variants, each diagnosis is invaluable to the patient and their family. As demonstrated, UTR variants are difficult to interpret and adding experimental functional validation would certainly increase diagnoses, although the assays required are not currently standard in a clinical genetics service. This would also include and elucidate impact of variants in the UTR where we do not necessarily understand the mechanism.

UTR variants can disrupt more than one UTR feature or gene. For example, a region may be annotated as 5'UTR in one gene and an important region of CDS in another, which can lead to difficulty in understanding a variant's full effect. Importantly, variants in UTRs can act through multiple different mechanisms including impacting transcript, RNA-processing, RNA stability and translation. A variant's impact cannot be fully ruled out unless all these mechanisms have been captured in the experiments performed [140]. Alternatively, if an assay focuses on a downstream impact such as cell viability, the exact mechanism of effect may remain unclear. Further, UTRs act as entire functional elements. Including the entire sequence in its native context is important for accurate functional characterisation.

7.2 Conflicting and counterintuitive findings in the 5'UTRs of genes intolerant to LoF

Through chapters 4 and 5, I found that genes that are intolerant to LoF and genes implicated in disease have 5'UTRs that are longer, have a greater propensity to form secondary structures and have more uORFs. DD recessive genes have remarkably different 5'UTRs to the other disease gene groups, and more likely reflect the average non-dosage sensitive gene.

There are some seemingly conflicting and counterintuitive findings in these genes 5'UTRs. uORFs in LoF intolerant genes are further away from the CDS implying they may be less repressive as the ribosome can then reinitiate and translate the CDS, but these genes 5'UTRs are also more structured which can impede scanning ribosomes. I hypothesise that there may be biological explanations for these trends. Just because uORFs are further away from the CDS does not necessarily mean they completely lack a repressive effect on the CDS. Rather, it suggests that they exert less repressive impact compared to uORFs closer to the CDS. For some genes, light repression is all that is required. Additionally, this finding was an overall trend, but many of these genes have multiple uAUGs so there may be many cases where there are several uORFs, in various positions within the 5'UTR. A more structured 5'UTR may impede scanning ribosomes, slowing them down, and this may also be a form of regulatory control; fine tuning the speed of protein production. However, there may be mechanisms to counter the effects of a highly structured 5'UTR; IRESs themselves are highly structured sequences, but they promote translation [190].

Why LoF intolerant genes would have 5'UTRs that have more uORFs first struck me as counterintuitive - if these genes are so important, i.e. the proteins they produce are essential, surely they would be optimised to promote efficient translation? I think there are two main things to keep in mind that may reconcile these. Firstly, the commonality these genes have is that they are intolerant to LoF, but LoF intolerant genes' functions are extremely varied. Secondly, there was still considerable variety amongst the 5'UTRs of these genes. The differences in 5'UTRs between these genes may well point to their different biological functions. The quantity of protein a gene needs to produce for optimal function can vary drastically between genes. Additionally, the rate of transcription will differ between genes, yet different genes may require similar amounts of protein. To achieve the correct protein levels, there are mechanisms to buffer the transcript quantities to ensure that highly transcribed genes reduce their transcript levels and genes with lower transcription rates maintain sufficient transcript levels. There may be a LoF intolerant gene that has a short 5'UTR so has efficient translation of the CDS, but this gene is only expressed during a short window of time during a certain stage of lifespan. There are so many additional factors to take into account, it is difficult to pinpoint or draw conclusions between genes with a wide range of function, which may explain the considerable variety within each of the trends and patterns found in these genes. Some LoF intolerant genes may be essential during early embryogenesis (expressed during a specific time frame) and be part of a larger complex gene cascade, whereas some others might be simpler, lone-acting proteins essential for basic cell function (expressed throughout lifetime).

In summary, LoF intolerant genes contain a range of genes that go from requiring high levels of gene expression, with shorter, less complex 5'UTRs, to genes which tend to be highly specialised; only expressed in certain times or places, requiring very specific amounts of protein to be produced, which might be controlled by uORFs and other features. It is important to keep in mind that there are other non-5'UTR features controlling the amount of protein produced, and all of these would need to be taken into account to ascertain the precise impact 5'UTR specific features have on gene expression.

There is a statistically significant, discernible 5'UTR profile for these LoF intolerant genes, but further work needs to be done to unpack this. For example, further stratifying the genes by function, temporal and tissue specific expression and examining transcripts that are differentially expressed with varying 5'UTRs might elucidate further trends between genes 5'UTRs. Investigating specific 5'UTR features in the context of wider gene features might also shed light on the importance of 5'UTR features. For example, in genes which appear to have shorter, less complex 5'UTRs, do they have longer or specific combinations of 3'UTR features?

7.3 Interpretation of variants that create or perturb start-stops is challenging

I investigated start-stop occurrences in 5'UTRs and found that they are a rare occurrence, however, they appear more frequently than would be expected by

chance in genes intolerant to LoF. I searched ClinVar for any variant that would create or remove a start-stop. The three variants I assessed were very difficult to ascertain whether they are likely to be pathogenic without functional studies. For example, the most plausible one, the *FANCI*:chr15:89247638A>G variant is a uAUG-creating variant classified as a variant of unknown significance (VUS) for Fanconi anaemia (FA). This variant also changes an existing out-of-frame oORF, into a start-stop. If the start-stop is more repressive to translation than the existing oORF and reduces *FANCI* protein levels (as FA is a LoF disease), this could be a causative variant. However, the converse is more likely to be true; that the oORF is more repressive. There were four variants in the GEL rare disease arm, two cases of which had been solved by alternative methods. The remaining two are in genes not linked to the phenotype of the proband and are not thought to be pathogenic. Finally, in the GEL cancer arm, I found that there were 37 patients with 30 unique start-stop creating or removing variants. There were 7 participants with the same stop-gain variant in an oncogene (*GRM3:*chr7:86644399C>T). This variant changes a uORF into a start-stop, and if the start-stop is less repressive than the uORF then this could be a plausible cause of an increase in protein levels and hence increased cell proliferation. Indeed, this may be expected given the uORF may be translated followed by a risk of ribosome dissociation whereas the start-stop would likely only act to pause the scanning ribosome. Assessing the impact of start-stop variants is difficult, especially as in the *GRM3* gene there are 13 possible uORFs, 3 which overlap the position of the identified variant. In summary, given the lack of knowledge of start-stop biology and function, without functional studies it is difficult to know

whether these variants are likely pathogenic. The repressive effect of a start-stop is likely to be specific to the gene they occur in and influenced by various factors such as where in the 5'UTR they appear. This demonstrates how interpreting variants in 5'UTRs can be challenging.

7.4 A note on uORFs

A large focus of my research has been on uORFs. There are many factors influencing the chance of a uORF being translated and how repressive it is of CDS translation including length of the uORF, start codon composition, start codon consensus sequence, position within the 5'UTR and whether there are other uAUGs in the 5'UTR. There are also factors that have conflicting reports of the extent of their impact or have likely not yet been fully elucidated.

Generally, it has been demonstrated that presence of a uORF reduces downstream CDS expression [37], and creating or removing a uORF can lead to disease [91,93]. However, there are also reports showing that uORF translation is associated with increased translation of the CDS; some Ribo-seq studies have suggested that uORF translation is generally positively regulated with translation of the CDS [31,191,192], however, this is at odds with uORFs that have been fully characterised by measuring protein output [37,193] and it is currently unclear if this reflects biology or is an artefact of Ribo-seq.

Other work showed that translated uORFs only have a modest inhibitory effect, due to possibility of reinitiation, weaker uAUGs or near-cognate starts which are scanned over ("leaky scanning") [194]. It was also shown that there is a correlation between the strength of the CDS start codon consensus sequence and the number of uAUGs. Genes with "weak" CDS start codon consensus sequences had longer 5'UTRs with multiple uAUGs, suggesting that this may be a mechanism to ensure low basal translational levels via a double negative control [195]. Additionally, a factor that influences translation initiation of non-AUG starts is the number of upstream starts preceding it, as the number of ribosomal pre-initiation complexes is reduced after each subsequent start, so start codons closer to the 5' end are more likely to be translated [196]. There may be other features which promote translation of uORFs which have near-cognate starts in unfavourable consensus contexts; some studies showed that secondary structures following a uORF could cause the scanning ribosome to slide back towards the uORF and make it translate the uORF [197,198], suggesting that ribosomes may also be able to scan in a 3' to 5' direction. Also, there is a possibility that IRESs might be widespread in humans and indeed stimulate translation past uORFs, in which case uORFs in that context are not even "seen" by a scanning ribosome and do not have an effect. Additionally, IRESs may exert control possibly only under certain conditions, for example with certain ribosome types or under stress conditions, where cap-independent translation is preferred, or only is bound or not bound by certain RNA-binding proteins.

