



University of Pennsylvania
ScholarlyCommons


Publicly Accessible Penn Dissertations

2021

Insights Into Functional Noncoding Rna Elements Through The Analysis Of Human Genetic Variation

David Sheng Ming Lee
University of Pennsylvania

Follow this and additional works at: <https://repository.upenn.edu/edissertations>

 Part of the [Biology Commons](#), and the [Genetics Commons](#)

Recommended Citation

Lee, David Sheng Ming, "Insights Into Functional Noncoding Rna Elements Through The Analysis Of Human Genetic Variation" (2021). *Publicly Accessible Penn Dissertations*. 4068.
<https://repository.upenn.edu/edissertations/4068>

This paper is posted at ScholarlyCommons. <https://repository.upenn.edu/edissertations/4068>
For more information, please contact repository@pobox.upenn.edu.

Insights Into Functional Noncoding Rna Elements Through The Analysis Of Human Genetic Variation

Abstract

Most of the human genome is noncoding but knowing how and when genetic variation in noncoding regions of the genome can impact biology and disease susceptibility remains challenging. Here, we apply an integrated genomics approach towards understanding and elucidating new patterns of functional genetic variation in untranslated regions of protein-coding messenger RNAs.

G-quadruplex (G4) sequences are abundant in untranslated regions (UTRs) of human messenger RNAs, but their functional importance remains unclear. In Part 1 of this dissertation, we integrate multiple sources of genetic and genomic data to show that putative G-quadruplex forming sequences (pG4) in 5' and 3' UTRs are selectively constrained and enriched for cis-eQTLs and RNA-binding protein (RBP) interactions. Using over 15,000 whole genome sequences, we find evidence of strong negative selection acting on central guanines of UTR pG4s. At multiple GWAS-implicated SNPs within pG4 UTR sequences, we find robust allelic imbalance in gene expression across diverse tissue contexts in GTEx, suggesting that variants affecting G4 formation in UTRs may also contribute to phenotypic variation. Our results establish UTR G4s as important cis-regulatory elements and point to a link between disruption of UTR pG4 and disease.

In Part 2 of this dissertation, we examine patterns of selective pressure in non-canonical open reading frames (ncORFs) mapped throughout the human genome. Ribosome-profiling has uncovered pervasive translation in ncORFs, however the biological significance of this phenomenon remains unclear. Using genetic variation from 71,702 human genomes, we assess patterns of selection in translated upstream open reading frames (uORFs) in 5'UTRs. We show that uORF variants introducing new stop codons, or strengthening existing stop codons, are under strong negative selection comparable to protein-coding missense variants. Using these variants, we map and validate new gene-disease associations in two independent biobanks containing exome sequencing from 10,900 and 32,268 individuals, respectively, and elucidate their impact of protein expression in human cells. Our results suggest new mechanisms relating uORF variation to reduced protein expression and demonstrate that translation at uORFs is genetically constrained in 50% of human genes.

Together, these studies help emphasize the importance of noncoding RNA regulatory elements in mediating post-transcriptional regulation of gene expression and illuminate new patterns of functional variation in UTRs with human disease relevance.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Genomics & Computational Biology

First Advisor

Yoseph Barash

Second Advisor

Louis R. Ghanem

Keywords

G-Quadruplex, Genetic Variation, non-canonical open reading frames, Regulatory elements, Ribosome Profiling, RNA

Subject Categories

Biology | Genetics

INSIGHTS INTO FUNCTIONAL NONCODING RNA ELEMENTS THROUGH THE ANALYSIS OF
HUMAN GENETIC VARIATION

David S. M. Lee

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2021

Supervisor of Dissertation

Yoseph Barash

Associate Professor of Genetics

Co-Supervisor of Dissertation

Louis R. Ghanem

Assistant Professor of Pediatrics

Graduate Group Chairperson

Benjamin F. Voight

Associate Professor of Systems Pharmacology and Translational Therapeutics

Dissertation Committee

Robert O. Heuckeroth, Professor of Pediatrics

Rachel Green, Bloomberg Distinguished Professor of Molecular Biology and Genetics

Zissimos Mourelatos, Professor of Pathology and Laboratory Medicine

Marylyn D. Ritchie, Professor of Genetics

Benjamin F. Voight, Associate Professor of Systems Pharmacology and Translational
Therapeutics

INSIGHTS INTO FUNCTIONAL NONCODING RNA ELEMENTS THROUGH THE ANALYSIS OF
HUMAN GENETIC VARIATION

COPYRIGHT

2021

David Sheng Ming Lee

ACKNOWLEDGMENTS

First and foremost, I am indebted to my thesis advisors Yoseph Barash and Lou Ghanem for their unwavering mentorship, feedback, and kindness. Graduate school has been an incredible adventure, and none of this would have been possible without either of your unwavering support and encouragement. I will never forget the feelings of excitement and exploration during our many meetings which were always a palpable presence in the room. I am also deeply grateful to my committee, Robert Heuckeroth, Rachel Green, Zissimos Mourelatos, Marylyn Ritchie, and Benjamin Voight, for their thoughtful comments and open doors over the past few years.

To Celeste Simon and Brian Keith, thank you for taking a chance on me as a chemistry major fresh out of Haverford College (whose most salient research experience to date was a far cry from cancer metabolism and hypoxia signaling). Working in your laboratory introduced me to the exciting world of translational research and convinced me that pursuing MD/PhD training was the right path for me.