Another not yet fully determined concept is whether some uORFs produce functional peptides. If so, is it possible that their influence on CDS protein production is a by-product, and not a regulatory feature? Additionally, what are the roles of these peptides? Are they essential parts of the proteome? The current literature has pointed to evidence from mass spectrometry that some uORFs encode a detectable peptide product (SEPs; smORF encoded peptides) [34]. Other work has demonstrated that some SEPs may have a biological function [173]. There has also been a more recent opinion that some uORFs might be small exons with splicing potential surrounding them, suggesting that they are part of the proteome [199], but further work needs to be done to establish on whether this is the case, what their functional role is and how common this is.

In summary, the many uORF properties add complexity to their influence on gene regulation and determine the specific mechanism of translational control of a given gene. As there are so many factors influencing uORFs translational profile, it is difficult to ascertain its impact by just examining its sequence.

7.5 Future Directions and final remarks

The main motivation for embarking on this DPhil journey was to delve into the intricacies of UTRs in order to contribute to our understanding of their composition and biological function, with a view to aid in rare disease diagnostics - the more we understand these regions the more information we have to include to analyse them in a diagnostic setting. Looking forward, the continued advancement of our

understanding of UTRs will refine diagnostic methods and also pave the way for innovation in the therapeutic space.

Gene therapy is fast evolving and holds great promise. Targeting UTRs can be an elegant way to up or down regulate gene expression without the risk of interfering with the coding regions. An emerging therapy is the use of antisense oligonucleotides (ASOs). These are short single stranded RNAs which can bind to mRNA and modulate gene expression through various mechanisms such as blocking translation or altering splicing patterns [200]. They can be engineered to target uORFs, reduce its effect and increase CDS translation [201,202]. Neatly, approaches targeted to UTR regulatory elements enable modulation of protein expression controlled by natural expression patterns, as they only act on an RNA which is already expressed. This prevents issues with ectopic over-expression that are seen for gene-replacement therapies.

As demonstrated, given the challenges in interpretation of UTR sequence variants, functional characterisation is important. In addition to experiments to decipher the effect of individual variants observed in patients, large-scale multiplexed assays of variant effect (MAVEs) hold great promise not only for variant interpretation, but also for us to learn more about gene regulation. MAVEs are a method to introduce one variant per cell that is assayed for a phenotype of interest such as expression of a reporter gene. This can be used to systematically introduce a variant at every position in a gene of interest to produce a variant effect map [203]. Although this is still in early

days, it holds great potential for understanding UTR features. Progress in this direction is evident with initiatives like the Atlas of Variant Effects project [204].

In summary, this thesis has provided an overview of the current knowledge on UTR variants in disease, a comprehensive analysis on 5'UTR features and how they vary between different disease gene sets. Additionally, start-stop creating and removing variants were investigated in large-scale sequencing studies. Overall, this work adds to our understanding of UTR features and their function and aids in our knowledge and awareness of interpreting 5'UTRs in a diagnostic setting.

References

1. Haberle V, Stark A. Eukaryotic core promoters and the functional basis of transcription initiation. Nat Rev Mol Cell Biol. 2018 Oct;19(10):621–37.
2. OGBOURNE S, ANTALIS TM. Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. Biochemical Journal. 1998 Apr 1;331(1):1–14.
3. Gasperini M, Tome JM, Shendure J. Towards a comprehensive catalogue of validated and target-linked human enhancers. Nat Rev Genet. 2020 May;21(5):292–310.
4. Yang J, Corces VG. Chromatin Insulators: A Role in Nuclear Organization and Gene Expression. Adv Cancer Res. 2011;110:43–76.
5. Szabo Q, Bantignies F, Cavalli G. Principles of genome folding into topologically associating domains. Sci Adv. 2019 Apr 10;5(4):eaaw1668.
6. Wilkinson ME, Charenton C, Nagai K. RNA Splicing by the Spliceosome. Annu Rev Biochem. 2020 Jun 20;89:359–88.
7. Pesole G, Mignone F, Gissi C, Grillo G, Licciulli F, Liuni S. Structural and functional features of eukaryotic mRNA untranslated regions. Gene. 2001 Oct 3;276(1):73–81.
8. Ramanathan A, Robb GB, Chan SH. mRNA capping: biological functions and applications. Nucleic Acids Res. 2016 Sep 19;44(16):7511–26.
9. Tian B, Manley JL. Alternative cleavage and polyadenylation: the long and short of it. Trends Biochem Sci. 2013 Jun;38(6):312–20.
10. Wang J, Horlacher M, Cheng L, Winther O. RNA trafficking and subcellular localization—a review of mechanisms, experimental and predictive methodologies. Briefings in Bioinformatics. 2023 Sep 1;24(5):bbad249.
11. Beermann J, Piccoli MT, Viereck J, Thum T. Non-coding RNAs in Development and Disease: Background, Mechanisms, and Therapeutic Approaches. Physiological Reviews. 2016 Oct;96(4):1297–325.
12. Shirokikh NE, Preiss T. Translation initiation by cap-dependent ribosome recruitment: Recent insights and open questions. Wiley Interdiscip Rev RNA. 2018 Jul;9(4):e1473.
13. Hinnebusch AG, Lorsch JR. The Mechanism of Eukaryotic Translation Initiation: New Insights and Challenges. Cold Spring Harb Perspect Biol. 2012 Oct;4(10):a011544.
14. Leppek K, Das R, Barna M. Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. Nat Rev Mol Cell Biol. 2018 Mar;19(3):158–74.
15. Ciolli Mattioli C, Rom A, Franke V, Imami K, Arrey G, Terne M, et al. Alternative 3′ UTRs direct localization of functionally diverse protein isoforms in neuronal compartments. Nucleic Acids Research. 2019 Mar 18;47(5):2560–73.
16. Hughes TA. Regulation of gene expression by alternative untranslated regions. Trends in Genetics. 2006 Mar 1;22(3):119–22.
17. Resch AM, Ogurtsov AY, Rogozin IB, Shabalina SA, Koonin EV. Evolution of

alternative and constitutive regions of mammalian 5'UTRs. BMC Genomics. 2009 Apr 16;10(1):162.
18. Mayr C. What Are 3′ UTRs Doing? Cold Spring Harb Perspect Biol. 2019 Oct;11(10):a034728.
19. Tierney JAS, Świrski M, Tjeldnes H, Mudge JM, Kufel J, Whiffin N, et al. Ribosome Decision Graphs for the Representation of Eukaryotic RNA Translation Complexity [Internet]. bioRxiv; 2023 [cited 2023 Dec 20]. p. 2023.11.10.566564. Available from: https://www.biorxiv.org/content/10.1101/2023.11.10.566564v1
20. Hinnebusch AG, Ivanov IP, Sonenberg N. Translational control by 5'-untranslated regions of eukaryotic mRNAs. Science. 2016 Jun 17;352(6292):1413–6.
21. Wolfe AL, Singh K, Zhong Y, Drewe P, Rajasekhar VK, Sanghvi VR, et al. RNA G-quadruplexes cause eIF4A-dependent oncogene translation in cancer. Nature. 2014 Sep;513(7516):65–70.
22. Wang X, Hou J, Quedenau C, Chen W. Pervasive isoform-specific translational regulation via alternative transcription start sites in mammals. Mol Syst Biol. 2016 Jul 18;12(7):875.
23. Bohlen J, Fenzl K, Kramer G, Bukau B, Teleman AA. Selective 40S Footprinting Reveals Cap-Tethered Ribosome Scanning in Human Cells. Molecular Cell. 2020 Aug 20;79(4):561-574.e5.
24. Wieder N, D'Souza EN, Martin-Geary AC, Lassen FH, Talbot-Martin J, Fernandes M, et al. Differences in 5'untranslated regions highlight the importance of translational regulation of dosage sensitive genes. Genome Biol. 2024 Apr 29;25(1):111.
25. Carninci P, Kasukawa T, Katayama S, Gough J, al et. The Transcriptional Landscape of the Mammalian Genome. Science. 2005 Sep 2;309(5740):1559–63.
26. Eden E, Brunak S. Analysis and recognition of 5′ UTR intron splice sites in human pre-mRNA. Nucleic Acids Research. 2004 Feb 1;32(3):1131–42.
27. Rosenstiel P, Huse K, Franke A, Hampe J, Reichwald K, Platzer C, et al. Functional characterization of two novel 5' untranslated exons reveals a complex regulation of NOD2 protein expression. BMC Genomics. 2007 Dec 20;8(1):472.
28. Wang G, Guo X, Floros J. Differences in the translation efficiency and mRNA stability mediated by 5'-UTR splice variants of human SP-A1 and SP-A2 genes. Am J Physiol Lung Cell Mol Physiol. 2005 Sep;289(3):L497-508.
29. Weatheritt RJ, Sterne-Weiler T, Blencowe BJ. The ribosome-engaged landscape of alternative splicing. Nat Struct Mol Biol. 2016 Dec;23(12):1117–23.
30. Floor SN, Doudna JA. Tunable protein synthesis by transcript isoforms in human cells. eLife. 5:e10921.
31. Chothani SP, Adami E, Widjaja AA, Langley SR, Viswanathan S, Pua CJ, et al. A high-resolution map of human RNA translation. Molecular Cell [Internet]. 2022 Jul 15 [cited 2022 Jul 25];0(0). Available from: https://www.cell.com/molecular-cell/abstract/S1097-2765(22)00606-2
32. Gunišová S, Hronová V, Mohammad MP, Hinnebusch AG, Valášek LS. Please do not recycle! Translation reinitiation in microbes and higher eukaryotes. FEMS Microbiol Rev. 2018 Mar 1;42(2):165–92.
33. Bottorff TA, Park H, Geballe AP, Subramaniam AR. Translational buffering by

ribosome stalling in upstream open reading frames. PLOS Genetics. 2022 Oct 31;18(10):e1010460.
34. Renz PF, Valdivia-Francia F, Sendoel A. Some like it translated: small ORFs in the 5′UTR. Experimental Cell Research. 2020 Nov 1;396(1):112229.
35. Iacono M, Mignone F, Pesole G. uAUG and uORFs in human and rodent 5′untranslated mRNAs. Gene. 2005 Apr 11;349:97–105.
36. Zhang H, Wang Y, Wu X, Tang X, Wu C, Lu J. Determinants of genome-wide distribution and evolution of uORFs in eukaryotes. Nat Commun. 2021 Feb 17;12(1):1076.
37. Calvo SE, Pagliarini DJ, Mootha VK. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc Natl Acad Sci U S A. 2009 May 5;106(18):7507–12.
38. Tanaka M, Sotta N, Yamazumi Y, Yamashita Y, Miwa K, Murota K, et al. The Minimum Open Reading Frame, AUG-Stop, Induces Boron-Dependent Ribosome Stalling and mRNA Degradation. Plant Cell. 2016 Nov;28(11):2830–49.
39. Rendleman J, Haizel S, Wu S, Liu J, Ge X, Zou H, et al. Regulatory start-stop elements in 5' untranslated regions pervasively modulate translation [Internet]. bioRxiv; 2023 [cited 2024 May 16]. p. 2021.07.26.453809. Available from: https://www.biorxiv.org/content/10.1101/2021.07.26.453809v3
40. Schleich S, Acevedo JM, Clemm von Hohenberg K, Teleman AA. Identification of transcripts with short stuORFs as targets for DENR•MCTS1-dependent translation in human cells. Sci Rep. 2017 Jun 16;7(1):3722.
41. Noderer WL, Flockhart RJ, Bhaduri A, Diaz de Arce AJ, Zhang J, Khavari PA, et al. Quantitative analysis of mammalian translation initiation sites by FACS-seq. Mol Syst Biol. 2014 Aug 28;10(8):748.
42. Kozak M. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Research. 1987 Oct 26;15(20):8125–48.
43. Churbanov A, Rogozin IB, Babenko VN, Ali H, Koonin EV. Evolutionary conservation suggests a regulatory function of AUG triplets in 5′-UTRs of eukaryotic genes. Nucleic Acids Res. 2005;33(17):5512–20.
44. Kastenmayer JP, Ni L, Chu A, Kitchen LE, Au WC, Yang H, et al. Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae. Genome Res. 2006 Mar;16(3):365–73.
45. Hanada K, Higuchi-Takeuchi M, Okamoto M, Yoshizumi T, Shimizu M, Nakaminami K, et al. Small open reading frames associated with morphogenesis are hidden in plant genomes. Proc Natl Acad Sci U S A. 2013 Feb 5;110(6):2395–400.
46. Hemm MR, Paul BJ, Schneider TD, Storz G, Rudd KE. Small membrane proteins found by comparative genomics and ribosome binding site models. Mol Microbiol. 2008 Dec;70(6):1487–501.
47. Vattem KM, Wek RC. Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. Proc Natl Acad Sci U S A. 2004 Aug 3;101(31):11269–74.
48. Hellen CUT, Sarnow P. Internal ribosome entry sites in eukaryotic mRNA molecules. Genes Dev. 2001 Jan 7;15(13):1593–612.
49. Macejak DG, Sarnow P. Internal initiation of translation mediated by the 5′

leader of a cellular mRNA. Nature. 1991 Sep;353(6339):90–4.
50. Jackson RJ. The Current Status of Vertebrate Cellular mRNA IRESs. Cold Spring Harb Perspect Biol. 2013 Jan 2;5(2):a011569.
51. Weingarten-Gabbay S, Khan D, Liberman N, Yoffe Y, Bialik S, Das S, et al. The translation initiation factor DAP5 promotes IRES-driven translation of p53 mRNA. Oncogene. 2014 Jan;33(5):611–8.
52. Pelletier J, Sonenberg N. Insertion mutagenesis to increase secondary structure within the 5′ noncoding region of a eukaryotic mRNA reduces translational efficiency. Cell. 1985 Mar 1;40(3):515–26.
53. Andrzejewska A, Zawadzka M, Pachulska-Wieczorek K. On the Way to Understanding the Interplay between the RNA Structure and Functions in Cells: A Genome-Wide Perspective. Int J Mol Sci. 2020 Sep 15;21(18):6770.
54. Georgakopoulos-Soares I, Parada GE, Hemberg M. Secondary structures in RNA synthesis, splicing and translation. Computational and Structural Biotechnology Journal. 2022 Jan 1;20:2871–84.
55. Mw H, Mu M, Nc A. Balancing acts: molecular control of mammalian iron metabolism. Cell [Internet]. 2004 Apr 30 [cited 2023 Aug 9];117(3). Available from: https://pubmed.ncbi.nlm.nih.gov/15109490/
56. Cockman E, Anderson P, Ivanov P. TOP mRNPs: Molecular Mechanisms and Principles of Regulation. Biomolecules. 2020 Jun 27;10(7):969.
57. Hochstoeger T, Chao JA. Towards a molecular understanding of the 5′TOP motif in regulating translation of ribosomal mRNAs. Seminars in Cell & Developmental Biology [Internet]. 2023 Jun 12 [cited 2023 Jun 19]; Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7407576/
58. Ørom UA, Nielsen FC, Lund AH. MicroRNA-10a Binds the 5′UTR of Ribosomal Protein mRNAs and Enhances Their Translation. Molecular Cell. 2008 May 23;30(4):460–71.
59. Mayr C. Regulation by 3′-Untranslated Regions. Annual Review of Genetics. 2017;51(1):171–94.
60. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. Nat Rev Genet. 2014 Dec;15(12):829–45.
61. Hong D, Jeong S. 3'UTR Diversity: Expanding Repertoire of RNA Alterations in Human mRNAs. Molecules and Cells. 2023 Jan 31;46(1):48–56.
62. Bicknell AA, Cenik C, Chua HN, Roth FP, Moore MJ. Introns in UTRs: Why we should stop ignoring them. BioEssays. 2012;34(12):1025–34.
63. Nagy E, Maquat LE. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. Trends in Biochemical Sciences. 1998 Jun 1;23(6):198–9.
64. Curinha A, Oliveira Braz S, Pereira-Castro I, Cruz A, Moreira A. Implications of polyadenylation in health and disease. Nucleus. 2014;5(6):508–19.
65. Proudfoot NJ. Ending the message: poly(A) signals then and now. Genes Dev. 2011 Sep 1;25(17):1770–82.
66. Mayr C, Bartel DP. Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell. 2009 Aug 21;138(4):673.
67. Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. Proliferating cells