Lastly, thank you and you, Mom and Dad, for all your loving support, and empowerment to pursue my passions. Thank you, Jamie for never failing to remind me what it feels like to be a kid again. Thank you, Hannah, for being my strongest shelter, my truest compass, and my tireless partner.

ABSTRACT

INSIGHTS INTO FUNCTIONAL NONCODING RNA ELEMENTS THROUGH THE ANALYSIS OF HUMAN GENETIC VARIATION

David S.M. Lee

Yoseph Barash

Louis R. Ghanem

Most of the human genome is noncoding but knowing how and when genetic variation in noncoding regions of the genome can impact biology and disease susceptibility remains challenging. Here, we apply an integrated genomics approach towards understanding and elucidating new patterns of functional genetic variation in untranslated regions of protein-coding messenger RNAs.

G-quadruplex (G4) sequences are abundant in untranslated regions (UTRs) of human messenger RNAs, but their functional importance remains unclear. In Part 1 of this dissertation, we integrate multiple sources of genetic and genomic data to show that putative G-quadruplex forming sequences (pG4) in 5' and 3' UTRs are selectively constrained and enriched for cis-eQTLs and RNA-binding protein (RBP) interactions. Using over 15,000 whole genome sequences, we find evidence of strong negative selection acting on central guanines of UTR pG4s. At multiple GWAS-implicated SNPs within pG4 UTR sequences, we find robust allelic imbalance in gene expression across diverse tissue contexts in GTEx, suggesting that variants affecting G4 formation in UTRs may also contribute to phenotypic variation. Our results establish UTR G4s as important cis-regulatory elements and point to a link between disruption of UTR pG4 and disease.

In Part 2 of this dissertation, we examine patterns of selective pressure in non-canonical open reading frames (ncORFs) mapped throughout the human genome. Ribosome-profiling has uncovered pervasive translation in ncORFs, however the biological significance of this phenomenon remains unclear. Using genetic variation from 71,702 human genomes, we assess patterns of selection in translated upstream open reading frames (uORFs) in 5'UTRs. We show that uORF variants introducing new stop codons, or strengthening existing stop codons, are under strong negative selection comparable to protein-coding missense variants. Using these variants, we map and validate new gene-disease associations in two independent biobanks containing exome sequencing from 10,900 and 32,268 individuals, respectively, and elucidate their impact of protein expression in human cells. Our results suggest new mechanisms relating uORF variation to reduced protein expression and demonstrate that translation at uORFs is genetically constrained in 50% of human genes.

Together, these studies help emphasize the importance of noncoding RNA regulatory elements in mediating post-transcriptional regulation of gene expression and illuminate new patterns of functional variation in UTRs with human disease relevance.

TABLE OF CONTENTS

ACKNOWLEDGMENTS	iii
ABSTRACT	iv
LIST OF TABLES	vii
LIST OF FIGURES	viii
CHAPTER 1: THE REGULATORY RENAISSANCE	1
1.1: How many genes?.....	1
1.2: Genetic variation and common disease	2
1.3: The challenge of interpreting noncoding variation	4
1.4: Untranslated regions in messenger RNA	5
1.5: Identifying variants of interest in the noncoding genome.....	7
CHAPTER 2: INTEGRATIVE ANALYSIS REVEALS RNA G-QUADRUPLEXES IN UTRS ARE SELECTIVELY CONSTRAINED AND ENRICHED FOR FUNCTIONAL ASSOCIATIONS	11
2.1: Secondary structures as RNA regulatory elements	11
2.2: G-quadruplexes are non-canonical secondary structures formed by nucleic acids	12
2.3: pG4 exhibit heightened selective pressure within UTRs	14
2.4: Most pG4 motifs in UTRs are isoform-restricted	20
2.5: pG4 motifs in the 5' and 3' UTR are enriched for cis-eQTLs	24
2.6: RNA-protein binding sites are enriched over UTR pG4 regions	26
2.7: 3'UTR pG4 in disease-causing genes are enriched for variants	30
2.8: Summary and future directions	33
2.9: Supplementary Materials to Integrative analysis reveals RNA G-Quadruplexes are selectively constrained and enriched for functional associations	36
CHAPTER 3: DISRUPTING UPSTREAM TRANSLATION IS ASSOCIATED WITH LOSS-OF-FUNCTION IN HUMAN DISEASE	42
3.1: Ribosome profiling and non-canonical open reading frames	42
3.2: Pervasive translation in non-canonical open reading frames.....	44
3.3: Disrupting upstream translation in mRNAs is associated with loss-of-function in human disease.....	46
3.3: Variants introducing new stop codons in uORFs are under strong negative selection	48
3.4: Translated uORFs use weak stop codons.....	50
3.5 Genomic positions that can create new stop codons in uORFs are conserved	52
3.6: Upstream open reading frames are not under strong selection to maintain amino acid identity.....	53
3.7: uORF start codons are conserved and under strong selective pressure	57
3.8: uORF-disrupting variants associate genes with new disease phenotypes	60
3.9: Disease-associated uORF variants change protein expression	64
3.10: Replication of novel associations by loss-of-function gene-burden studies	65
3.11: Summary and future directions	67
3.12: Supplementary Figures and Tables for Disrupting upstream translation in mRNAs is associated with human disease.....	72