express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. Science. 2008 Jun 20;320(5883):1643–7.
68. Berkovits BD, Mayr C. Alternative 3' UTRs act as scaffolds to regulate membrane protein localization. Nature. 2015 Jun 18;522(7556):363–7.
69. O'Brien J, Hayder H, Zayed Y, Peng C. Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. Frontiers in Endocrinology [Internet]. 2018 [cited 2024 Feb 7];9. Available from: https://www.frontiersin.org/journals/endocrinology/articles/10.3389/fendo.2018.00402
70. Yekta S, Shih I hung, Bartel DP. MicroRNA-Directed Cleavage of HOXB8 mRNA. Science. 2004 Apr 23;304(5670):594–6.
71. Baou M, Jewell A, Murphy JJ. TIS11 Family Proteins and Their Roles in Posttranscriptional Gene Regulation. J Biomed Biotechnol. 2009;2009:634520.
72. Sidali A, Teotia V, Solaiman NS, Bashir N, Kanagaraj R, Murphy JJ, et al. AU-Rich Element RNA Binding Proteins: At the Crossroads of Post-Transcriptional Regulation and Genome Integrity. Int J Mol Sci. 2021 Dec 22;23(1):96.
73. Wu Q, Wright M, Gogol MM, Bradford WD, Zhang N, Bazzini AA. Translation of small downstream ORFs enhances translation of canonical main open reading frames. EMBO J. 2020 Sep 1;39(17):e104763.
74. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009 Apr 10;324(5924):218–23.
75. van den Akker GGH, Zacchini F, Housmans BAC, van der Vloet L, Caron MMJ, Montanaro L, et al. Current Practice in Bicistronic IRES Reporter Use: A Systematic Review. Int J Mol Sci. 2021 May 14;22(10):5193.
76. Hoque M, Li W, Tian B. Accurate Mapping of Cleavage and Polyadenylation Sites by 3′ Region Extraction and Deep Sequencing. In: Rorbach J, Bobrowicz AJ, editors. Polyadenylation: Methods and Protocols [Internet]. Totowa, NJ: Humana Press; 2014 [cited 2024 Apr 3]. p. 119–29. Available from: https://doi.org/10.1007/978-1-62703-971-0_10
77. Zheng D, Tian B. Polyadenylation Site-Based Analysis of Transcript Expression by 3'READS. Methods Mol Biol. 2017;1648:65–77.
78. Siddika T, Heinemann IU. Bringing MicroRNAs to Light: Methods for MicroRNA Quantification and Visualization in Live Cells. Front Bioeng Biotechnol. 2021 Jan 18;8:619583.
79. Ule J, Jensen K, Mele A, Darnell RB. CLIP: A method for identifying protein–RNA interaction sites in living cells. Methods. 2005 Dec 1;37(4):376–86.
80. Bartel DP. MicroRNA Target Recognition and Regulatory Functions. Cell. 2009 Jan 23;136(2):215–33.
81. Stark Z, Boughtwood T, Haas M, Braithwaite J, Gaff CL, Goranitis I, et al. Australian Genomics: Outcomes of a 5-year national program to accelerate the integration of genomics in healthcare. The American Journal of Human Genetics. 2023 Mar 2;110(3):419–26.
82. Clark MM, Stark Z, Farnaes L, Tan TY, White SM, Dimmock D, et al. Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. npj Genomic

Med. 2018 Jul 9;3(1):1–10.
83. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care — Preliminary Report. New England Journal of Medicine. 2021 Nov 11;385(20):1868–80.
84. Zhang F, Lupski JR. Non-coding genetic variants in human disease. Human Molecular Genetics. 2015 Oct 15;24(R1):R102–10.
85. Ellingford JM, Ahn JW, Bagnall RD, Baralle D, Barton S, Campbell C, et al. Recommendations for clinical interpretation of variants found in non-coding regions of the genome [Internet]. 2021 Dec [cited 2022 Jan 13] p. 2021.12.28.21267792. Available from: https://www.medrxiv.org/content/10.1101/2021.12.28.21267792v1
86. Wright CF, Quaife NM, Ramos-Hernández L, Danecek P, Ferla MP, Samocha KE, et al. Non-coding region variants upstream of MEF2C cause severe developmental disorder through three distinct loss-of-function mechanisms. The American Journal of Human Genetics. 2021 Jun 3;108(6):1083–94.
87. Rey AD, Valero M del P, Bouckaert M, Broeck FVD, Varela MD, Heetvelde MV, et al. Combining a prioritization strategy and functional studies nominates 5'UTR variants underlying inherited retinal disease [Internet]. medRxiv; 2023 [cited 2024 Jan 18]. p. 2023.06.19.23291376. Available from: https://www.medrxiv.org/content/10.1101/2023.06.19.23291376v1
88. Whiffin N, Karczewski KJ, Zhang X, Chothani S, Smith MJ, Evans DG, et al. Characterising the loss-of-function impact of 5' untranslated region variants in 15,708 individuals. Nature Communications. 2020 May 27;11(1):2523.
89. Ferdinandusse S, te Brinke H, Ruiter JPN, Haasjes J, Oostheim W, van Lenthe H, et al. A mutation creating an upstream translation initiation codon in SLC22A5 5′UTR is a frequent cause of primary carnitine deficiency. Hum Mutat. 2019 Oct;40(10):1899–904.
90. Coursimault J, Rovelet-Lecrux A, Cassinari K, Brischoux-Boucher E, Saugier-Veber P, Goldenberg A, et al. uORF-introducing variants in the 5′UTR of the NIPBL gene as a cause of Cornelia de Lange syndrome. Human Mutation. 2022;43(9):1239–48.
91. Zhou Y, Koelling N, Fenwick AL, McGowan SJ, Calpena E, Wall SA, et al. Disruption of TWIST1 translation by 5′ UTR variants in Saethre-Chotzen syndrome. Human Mutation. 2018 Oct;39(10):1360–5.
92. Cho TJ, Lee KE, Lee SK, Song SJ, Kim KJ, Jeon D, et al. A Single Recurrent Mutation in the 5′-UTR of IFITM5 Causes Osteogenesis Imperfecta Type V. The American Journal of Human Genetics. 2012 Aug 10;91(2):343–8.
93. Schulz J, Mah N, Neuenschwander M, Kischka T, Ratei R, Schlag PM, et al. Loss-of-function uORF mutations in human malignancies. Sci Rep. 2018 Feb 5;8(1):2395.
94. Ghilardi N, Wiestner A, Kikuchi M, Ohsaka A, Skoda RC. Hereditary thrombocythaemia in a Japanese family is caused by a novel point mutation in the thrombopoietin gene. British Journal of Haematology. 1999;107(2):310–6.
95. Ghilardi N, Wiestner A, Skoda RC. Thrombopoietin Production Is Inhibited by a Translational Mechanism. Blood. 1998 Dec 1;92(11):4023–30.
96. Evans DG, Bowers N, Burkitt-Wright E, Miles E, Garg S, Scott-Kitching V, et al. Comprehensive RNA Analysis of the NF1 Gene in Classically Affected NF1

Affected Individuals Meeting NIH Criteria has High Sensitivity and Mutation Negative Testing is Reassuring in Isolated Cases With Pigmentary Features Only. eBioMedicine. 2016 May 1;7:212–20.
97. Filatova AY, Vasilyeva TA, Marakhonov AV, Sukhanova NV, Voskresenskaya AA, Zinchenko RA, et al. Upstream ORF frameshift variants in the PAX6 5'UTR cause congenital aniridia. Hum Mutat. 2021 Aug;42(8):1053–65.
98. Li M, Yin M, Yang L, Chen Z, Du P, Sun L, et al. A novel splicing mutation in 5'UTR of GJB1 causes X-linked Charcot—Marie–tooth disease. Molecular Genetics & Genomic Medicine. n/a(n/a):e2108.
99. Dofash LNH, Monahan GV, Servián-Morilla E, Rivas E, Faiz F, Sullivan P, et al. A KLHL40 3' UTR splice-altering variant causes milder NEM8, an under-appreciated disease mechanism. Human Molecular Genetics. 2023 Apr 1;32(7):1127–36.
100. B P, Ac B, Kj C, J O, O EM. Evidence of pathogenicity of a mutation in 3' untranslated region causing mild haemophilia A. Haemophilia : the official journal of the World Federation of Hemophilia [Internet]. 2016 Jul [cited 2023 Jul 13];22(4). Available from: https://pubmed.ncbi.nlm.nih.gov/27216882/
101. Hudder A, Werner R. Analysis of a Charcot-Marie-Tooth Disease Mutation Reveals an Essential Internal Ribosome Entry Site Element in the Connexin-32 Gene *. Journal of Biological Chemistry. 2000 Nov 3;275(44):34586–91.
102. Grosz BR, Svaren J, Perez-Siles G, Nicholson GA, Kennerson ML. Revisiting the pathogenic mechanism of the GJB1 5' UTR c.-103C > T mutation causing CMTX1. Neurogenetics. 2021 Jul;22(3):149–60.
103. Cobbold LC, Wilson LA, Sawicka K, King HA, Kondrashov AV, Spriggs KA, et al. Upregulated c-myc expression in multiple myeloma by internal ribosome entry results from increased interactions with and expression of PTB-1 and YB-1. Oncogene. 2010 May 13;29(19):2884–91.
104. Brouwer JR, Willemsen R, Oostra BA. Microsatellite repeat instability and neurological disease. BioEssays. 2009;31(1):71–83.
105. Swinnen B, Robberecht W, Van Den Bosch L. RNA toxicity in non-coding repeat expansion disorders. The EMBO Journal. 2020 Jan 2;39(1):e101112.
106. Paulson H. Repeat expansion diseases. Handb Clin Neurol. 2018;147:105–23.
107. Green KM, Linsalata AE, Todd PK. RAN translation—what makes it run? Brain Res. 2016 Sep 15;1647:30–42.
108. Nolin SL, Brown WT, Glicksman A, Houck, Jr. GE, Gargano AD, Sullivan A, et al. Expansion of the Fragile X CGG Repeat in Females with Premutation or Intermediate Alleles. Am J Hum Genet. 2003 Feb;72(2):454–64.
109. Zhou Y, Kumari D, Sciascia N, Usdin K. CGG-repeat dynamics and FMR1 gene silencing in fragile X syndrome stem cells and stem cell-derived neurons. Molecular Autism. 2016 Oct 6;7(1):42.
110. Xi J, Wang X, Yue D, Dou T, Wu Q, Lu J, et al. 5′ UTR CGG repeat expansion in GIPC1 is associated with oculopharyngodistal myopathy. Brain. 2021 Feb 1;144(2):601–14.
111. Tietze JK, Pfob M, Eggert M, von Preußen A, Mehraein Y, Ruzicka T, et al. A non-coding mutation in the 5' untranslated region of patched homologue 1 predisposes to basal cell carcinoma. Exp Dermatol. 2013 Dec;22(12):834–5.

112. Timchenko LT, Miller JW, Timchenko NA, DeVore DR, Datar KV, Lin L, et al. Identification of a (CUG)n triplet repeat RNA-binding protein and its expression in myotonic dystrophy. Nucleic Acids Res. 1996 Nov 15;24(22):4407–14.
113. Cruchten RTP van, Wieringa B, Wansink DG. Expanded CUG repeats in DMPK transcripts adopt diverse hairpin conformations without influencing the structure of the flanking sequences. RNA. 2019 Jan 4;25(4):481–95.
114. Daughters RS, Tuttle DL, Gao W, Ikeda Y, Moseley ML, Ebner TJ, et al. RNA Gain-of-Function in Spinocerebellar Ataxia Type 8. PLoS Genet. 2009 Aug 14;5(8):e1000600.
115. Zu T, Gibbens B, Doty NS, Gomes-Pereira M, Huguet A, Stone MD, et al. Non-ATG–initiated translation directed by microsatellite expansions. Proceedings of the National Academy of Sciences. 2011 Jan 4;108(1):260–5.
116. Evans DGR, van Veen EM, Byers HJ, Wallace AJ, Ellingford JM, Beaman G, et al. A Dominantly Inherited 5′ UTR Variant Causing Methylation-Associated Silencing of BRCA1 as a Cause of Breast and Ovarian Cancer. The American Journal of Human Genetics. 2018 Aug 2;103(2):213–20.
117. de Jong VMT, Pruntel R, Steenbruggen TG, Bleeker FE, Nederlof P, Hogervorst FBL, et al. Identifying the BRCA1 c.-107A > T variant in Dutch patients with a tumor BRCA1 promoter hypermethylation. Fam Cancer. 2023 Apr;22(2):151–4.
118. Cazzola M, Skoda RC. Translational pathophysiology: a novel molecular mechanism of human disease. Blood. 2000 Jun 1;95(11):3280–8.
119. Luscieti S, Tolle G, Aranda J, Campos CB, Risse F, Morán É, et al. Novel mutations in the ferritin-L iron-responsive element that only mildly impair IRP binding cause hereditary hyperferritinaemia cataract syndrome. Orphanet J Rare Dis. 2013 Feb 19;8:30.
120. Suganuma M, Kono M, Yamanaka M, Akiyama M. Pathogenesis of a variant in the 5′ untranslated region of ADAR1 in dyschromatosis symmetrica hereditaria. Pigment Cell & Melanoma Research. 2020;33(4):591–600.
121. Maiti B, Arbogast S, Allamand V, Moyle MW, Anderson CB, Richard P, et al. A mutation in the SEPN1 selenocysteine redefinition element (SRE) reduces selenocysteine incorporation and leads to SEPN1-related myopathy. Hum Mutat. 2009 Mar;30(3):411–6.
122. Allamand V, Richard P, Lescure A, Ledeuil C, Desjardin D, Petit N, et al. A single homozygous point mutation in a 3′untranslated region motif of selenoprotein N mRNA causes SEPN1-related myopathy. EMBO Rep. 2006 Apr;7(4):450–4.
123. Villar-Quiles RN, von der Hagen M, Métay C, Gonzalez V, Donkervoort S, Bertini E, et al. The clinical, histologic, and genotypic spectrum of SEPN1-related myopathy. Neurology. 2020 Sep 15;95(11):e1512–27.
124. Mohan RA, van Engelen K, Stefanovic S, Barnett P, Ilgun A, Baars MJH, et al. A mutation in the Kozak sequence of GATA4 hampers translation in a family with atrial septal defects. Am J Med Genet A. 2014 Nov;164A(11):2732–8.
125. Nicolle R, Altin N, Siquier-Pernet K, Salignac S, Blanc P, Munnich A, et al. A non-coding variant in the Kozak sequence of RARS2 strongly decreases protein levels and causes pontocerebellar hypoplasia. BMC Medical Genomics. 2023 Jun 21;16(1):143.