3.13: Supplementary Note: Estimating the proportion of uORFs that may cause pathogenic loss-of-function equivalent consequences in ClinVar disease genes	86
CHAPTER 4: TRANSLATING REGULATORY INSIGHTS INTO THERAPIES	90
4.1: G-quadruplexes and upstream open reading frames in UTRs	90
4.2: From regulation to therapy.....	93
APPENDIX A: METHODS FOR INTEGRATIVE ANALYSIS REVEALS RNA G-QUADRUPLEXES IN UTRS ARE SELECTIVELY CONSTRAINED AND ENRICHED FOR FUNCTIONAL ASSOCIATIONS	95
APPENDIX B: METHODS FOR DISRUPTING UPSTREAM TRANSLATION IN MRNAS IS ASSOCIATED WITH LOSS-OF-FUNCTION IN HUMAN DISEASE	105
BIBLIOGRAPHY	118

LIST OF TABLES

Table 3.1: Significant novel associations in PheWAS of Penn Medicine BioBank.....	61
Table 3.2: uORF UTC / Stop-strengthening MAPS analysis with all CDS-overlapping variants removed.....	82
Table 3.3: Relative frequencies of TGA, TAG, and UAA trinucleotides across different 5'UTR sequence contexts	82
Table 3.4: Minor allele frequencies for all PheWAS-significant variants tested in discovery and replication analyses.....	82
Table 3.5: PheWAS replication analyses phenotypes tested	83
Table 3.6: 5'UTR Fragments used in expression constructs.....	83
Table 3.7: Nominal cardiac and movement disorder associations with SHMT2 stop-strengthening variant uncovered through PheWAS in Penn Medicine Biobank.	85

LIST OF FIGURES

Figure 2.1: Canonical G-quadruplex motif.	12
Figure 2.2: Allele frequencies for G-run disrupting variants in gnomAD. ...	16
Figure 2.3: pG4-forming G-tracts are depleted of polymorphic sites compared to intervening sequences.	17
Figure 2.4: Mutability-adjusted proportion of singletons (MAPS) for each set of variants affecting trinucleotide guanines within the meta-pG4 sequence motif.	19
Figure 2.5: Distribution of 5' and 3' UTR transcript isoforms with constitutive or non-constitutive pG4 sequence motifs.	21
Figure 2.6: Median expression (TPM) of each pG4-transcript or non-pG4 transcript was assessed for each tissue context.	23
Figure 2.7: Analysis of frequency of variants in UTR pG4 also being GTEx cis-eQTLs compared to non-pG4 UTR variants.	25
Figure 2.8: Annotated cis-eQTLs affecting 3'UTR pG4 sequences tend to increase gene expression.	25
Figure 2.9: Density of RBP-binding sites per kilobase of pG4 sequence compared to non-pG4 regions of the UTR.	27
Figure 2.10: Enrichment of specific protein–pG4 binding sites using CLIP-seq data from ENCODE.	29
Figure 2.11: UTR pG4 sequences are enriched for known pathogenic, and putative disease-associated genetic variants.	32
Figure 2.12: Distribution of binned distances for mapped canonical 5' (left) and 3' (right) UTR pG4 sequences with respect to protein-coding sequences across pG4-UTR containing mRNA transcripts.	36
Figure 2.13: Quantile-quantile plot showing matching between pG4 and non-pG4 containing transcripts based on LOEUF scores for 5'UTR (left) and 3'UTR (right) transcripts.	36
Figure 2.14: The empirical distribution of observed vs. expected number of substitutions across 10,000 bootstrapped 5' and 3' UTR regions in the European subpopulation of the 1000 Genomes Project Phase 1 release. ...	37
Figure 2.15: MAPS scores for All (black), G4 (red), and non-G4 GGG/CCC (blue) variants across multiple gene sets.	38
Figure 2.16: Distribution of odds ratio with error bars representing 95% confidence interval for changing gene expression of a pG4 containing gene	

versus a non-pG4 containing gene with sh-RBP knockdown in ENCODE as determined by Fisher's Exact Test.	39
Figure 2.17: Boxplot of distribution of distances between 5' UTR (a) and 3'UTR (b) pG4 sequences and nearest annotated protein-coding exons in ClinVar disease-associated genes showing the 1.5 times the interquartile range and median values.....	40
Figure 2.18: Additional common SNPs in high LD ($r^2 > 0.85$ in the 1000 Genomes GBR population) with GWAS tag SNPs exhibiting evidence of allelic imbalance in UTR pG4 sequences.	41
Figure 3.1: Basic experimental workflow for ribosome profiling.	44
Figure 3.2: Stop-introducing and stop strengthening mutations in translated uORFs are under strong negative selection.	50
Figure 3.3: Translated uORFs tend to use weak stop codons.....	51
Figure 3.4: ncORFs do not exhibit strong selective pressure to maintain amino acid identity.	55
Figure 3.5: Synonymous and missense variants in translated uORFs are under selective pressure to maintain codon optimality.....	56
Figure 3.6: Start codon usage for uORFs mapped by ribosome profiling...	58
Figure 3.7: MAPS scores for uORF start disrupting variants.	59
Figure 3.8: PhyloP estimates for possible start codon disrupting positions.	60
Figure 3.9: Phenome-wide association study (PheWAS) of predicted stop-strengthening variant in a translated uORF in PMVK.	63
Figure 3.10: Reporter gene assays for translated uORF stop-introducing and stop-strengthening variants.....	65
Figure 3.11: Distribution of protein coding ORFs, uORFs, and other non-canonical ORFs mapped by ribosome profiling from Ji et. al paper [121]. .	72
Figure 3.12: MAPS scores for uORF UTC-creating and stop-strengthening variants compared to non UTC-creating or stop-strengthening uORF variants matched by trinucleotide mutation context.....	73
Figure 3.13: PhyloP scores for possible UTC-creating positions and gnomAD protein-coding constraint.	73
Figure 3.14: Optimality changing MAPS scores for SNVs in dORFs (3'UTRs), pseudogenes, and long-noncoding RNAs (lncRNAs).	74
Figure 3.15: MAPS score for optimality changing variants using different optimality scores.	75