126. Dr H, Se G, J L, Jb C, Dj W, Nj P. Alpha-thalassaemia caused by a polyadenylation signal mutation. Nature [Internet]. 1983 Nov [cited 2023 Jun 21];306(5941). Available from: https://pubmed.ncbi.nlm.nih.gov/6646217/
127. Johnston JJ, Williamson KA, Chou CM, Sapp JC, Ansari M, Chapman HM, et al. NAA10 polyadenylation signal variants cause syndromic microphthalmia. J Med Genet. 2019 Jul;56(7):444–52.
128. Gehring NH, Frede U, Neu-Yilik G, Hundsdoerfer P, Vetter B, Hentze MW, et al. Increased efficiency of mRNA 3' end formation: a new genetic mechanism contributing to hereditary thrombophilia. Nat Genet. 2001 Aug;28(4):389–92.
129. Morales J, Russell JE, Liebhaber SA. Destabilization of Human α-Globin mRNA by Translation Anti-termination Is Controlled during Erythroid Differentiation and Is Paralleled by Phased Shortening of the Poly(A) Tail *. Journal of Biological Chemistry. 1997 Mar 7;272(10):6607–13.
130. Weiss IM, Liebhaber SA. Erythroid Cell-Specific mRNA Stability Elements in the α2-Globin 39 Nontranslated Region. Molecular and Cellular Biology. 1995 May 1;15(5):2457–65.
131. Conne B, Stutz A, Vassalli JD. The 3′ untranslated region of messenger RNA: A molecular 'hotspot' for pathology? Nat Med. 2000 Jun;6(6):637–41.
132. Hughes AE, Bradley DT, Campbell M, Lechner J, Dash DP, Simpson DA, et al. Mutation Altering the miR-184 Seed Region Causes Familial Keratoconus with Cataract. Am J Hum Genet. 2011 Nov 11;89(5):628–33.
133. Conte I, Hadfield KD, Barbato S, Carrella S, Pizzo M, Bhat RS, et al. MiR-204 is responsible for inherited retinal dystrophy associated with ocular coloboma. Proceedings of the National Academy of Sciences. 2015 Jun 23;112(25):E3236–45.
134. Jedlickova J, Vajter M, Barta T, Black GCM, Perveen R, Mares J, et al. MIR204 n.37C>T variant as a cause of chorioretinal dystrophy variably associated with iris coloboma, early-onset cataracts and congenital glaucoma. Clin Genet. 2023 Jun 15;
135. Verdura E, Hervé D, Bergametti F, Jacquet C, Morvan T, Prieto-Morin C, et al. Disruption of a miR-29 binding site leading to COL4A1 upregulation causes pontine autosomal dominant microangiopathy with leukoencephalopathy. Ann Neurol. 2016 Nov;80(5):741–53.
136. Beetz C, Schüle R, Deconinck T, Tran-Viet KN, Zhu H, Kremer BPH, et al. REEP1 mutation spectrum and genotype/phenotype correlation in hereditary spastic paraplegia type 31. Brain. 2008 Apr;131(Pt 4):1078–86.
137. K D, K M, D D, M V, J G, S H, et al. Diagnostic high-throughput sequencing of 2396 patients with bleeding, thrombotic, and platelet disorders. Blood [Internet]. 2019 May 12 [cited 2023 Jun 5];134(23). Available from: https://pubmed.ncbi.nlm.nih.gov/31064749/
138. Zidan NI, AbdElmonem DM, Elsheikh HM, Metwally EA, Mokhtar WA, Osman GM. Relation between mutations in the 5′ UTR of ANKRD26 gene and inherited thrombocytopenia with predisposition to myeloid malignancies. An Egyptian study. Platelets. 2021 Jul 4;32(5):642–50.
139. Ruggieri A, Ramachandran N, Wang P, Haan E, Kneebone C, Manavis J, et al. Non-coding VMA21 deletions cause X-linked myopathy with excessive

autophagy. Neuromuscul Disord. 2015 Mar;25(3):207–11.
140. Ellingford JM, Ahn JW, Bagnall RD, Baralle D, Barton S, Campbell C, et al. Recommendations for clinical interpretation of variants found in non-coding regions of the genome. Genome Medicine. 2022 Jul 19;14(1):73.
141. Steri M, Idda ML, Whalen MB, Orrù V. Genetic variants in mRNA untranslated regions. WIREs RNA. 2018;9(4):e1474.
142. Griesemer D, Xue JR, Reilly SK, Ulirsch JC, Kukreja K, Davis JR, et al. Genome-wide functional screen of 3′UTR variants uncovers causal variants for human disease and evolution. Cell. 2021 Sep 30;184(20):5247-5260.e19.
143. Ryczek N, Łyś A, Makałowska I. The Functional Meaning of 5′UTR in Protein-Coding Genes. International Journal of Molecular Sciences. 2023 Jan;24(3):2976.
144. Morales J, Pujar S, Loveland JE, Astashyn A, Bennett R, Berry A, et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature. 2022 Apr;604(7905):310–5.
145. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010 Jan 1;20(1):110–21.
146. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Research. 2019 Jan 8;47(D1):D886–94.
147. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020 May;581(7809):434–43.
148. S G, M SB, Na W, W P, Jk G, M S, et al. Variant interpretation using population databases: Lessons from gnomAD. Human mutation [Internet]. 2022 Aug [cited 2024 Apr 10];43(8). Available from: https://pubmed.ncbi.nlm.nih.gov/34859531/
149. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013 Jun;45(6):580–5.
150. Forrest ARR, Kawaji H, Rehli M, Kenneth Baillie J, de Hoon MJL, Haberle V, et al. A promoter-level mammalian expression atlas. Nature. 2014 Mar;507(7493):462–70.
151. Hart T, Tong AHY, Chan K, Van Leeuwen J, Seetharaman A, Aregger M, et al. Evaluation and Design of Genome-Wide CRISPR/SpCas9 Knockout Screens. G3 (Bethesda). 2017 Aug 7;7(8):2719–27.
152. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012 Aug 17;337(6096):816–21.
153. McRae JF, Clayton S, Fitzgerald TW, Kaplanis J, Prigmore E, Rajan D, et al. Prevalence and architecture of de novo mutations in developmental disorders. Nature. 2017 Feb;542(7642):433–8.
154. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer. 2018 Nov;18(11):696–705.
155. Collins RL, Glessner JT, Porcu E, Lepamets M, Brandon R, Lauricella C, et al. A cross-disorder dosage sensitivity map of the human genome. Cell. 2022 Aug

4;185(16):3041-3055.e25.
156. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Research. 2014 Jan 1;42(D1):D980–5.
157. Turnbull C, Scott RH, Thomas E, Jones L, Murugaesu N, Pretty FB, et al. The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. BMJ. 2018 Apr 24;361:k1687.
158. Caulfield M, Davies J, Dennys M, Elbahy L, Fowler T, Hill S, et al. The 100,000 Genomes Project Protocol. 2017 Dec 21 [cited 2024 Apr 10]; Available from: https://figshare.com/articles/journal_contribution/GenomicEnglandProtocol_pdf/4530893/4
159. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Research. 2019 Jan 8;47(D1):D886–94.
160. Taliaferro JM, Lambert NJ, Sudmant PH, Dominguez D, Merkin JJ, Alexis MS, et al. RNA Sequence Context Effects Measured In Vitro Predict In Vivo Protein Binding and Regulation. Molecular Cell. 2016 Oct 20;64(2):294–306.
161. Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. The Vienna RNA Websuite. Nucleic Acids Research. 2008 Jul 1;36(suppl_2):W70–4.
162. Capell A, Fellerer K, Haass C. Progranulin transcripts with short and long 5' untranslated regions (UTRs) are differentially expressed via posttranscriptional and translational repression. J Biol Chem. 2014 Sep 12;289(37):25879–89.
163. Jiang M, Anderson J, Gillespie J, Mayne M. uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. BMC Bioinformatics. 2008 Apr 11;9:192.
164. Kozak M. Context Effects and Inefficient Initiation at Non-AUG Codons in Eucaryotic Cell-Free Translation Systems. Molecular and Cellular Biology. 1989 Nov 1;9(11):5073–80.
165. Dever TE. Gene-specific regulation by general translation factors. Cell. 2002 Feb 22;108(4):545–56.
166. Kozak M. Constraints on reinitiation of translation in mammals. Nucleic Acids Research. 2001 Dec 12;29(24):5226.
167. Pöyry TAA, Kaminski A, Jackson RJ. What determines whether mammalian ribosomes resume scanning after translation of a short upstream open reading frame? Genes Dev. 2004 Jan 1;18(1):62–75.
168. Chew GL, Pauli A, Schier AF. Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish. Nat Commun. 2016 May 24;7(1):11663.
169. Pechmann S, Frydman J. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. Nat Struct Mol Biol. 2013 Feb;20(2):237–43.
170. Forrest ME, Pinkard O, Martin S, Sweet TJ, Hanson G, Coller J. Codon and amino acid content are associated with mRNA stability in mammalian cells. PLoS One. 2020 Feb 13;15(2):e0228730.
171. Morris DR, Geballe AP. Upstream open reading frames as regulators of mRNA translation. Mol Cell Biol. 2000 Dec;20(23):8635–42.