Figure 3.16: PheWAS plot of VPS53 stop-strengthening variant.	76
Figure 3.17: Change in PMVK 5'UTR annotation as of September 2019 Gencode 32 release.	77
Figure 3.18: PheWAS plot of BCL2L13 stop-strengthening variant.	77
Figure 3.19: PheWAS plot of NALCN UAA UTC variant.	78
Figure 3.20: PheWAS plot of SHMT2 stop-strengthening variant.	79
Figure 3.21: PheWAS plot of MOAP1 UAA UTC variant.	80
Figure 3.22: Luciferase experiments for PMVK and VPS53 plasmid Constructs showing similar direction of effect for UTC and stop- strengthening variants using HeLa cells for transfection.	81
Figure 3.23: Sampling procedure to generate MAPS scores to model the proportion of uORFs where UTC or stop-strengthening variants are capable of having pathogenic consequences.	87
Figure 3.24: Relationship between the fraction of true LOF variants in ClinVar pathogenic genes and MAPS scores adjusted for uORF- synonymous variants baseline.	88
Figure 3.25: Relationship between the fraction of true missense variants in all protein-coding genes and MAPS scores adjusted for uORF-synonymous variants baseline.	89

CHAPTER 1: THE REGULATORY RENAISSANCE

1.1: How many genes?

In 1999 the New York Times published estimates from scientists at the Incyte Corporation that the human genome contained approximately 140,000 genes. This represented a ~40% increase over the consensus estimate by many researchers at the time, with most previous estimates falling between 50,000-100,000. Many assumed that more complex organisms, like humans, necessitated a greater number of genes to support their biological complexity. As early as 1951 it had been observed that genome size appeared to increase with organismal complexity from invertebrates to vertebrates [1]. However, when the first working draft human genome was published in 2001, many were surprised to find evidence supporting the existence of only 30-40,000 protein coding genes [2]. Three years later, this estimate was further reduced to 21,000 - almost 20% less than in the zebrafish genome (~26,000) [3,4]. The paucity of protein coding genes in the draft human genome challenged the view that more complex organisms encoded more proteins in their genomes, and raised an important question: What was the source of human biological complexity?

The discrepancy between perceived organismal complexity and the number of protein coding genes came to be known as the "G-value Paradox". One proposed resolution to this paradox hypothesized that biological complexity more likely arose from a milieu of regulatory interactions between a limited set of protein-coding genes rather simply encoding more genes in the genome. These ideas were pioneered by Roy Britten and Eric Davidson in a 1969 essay published in Science [5], writing that:

Nonetheless, it seems unlikely that the 30-fold increase from poriferan to mammal can be attributed to a 30-fold increase in the number of producer [protein-coding] genes ... Quite possibly, the principal difference between a poriferan and a mammal could lie in the

degree of integrated cellular activity, and thus in a vastly increased complexity of regulation rather than a vastly increased number of producer genes.

We now recognize that substantial biological complexity arises through multiple layers of regulation - from the intracellular processes that govern gene expression and alternative splicing, to tissue-level and whole-organism levels of regulation that contribute to physiology and pathology. Many diseases result from otherwise protective physiologic processes - including fibrosis, inflammation, and immune activation - that become co-opted and dysregulated to produce pathology. Understanding the contribution of our genetics to how these processes become dysregulated can not only help expand our appreciation of disease biology, but also inform new targeted approaches towards therapeutic development.

1.2: Genetic variation and common disease

A crucial motivation for sequencing the human genome was to better understand the relationship between genes and disease. Shortly after the first draft human sequence was completed, the International HapMap was formed to create a database of common single nucleotide polymorphisms (SNPs) across diverse human populations [6]. The formation of this database facilitated a wave of genome wide association studies (GWAS) through which researchers sought to associate common genetic variants with numerous human phenotypes [7]. A typical GWAS required genotyping large cohorts of individuals at 300,000 - 5 million SNPs selected for their ability to provide broad coverage across the human genome by being in linkage disequilibrium with as many nearby genetic variants as possible [8]. Each genotyped SNP therefore represented a block of commonly co-inherited genetic variants, all in linkage disequilibrium. Association tests were then performed between genotyped SNPs and phenotypes of interest to determine whether specific alleles at these variants were. Phenotypes of interest could include true disease cases

versus controls, or with continuous traits like height or QT interval. When statistically significant associations were uncovered, discriminating between one or multiple true causal variants among co-inherited blocks of SNPs required additional fine-mapping and functional studies. Since most genetic variants uncovered through association studies resided in noncoding regions of the genome, the relationship between these variants, the genes they impacted, and how they related to the phenotypes being studied were rarely immediately obvious [8].