172. Policastro RA, Zentner GE. Global approaches for profiling transcription initiation. Cell Rep Methods. 2021 Sep 16;1(5):100081.
173. McGillivray P, Ault R, Pawashe M, Kitchen R, Balasubramanian S, Gerstein M. A comprehensive catalog of predicted functional upstream open reading frames in humans. Nucleic Acids Research. 2018 Apr 20;46(7):3326–38.
174. Stankiewicz P, Sen P, Bhatt SS, Storer M, Xia Z, Bejjani BA, et al. Genomic and genic deletions of the FOX gene cluster on 16q24.1 and inactivating mutations of FOXF1 cause alveolar capillary dysplasia and other malformations. Am J Hum Genet. 2009 Jun;84(6):780–91.
175. Alveolar Capillary Dysplasia with Misalignment of Pulmonary Veins - Symptoms, Causes, Treatment | NORD [Internet]. [cited 2024 May 21]. Available from: https://rarediseases.org/rare-diseases/alveolar-capillary-dysplasia/
176. Zhang X, Wakeling M, Ware J, Whiffin N. Annotating high-impact 5′untranslated region variants with the UTRannotator. Bioinformatics. 2021 Apr 15;37(8):1171–3.
177. Calviello L, Mukherjee N, Wyler E, Zauber H, Hirsekorn A, Selbach M, et al. Detecting actively translated open reading frames in ribosome profiling data. Nat Methods. 2016 Feb;13(2):165–70.
178. Gaertner B, van Heesch S, Schneider-Lunitz V, Schulz JF, Witte F, Blachut S, et al. A human ESC-based screen identifies a role for the translated lncRNA LINC00261 in pancreatic endocrine differentiation. Sussel L, Stainier DY, editors. eLife. 2020 Aug 3;9:e58659.
179. Chen J, Brunner AD, Cogan JZ, Nuñez JK, Fields AP, Adamson B, et al. Pervasive functional translation of noncanonical human open reading frames. Science. 2020 Mar 6;367(6482):1140–6.
180. Ji Z, Song R, Regev A, Struhl K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. Sonenberg N, editor. eLife. 2015 Dec 19;4:e08890.
181. Whiffin N, Minikel E, Walsh R, O'Donnell-Luria AH, Karczewski K, Ing AY, et al. Using high-resolution variant frequencies to empower clinical genome interpretation. Genet Med. 2017 Oct;19(10):1151–8.
182. Pejaver V, Byrne AB, Feng BJ, Pagel KA, Mooney SD, Karchin R, et al. Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. Am J Hum Genet. 2022 Dec 1;109(12):2163–77.
183. Yi H, Geng L, Black A, Talmon G, Berim L, Wang J. The miR-487b-3p/GRM3/TGFβ signaling axis is an important regulator of colon cancer tumorigenesis. Oncogene. 2017 Jun 15;36(24):3477–89.
184. Prickett TD, Wei X, Cardenas-Navia I, Teer JK, Lin JC, Walia V, et al. Exon capture analysis of G protein-coupled receptors identifies activating mutations in GRM3 in melanoma. Nat Genet. 2011 Sep 25;43(11):1119–26.
185. De novo variant research dataset - Genomics England Research Environment User Guide [Internet]. [cited 2024 Mar 27]. Available from: https://re-docs.genomicsengland.co.uk/de_novo_data/
186. Labkey API - Genomics England Research Environment User Guide [Internet]. [cited 2024 Mar 27]. Available from: https://re-

docs.genomicsengland.co.uk/labkey_api/
187. Somatic aggregated variant call (somAgg v0.2 [ALPHA version]) - Genomics England Research Environment User Guide [Internet]. [cited 2024 Apr 3]. Available from: https://re-docs.genomicsengland.co.uk/somAgg/
188. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biology. 2016 Jun 6;17(1):122.
189. Martin-Geary AC, Blakes AJM, Dawes R, Findlay SD, Lord J, Walker S, et al. Systematic identification of disease-causing promoter and untranslated region variants in 8,040 undiagnosed individuals with rare disease [Internet]. medRxiv; 2023 [cited 2023 Oct 11]. p. 2023.09.12.23295416. Available from: https://www.medrxiv.org/content/10.1101/2023.09.12.23295416v1
190. Lozano G, Francisco-Velilla R, Martinez-Salas E. Deconstructing internal ribosome entry site elements: an update of structural motifs and functional divergences. Open Biol. 2018 Nov 28;8(11):180155.
191. van Heesch S, Witte F, Schneider-Lunitz V, Schulz JF, Adami E, Faber AB, et al. The Translational Landscape of the Human Heart. Cell. 2019 Jun 27;178(1):242-260.e29.
192. Duffy EE, Finander B, Choi G, Carter AC, Pritisanac I, Alam A, et al. Developmental dynamics of RNA translation in the human brain. Nat Neurosci. 2022 Oct;25(10):1353–65.
193. Ye Y, Liang Y, Yu Q, Hu L, Li H, Zhang Z, et al. Analysis of human upstream open reading frames and impact on gene expression. Hum Genet. 2015 Jun;134(6):605–12.
194. Andreev DE, Loughran G, Fedorova AD, Mikhaylova MS, Shatsky IN, Baranov PV. Non-AUG translation initiation in mammals. Genome Biology. 2022 May 9;23(1):111.
195. Rogozin IB, Kochetov AV, Kondrashov FA, Koonin EV, Milanesi L. Presence of ATG triplets in 5' untranslated regions of eukaryotic cDNAs correlates with a 'weak' context of the start codon. Bioinformatics. 2001 Oct;17(10):890–900.
196. Michel AM, Andreev DE, Baranov PV. Computational approach for calculating the probability of eukaryotic translation initiation from ribo-seq data that takes into account leaky scanning. BMC Bioinformatics. 2014 Nov 21;15(1):380.
197. Guenther UP, Weinberg DE, Zubradt MM, Tedeschi FA, Stawicki BN, Zagore LL, et al. The helicase Ded1p controls use of near-cognate translation initiation codons in 5' UTRs. Nature. 2018 Jul;559(7712):130–4.
198. Murat P, Marsico G, Herdy B, Ghanbarian AT, Portella G, Balasubramanian S. RNA G-quadruplexes at upstream open reading frames cause DHX36- and DHX9-dependent translation of human mRNAs. Genome Biol. 2018 Dec 27;19(1):229.
199. Ji HJ, Salzberg SL. Upstream open reading frames may contain hundreds of novel human exons [Internet]. bioRxiv; 2024 [cited 2024 Apr 11]. p. 2024.03.22.586333. Available from: https://www.biorxiv.org/content/10.1101/2024.03.22.586333v2
200. Collotta D, Bertocchi I, Chiapello E, Collino M. Antisense oligonucleotides: a novel Frontier in pharmacological strategy. Front Pharmacol. 2023 Nov 17;14:1304342.
201. Liang X hai, Shen W, Sun H, Migawa MT, Vickers TA, Crooke ST. Translation

efficiency of mRNAs is increased by antisense oligonucleotides targeting upstream open reading frames. Nat Biotechnol. 2016 Aug;34(8):875–80.
202. Liang X hai, Sun H, Shen W, Wang S, Yao J, Migawa MT, et al. Antisense oligonucleotides targeting translation inhibitory elements in 5′ UTRs can selectively increase protein levels. Nucleic Acids Res. 2017 Sep 19;45(16):9528–46.
203. Kinney JB, McCandlish DM. Massively Parallel Assays and Quantitative Sequence–Function Relationships. Annual Review of Genomics and Human Genetics. 2019 Aug 31;20(Volume 20, 2019):99–127.
204. Fowler DM, Adams DJ, Gloyn AL, Hahn WC, Marks DS, Muffley LA, et al. An Atlas of Variant Effects to understand the genome at nucleotide resolution. Genome Biol. 2023 Jul 3;24(1):147.
205. Topisirovic I, Sonenberg N. Distinctive tRNA Repertoires in Proliferating versus Differentiating Cells. Cell. 2014 Sep 11;158(6):1238–9.

Appendix

LOEUF decile	mean (bps)	median (bps)	range (bps)	sd
10	284.57	216	3-2253	253.91
20	253.21	194	4-2252	221.57
30	237.96	172	5-2761	228.37
40	212.50	155	4-1900	204.95
50	177.88	128	3-1809	168.84
60	172.34	117	3-2732	195.21
70	164.66	111	2-2545	186.12
80	158.69	104	2-1614	168.50
90	161.42	100	1-3548	211.40
100	163.70	93	4-3561	235.01

Appendix Table 1: Data for Figure 4.2, 5'UTR length across LOEUF deciles.

LOEUF decile	ORF type	count	total in decile	%
10	oORF	305	1841	16.57
10	Start-Stop	129	1841	7.01
10	uORF	874	1841	47.47
20	oORF	285	1854	15.37
20	Start-Stop	123	1854	6.63
20	uORF	832	1854	44.88
30	oORF	267	1853	14.41
30	Start-Stop	112	1853	6.04
30	uORF	739	1853	39.88
40	oORF	265	1850	14.32
40	Start-Stop	85	1850	4.59
40	uORF	644	1850	34.81
50	oORF	249	1826	13.64
50	Start-Stop	58	1826	3.18
50	uORF	593	1826	32.48
60	oORF	236	1814	13.01
60	Start-Stop	70	1814	3.86
60	uORF	533	1814	29.38
70	oORF	257	1783	14.41
70	Start-Stop	82	1783	4.60
70	uORF	505	1783	28.32
80	oORF	252	1727	14.59
80	Start-Stop	65	1727	3.76
80	uORF	471	1727	27.27
90	oORF	251	1600	15.69
90	Start-Stop	70	1600	4.38
90	uORF	443	1600	27.69
100	oORF	216	1189	18.17
100	Start-Stop	55	1189	4.63
100	uORF	331	1189	27.84

Appendix Table 2: Data for Figure 4.5, uAUG type across LOEUF deciles.

LOEUF decile	start type	count	total in decile	%
10	canonical	532	912	58.33
10	non-canonical	380	912	41.67
20	canonical	432	762	56.69
20	non-canonical	330	762	43.31
30	canonical	369	661	55.82
30	non-canonical	292	661	44.18
40	canonical	282	538	52.42
40	non-canonical	256	538	47.58
50	canonical	277	495	55.96
50	non-canonical	218	495	44.04
60	canonical	210	388	54.12
60	non-canonical	178	388	45.88
70	canonical	204	390	52.31
70	non-canonical	186	390	47.69
80	canonical	156	317	49.21
80	non-canonical	161	317	50.79
90	canonical	130	235	55.32
90	non-canonical	105	235	44.68
100	canonical	69	136	50.74
100	non-canonical	67	136	49.26

Appendix Table 3: Data for Figure 4.10, uORF start codons across LOEUF deciles.

group	LOEUF decile	mean	median	sd	min	max
5'UTR	10	1.213	1.008	0.999	-1.150	7.483
5'UTR	20	0.991	0.813	0.941	-1.438	7.714
5'UTR	30	0.738	0.580	0.828	-1.500	6.144
5'UTR	40	0.521	0.404	0.797	-3.083	6.615
5'UTR	50	0.349	0.268	0.691	-1.691	6.060
5'UTR	60	0.220	0.148	0.672	-1.750	8.012
5'UTR	70	0.123	0.059	0.653	-2.417	6.784
5'UTR	80	0.109	0.054	0.632	-1.917	5.381
5'UTR	90	0.049	0.000	0.637	-2.000	5.849
5'UTR	100	-0.036	-0.069	0.578	-3.840	5.348
uORF Start	10	2.148	1.833	1.823	-2.667	9.000
uORF Start	20	1.888	1.667	1.809	-2.667	9.000
uORF Start	30	1.393	1.000	1.688	-3.000	7.667
uORF Start	40	1.273	1.000	1.600	-2.833	9.000
uORF Start	50	0.798	0.500	1.454	-2.667	9.000
uORF Start	60	0.453	0.167	1.364	-2.667	8.333
uORF Start	70	0.334	0.167	1.234	-3.000	8.000
uORF Start	80	0.277	0.000	1.215	-3.500	7.667
uORF Start	90	0.134	0.000	1.159	-3.333	6.667
uORF Start	100	-0.020	0.000	1.162	-3.833	7.333
uORF Stop	10	2.068	1.833	1.695	-3.167	9.000
uORF Stop	20	1.815	1.500	1.652	-3.333	9.000
uORF Stop	30	1.367	1.000	1.514	-2.333	8.333
uORF Stop	40	1.234	1.000	1.470	-2.667	8.333
uORF Stop	50	0.827	0.667	1.271	-2.667	9.000
uORF Stop	60	0.565	0.333	1.210	-3.833	8.000
uORF Stop	70	0.465	0.333	1.077	-3.167	6.667
uORF Stop	80	0.439	0.167	1.157	-3.333	7.667
uORF Stop	90	0.291	0.167	1.052	-2.667	5.333
uORF Stop	100	0.136	0.000	0.977	-3.333	8.000
Start-Stop	10	2.429	2.333	1.698	-0.917	8.000
Start-Stop	20	1.781	1.500	1.551	-1.167	8.333
Start-Stop	30	1.593	1.333	1.480	-0.583	6.000

Start-Stop	40	1.195	0.917	1.308	-0.750	5.167
Start-Stop	50	1.014	0.583	1.467	-0.750	7.667
Start-Stop	60	0.491	0.250	1.124	-1.000	4.833
Start-Stop	70	0.293	0.083	1.175	-1.500	7.667
Start-Stop	80	0.270	0.042	0.944	-1.667	3.417
Start-Stop	90	0.356	0.000	1.183	-1.250	5.667
Start-Stop	100	0.149	0.000	0.897	-1.833	4.000

Appendix Table 4: Data for Figure 4.14, PhyloP across LOEUF deciles.

LOEUF decile	CAGE peaks	count	total in decile	%
10	1	133	1809	7.352
20	1	161	1824	8.827
30	1	191	1810	10.552
40	1	237	1803	13.145
50	1	268	1765	15.184
60	1	314	1716	18.298
70	1	329	1676	19.630
80	1	320	1567	20.421
90	1	372	1375	27.055
100	1	253	888	28.491
10	2-5	815	1809	45.053
20	2-5	903	1824	49.507
30	2-5	894	1810	49.392
40	2-5	975	1803	54.077
50	2-5	1017	1765	57.620
60	2-5	998	1716	58.159
70	2-5	981	1676	58.532
80	2-5	925	1567	59.030
90	2-5	745	1375	54.182
100	2-5	525	888	59.122
10	≥6	861	1809	47.595
20	≥6	760	1824	41.667
30	≥6	725	1810	40.055
40	≥6	591	1803	32.779
50	≥6	480	1765	27.195
60	≥6	404	1716	23.543
70	≥6	366	1676	21.838
80	≥6	322	1567	20.549
90	≥6	258	1375	18.764
100	≥6	110	888	12.387

Appendix Table 5: Data for Figure 4.16, CAGE peaks across LOEUF deciles.

LOEUF decile	mean	median	range	sd
10	-122.70	-90.60	-844.4-0	116.31
20	-107.64	-81.25	-972.7-0	99.24
30	-99.12	-70.70	-871.7-0	99.16
40	-87.96	-61.15	-745.5-0	91.23
50	-69.59	-45.70	-1023.8-0	75.03
60	-65.06	-40.30	-1051-0	82.42
70	-59.91	-37.80	-828.4-0	73.52
80	-55.72	-34.20	-770.8-0	66.92
90	-55.73	-31.05	-1302.8-0	80.38
100	-53.55	-26.20	-1185-0	86.11

Appendix Table 6: Data for Figure 4.18, MFE scores for 5'UTRs across LOEUF deciles.