Even when GWAS lead variants fell within the boundaries of annotated genes, the possibility of long-range interactions made simple approaches including connecting the biological impact of tag GWAS SNPs to their nearest genes dubious. In a well-known example, a functional SNP associated with lactose intolerance (rs4988235) was found to reside within an intron of the minichromosome maintenance complex component 6 (*MCM6*) gene. Although it was known that expression of the lactase (*LCT*) gene was crucial for maintaining lactose tolerance, rs4988235 was located 13.9 kilobases upstream of *LCT* [9]. Despite being in the intron of *MCM6*, subsequent molecular studies showed that rs4988235 was indeed capable of modulating *LCT* expression by disrupting a distal enhancer element that is active only in lactase-producing enterocytes [10]. Thus, although GWAS were highly effective at uncovering sets of candidate variants associated with common diseases, identifying the true subset of causal genetic variants, and understanding how these variants connected specific genes to disease phenotypes remained challenging, particularly for disease contexts where no clear connection could be made between gene function and disease pathology.

Indeed, over 90% of disease-associated genetic variants identified through genome-wide association studies occupy noncoding regions of the genome [8]. These variants are broadly hypothesized to disrupt regulatory processes important for controlling normal biological functions, however interpreting their potential biological impact has remained challenging. In protein-coding DNA sequences, biological information is encoded in the form of codons that directly correspond

to amino acids. In contrast, noncoding regions of the genome can contain regulatory information that is often encoded as DNA sequence motifs. Making informed hypotheses on the mechanistic impact of genetic variation in noncoding DNA - in particular small insertions, deletions, or substitutions - is challenging both because much of the noncoding genome may not be functional, because many biological motifs are degenerate, and the specific mechanisms by which genetic variation impacts the activity of functional noncoding motifs are often unclear [11]. Thus, the promise of identifying specific biological mechanisms relating genetic variation to disease remained difficult for the vast majority of GWAS-uncovered SNPs.

1.3: The challenge of interpreting noncoding variation

Nevertheless, prior to the age of GWAS, traditional molecular biology approaches characterized numerous examples of noncoding genetic variants affecting DNA regulatory elements, including transcription factor binding sites [12], distal-acting transcriptional enhancers or silencers [13], and epigenetic regulatory marks that drive the pathogenesis disease [14]. Early studies of the β -globin gene were the first to demonstrate that genetic variants disrupting noncoding regulatory DNA could cause human disease [15]. It had been known previously that large deletions affecting the β -globin gene produced thalassemias in patients. Unexpectedly, two patients were found to also manifest clinical and histological symptoms of classic β -thalassemia despite having a completely intact β -globin gene [16]. Restriction mapping revealed large deletions in these patients occupying noncoding DNase I hypersensitive sites far upstream of the β -globin gene. Further molecular studies elucidated that deletion of these DNase I hypersensitive sites led to loss of β -globin expression, confirming a role for noncoding DNA in regulating gene expression, and also directly linking the disruption of noncoding regulatory DNA elements far upstream of encoded proteins to disease phenotypes [17].

The recognition that noncoding genetic variants could affect gene expression, sometimes over long ranges, provided one avenue for linking noncoding variants to their impact on genes, and ultimately genes to disease. The ability to capture global profiles of gene expression from biological samples through microarrays and high-throughput sequencing facilitated association studies between catalogued common genetic variants and changes in gene expression. Common variants found to be significantly associated with changes in gene expression, or expression quantitative trait loci (eQTLs) could now be mapped and identified for all 20,000 protein coding genes. The largest of these studies, the Genotype-Tissue Expression (GTEx) project, accumulated genotype, and gene expression profiles for thousands of individuals across 49 tissues [18,19]. These eQTLs could provide a putative link between changes in noncoding DNA and the activity of specific genes, and generate new hypotheses linking genes and disease. Similar approaches were developed for other measurable quantitative molecular traits - including levels of alternative splicing, degrees of chromatin accessibility, DNA methylation, protein expression, and metabolites among others [20]. Although these approaches could associate noncoding genetic variants with molecular phenotypes in cells, ascertaining precisely how these variants could change biological mechanisms remained elusive.

1.4: Untranslated regions in messenger RNA

While there is growing appreciation for the impact of noncoding genetic variation in human disease, most previous work has focused on understanding the impact of this variation on regulatory elements in DNA; in contrast, genetic variants affecting RNA have received comparatively less focus. As genetic information in the genome is first transcribed from DNA to RNA, and then translated from RNA to protein, understanding how genetic variation impacts RNA function is central to forming a more complete picture of how genetic variation can affect protein expression.

The 5' and 3' untranslated regions (UTRs) are core components of all mature messenger RNAs (mRNAs) that contain regulatory elements controlling diverse post-transcriptional processes. Indeed, it has been observed in yeast that the majority of the variance in mRNA stability can be explained by cis-regulatory elements in UTRs and coding sequences [21]. Although these distinctions are not absolute, the 5'UTR is broadly thought to influence mRNA translation by modifying ribosome loading and the efficiency of translation initiation through structural elements such as hairpins [22], internal ribosome entry sites (IRES) [23], and sequences capable of initiating translation which include cognate or near-cognate start codons [24], and long repetitive trinucleotide repeats [25,26]. In contrast, regulatory elements in 3'UTRs - including secondary structure forming motifs and microRNA binding sites -are more broadly thought to modify post-transcriptional mRNA stability, determine subcellular localization, and influence ribosomal recycling through interactions with the 5'UTR [27]. Moreover, sequence motifs in both 5' and 3' UTRs can serve as substrates for RNA binding proteins that greatly expand the repertoire of possible post-transcriptional RNA interactions impacting how and when mRNAs are translated or degraded. Thus, although genetic variants in both 5' and 3' UTRs have the capacity to affect protein expression through diverse mechanisms, dissecting which UTR variants can functionally impact biology from those that are silent remains challenging.

Examples of genetic variants in UTRs causing disease have been identified by traditional linkage mapping approaches with experimental validations. Well-known functional elements within UTRs with disease relevance include 3'UTR polyadenylation signals [28], microRNA binding sites [29], and upstream open reading frames (uORFs) in 5'UTRs which modify translation [30]. Yet, broadly understanding how mutations in these elements can affect the biological processes which lead to pathology is difficult even when the functional impact of a mutation on a regulatory element is known. As an example, systematic mutagenesis of the canonical polyadenylation signal AAUAAA has elucidated the functional effects of all possible mutations on the motif's function using synthetic 3'UTR constructs [31]. Yet, because 3'UTR regulatory elements are often repeated and

can act synergistically [32], the ultimate effect of mutations in one of several potential polyadenylation signals within a 3'UTR cannot be known without performing targeted biochemical studies. A consequence of the ambiguity inherent in interpreting noncoding genetic variation is that most known pathogenic variants are in the coding genome, even though many exome sequencing approaches can capture variation in UTRs. Because of this, expanding our capacity to interpret genetic variation in both 5' and 3' UTRs can have an immediate and direct impact in clinical applications and our understanding of disease biology.

1.5: Identifying variants of interest in the noncoding genome

The falling cost of sequencing has led to the creation of large databases of human genetic variation, facilitating new approaches to identify functional noncoding genetic variation. Under the principle that functional elements within the human genome tend to be less tolerant to genetic variation, identifying regions of the genome that are depleted of variation can help unmask functional regulatory elements in noncoding DNA. Many studies have analyzed allele frequencies to infer the action of natural selection on putative and predicted noncoding regulatory elements throughout the genome. Broadly, allele frequencies in an isolated population are shaped by three forces: mutation, drift, and selection. Genetic variants are introduced into the population through random mutation - a process that tends to introduce variation throughout the genome at a nearly constant rate [33]. Under the neutral theory of molecular evolution, most mutations have neutral or negligible effects on an organism's fitness and therefore will either be randomly removed from the population, or rarely, will become fixed in a population due to genetic drift [34]. Even more rarely, a mutation can be beneficial to an organism's fitness, and selection may cause its allele frequency to increase within a population in a process known as positive selection. A hard sweep occurs when a beneficial mutation is introduced into a population by random mutation and rapidly reaches fixation [35]. Alternatively, positive selection may manifest as a soft sweep - when

changes in the environment make a previously neutral variant present within the population beneficial and its allele frequency increases [36].

In contrast, when a mutation has deleterious effects on an organism's fitness, negative, or purifying selection, tends to remove it from the population [33]. Thus allele frequencies for deleterious variants tend to remain low within a population, and by extension deleterious classes of mutations - such as protein truncating mutations, or mutations affecting essential splice sites - will exhibit enrichment in rare allele frequencies compared to classes of genetic variation that are selectively neutral [37]. This relationship between selection and allele frequencies can be used to explore hypotheses and identify new classes of functional genetic variation in the human genome.

Prior to the generation of large databases of human genetic variation, putative noncoding regulatory elements were identified by comparing genome alignments from multiple species. Many early approaches relied on detecting non-neutral substitution rates over particular genome segments compared to expectation based on genetic drift [38]. Although effective at identifying long stretches of conserved sequences in the human genome, the power of these methods to detect non-neutral substitution rates at the individual nucleotides was limited by comparison, particularly at moderately conserved bases [38]. By comparing strongly conserved sequences across 29 different mammals, Lindblad-Toh and colleagues reported a map of human constrained elements which also appeared to be depleted of single-nucleotide polymorphisms in human data [39]. Thus, mutations affecting highly conserved noncoding DNA also appeared to be strongly depleted of variation in human populations [39]. The increasing availability of sequenced human genomes made it possible to implement approaches for inferring selection on specific genetic variants or groups of nucleotide positions in the human genome based on their allele frequency spectrum. Because functional regulatory elements in noncoding DNA are likely under a greater degree of selective constraint compared to neutrally evolving DNA segments, the site frequency

spectrum of variants with deleterious impacts on these functional elements are also expected to be enriched for rare allele frequencies.

Studies inferring the importance of microRNA regulatory sites in humans were among the first to apply the relationship between selection and allele frequencies in human genomes. Using allele frequencies from the 1000 Genomes Project Chen et al. found that SNPs affecting computationally predicted conserved microRNA binding sites throughout the genome were enriched for rare allele frequencies compared to other conserved sequence motifs in 3'UTR sequences, suggesting that these sites were under a greater degree of negative selection [40]. By analyzing SNP density and the site frequency spectrum of genetic variants affecting ~22,000 predicted microRNA binding sites conserved across 5 mammals, Chen and colleagues uncovered a significant depletion of genetic variation affecting these elements, suggesting that they were also more likely to be functionally important in 3'UTRs. Similar approaches have been used to examine patterns of selection on RNA-protein binding sites in RNAs [41], long-noncoding RNAs [42], and sites of m6A methylation in human mRNAs [43].

As the volume of sequencing data has increased, subsequent refinements have been made to this approach, most notably with the publication of the ExAC database in 2014 [37]. Here, Lek et al. developed a method to quantify the enrichment of rare allele frequencies for a given class of genetic variation using synonymous coding variants as a baseline for neutral selection, while adjusting for different rates of mutation based on local sequence context. This metric - termed the Mutability Adjusted Proportion of Singletons (MAPS) - has been applied to elucidate new classes of putatively functional genetic variation by identifying groups of variants under stronger negative selection compared to synonymous variants in coding regions of the genome. Indeed this approach has been applied to identifying deleterious classes of coding mutations [44], new putative splice-site disrupting mutations [45], and genetic variants creating new upstream open reading frames in mRNA 5'UTRs [46]. A key advantage of this approach is that biologically-

informed hypotheses made about the potential functionality of a given class of genetic variants can be evaluated by analyzing their allele frequency spectrum in large population-scale databases.

Here, we have applied metrics of negative selection across human cohorts to study noncoding regulatory elements in RNAs. To illuminate possible mechanistic relationships between genetic variation and gene regulation, we take a hypothesis-driven approach to assess the significance of specific classes of variation affecting both predicted and experimentally mapped RNA regulatory elements. We have additionally used a combination of public genotype and phenotype databases to explore the relevance of these variants to human disease. We have selected putative regulatory elements based on the availability of experimental data to support their existence across a large fraction of human mRNAs, and previously published literature supporting their potential functionality. The focus of this thesis is twofold: In Chapter 2, we investigate whether there is evidence for a functional role of G-quadruplex forming sequences in mRNA UTRs. In Chapter 3, we employ the same investigation framework to non-canonical open reading frames (ncORFs). Using public repositories of ribosome profiling to identify ncORFs, we elucidate selective pressures acting within these translated noncoding sequences and identify new patterns of functional variation in upstream open reading frames. For a subset of the variants we identified as functional and associated with disease, we performed luciferase assays to validate their effect on the translation of the downstream gene. Importantly, beyond the functional variants identified in this work, these studies represent a new approach to study and assess the impact of genetic variation on cis-regulatory elements in mRNA UTRs.

CHAPTER 2: INTEGRATIVE ANALYSIS REVEALS RNA G-QUADRUPLEXES IN UTRS ARE SELECTIVELY CONSTRAINED AND ENRICHED FOR FUNCTIONAL ASSOCIATIONS *

2.1: Secondary structures as RNA regulatory elements

Unlike DNA, the single-stranded nature of RNAs significantly expands the possibility for intrastrand base-pairing. RNA secondary structures have long been implicated in regulating gene expression through diverse post-transcriptional mechanisms. Strong secondary structures in 5'UTRs are known to repress translation of downstream coding sequences by blocking the formation of ribosome translation initiation complexes [22,47]. Specific 5'UTR secondary structures have also been observed to facilitate increased translation of downstream protein coding sequences through serving as internal ribosome entry sites (IRES) [23], or possibly through blocking translation initiation at inhibitory upstream open reading frames (uORFs) [22]. In contrast, 3'UTR secondary structures have been found to mask microRNA binding sites [47,48], facilitate interactions with RNA binding proteins [49,50], and regulate mRNA stability [51], or subcellular localization [52–54].

Several experimental approaches have been developed to map secondary structures transcriptome-wide [55–57]. These studies have generally uncovered an enrichment of secondary structures in both 5' and 3' UTRs compared to protein coding sequences where strong structure formation could disrupt translation elongation by ribosomes [58]. RNA secondary structures can also facilitate interactions with RNA binding proteins - either through helicases that unwind and

*Published as Lee D.S.M. et al. Nat. Commun. 2020. [138]

resolve secondary structure elements in the 5'UTR to facilitate translation initiation, or through binding proteins which can mediate RNA subcellular localization [59].

2.2: G-quadruplexes are non-canonical secondary structures formed by nucleic acids

Guanine rich nucleic acid sequences can form non-canonical secondary structures known as G-quadruplexes (G4s) in both DNA and RNA [60]. In contrast to DNA G4s, RNA G4s are thought to form more readily in vitro due to their increased thermodynamic stability and reduced steric hindrance [61,62]. G4 secondary structures are formed through non-canonical base pairing of guanine side chains in G-rich sequences. The canonical G-quadruplex forming sequence consists of four trinucleotide G-runs separated by 1-7 nucleotides (Fig. 2.1).

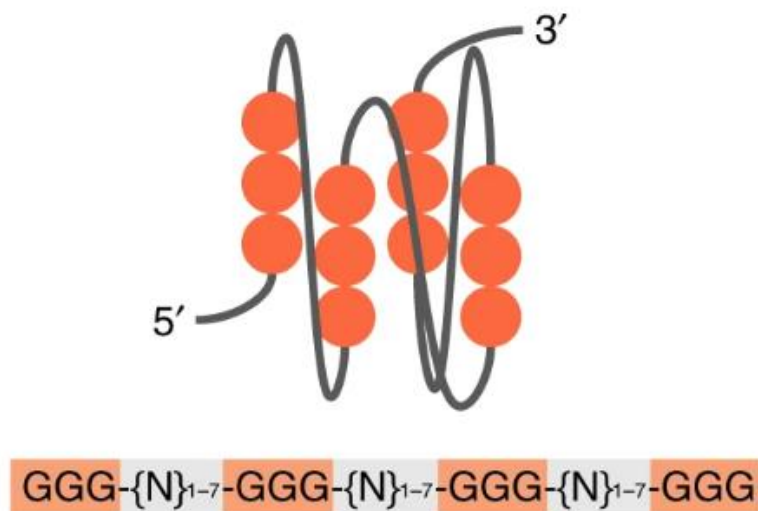


Figure 2.1: Canonical G-quadruplex motif. Schematic depicting a folded RNA parallel G-quadruplex with the accompanying canonical G4 forming sequence.

Transcriptome-wide G4 mapping studies have uncovered evidence for widespread G4 formation in both 5' and 3'UTR sequences [63,64]. In one approach, Kwok and colleagues developed a method to map transcriptome-wide G4 structures (rG4) by measuring reverse transcriptase stalling [63]. The method relies on isolating cellular mRNAs *in vitro*, while creating ionic conditions that are favorable to RNA G4 formation. These RNAs are then treated with reverse transcriptase (RT) to produce cDNA fragments which are isolated and subsequently sequenced. Because G4s will induce RT-stoppage due to steric blockage, evidence of significant RT stopping can be used to map secondary structures and identify G4 forming sequences within mRNAs transcriptome-wide [63]. Results from these experiments uncovered specific enrichment of G4 structures in 5' and 3'UTRs, and evidence that the capacity for G4 structure formation in mRNAs extends beyond the canonical G4 motif to include structures with extended loops, bulges in G-runs, and two quartets [63].

Although G4 formation has been studied extensively *in vitro*, whether G4s in mRNAs exist *in vivo* has been an active area of debate. The single-stranded nature of RNAs is thought to favor G4 formation due to reduced steric hindrance, however some *in cellulo* structural probing experiments using RT stopping have suggested that G4s typically exist unfolded at steady-state in most eukaryotic cells [65]. Guo et al. modified the RT-stopping approach to map RNA sequences capable of forming G-quadruplexes *in vivo* to probe for G4 formation *in cellulo*. Strikingly, they observed that while most predicted RNA G4-forming (pG4) sequences were unfolded at the steady-state, these same RNAs could form secondary structures when expressed in prokaryotes, suggesting that eukaryotic cells harbored factors capable of unwinding these RNA secondary structures [65].

While specific RNA G4s have been associated with diverse biological functions, including mediating translational control [66,67], alternative splicing [68], subcellular localization [69], and RNA stability [70,71], the transcriptome-wide functional importance of UTR G4s has largely been

extrapolated from a limited number of experimental studies. To address this question, we combine several large-scale genomic and genetic data resources to assess evidence for evolutionary constraint on UTR pG4 sequences in humans, and enrichment for functional associations, including cis-eQTLs and protein binding sites. We show that UTR pG4 sequences are subjected to heightened selective pressures, have enrichment for cis-eQTL variants as identified by GTEx, and enrichment of RNA-protein binding interactions mapped by ENCODE. Taken together, our results support the biological significance of UTR pG4 sequences and highlight the importance of considering secondary structures in determining biological function in noncoding regions of the genome.

2.3: pG4 exhibit heightened selective pressure within UTRs

Putative G-quadruplex (pG4) forming sequences are enriched within untranslated regions of human messenger RNAs [63]. If these sequences are functional, they should exhibit patterns of genetic variation consistent with heightened evolutionary constraint. To test this hypothesis, we evaluated the distribution and frequency of single nucleotide variants occurring within UTR pG4 sequences using whole-genome sequencing data from over 15,000 individuals from the public gnomAD release (version 2.2.1) [44]. We mapped pG4 sequences transcriptome-wide within annotated UTRs using the canonical G4 motif - GGG-{N-1:7}(3)-GGG (Fig. 2, 1). Consistent with previous UTR G4 mapping efforts [72], we identified 2967 unique protein-coding genes encoding for at least one transcript isoform containing a pG4 sequence within the 5'UTR, and 2835 protein-coding genes encoding a pG4 sequence within the 3'UTR. To further increase the specificity of pG4 sequences, we additionally defined a subset of experimentally supported rG4 sequences (466 in the 5'UTR, 1743 in the 3'UTR), consisting of canonical pG4 sequences with evidence of secondary structure formation as determined by biochemical structure mapping approaches [63]. Under the expectation that deleterious variation is continuously removed from the population, we

expect allele frequencies for variants affecting UTR pG4 sequences to be skewed towards more rare variation compared to non-pG4 UTR variants, reflecting their greater functional importance [40,73,74]. Because allele frequencies throughout the genome are affected both by local sequence context, which influences the mutability of a base at a given position, and nearby constrained functional elements that are under linked selection, we compared only single nucleotide variants affecting pG4 G-tracts to non-pG4 G-tracts (3 or more Gs) within UTRs belonging to a subset of transcripts whose estimated levels of overall constraint matched our UTR pG4-containing transcripts. This set of comparator transcripts was selected using the upper 90% bound of the observed vs. expected (LOEUF) metric, as published by gnomAD [44]. This analysis revealed a significant depletion of variants in pG4 and rG4-seq G4s (Fig 2.2). For rG4 sequences, we found mean allele frequencies were approximately one-third of that compared to non-pG4 G-tracts in constraint-matched transcripts in the 5'UTR, and 30% lower for the 3'UTR. For pG4 sequences without direct experimental support, G-tract variant frequency differences were similarly reduced ($P < 2.2 \times 10^{-16}$ for 5'UTR and 3'UTR; Fisher's Exact Test). Taken together, this reduction in mean allele frequencies for variants in 5' and 3'UTR pG4 sequences relative to those not affecting pG4 sequences is consistent with the effects of negative selection.

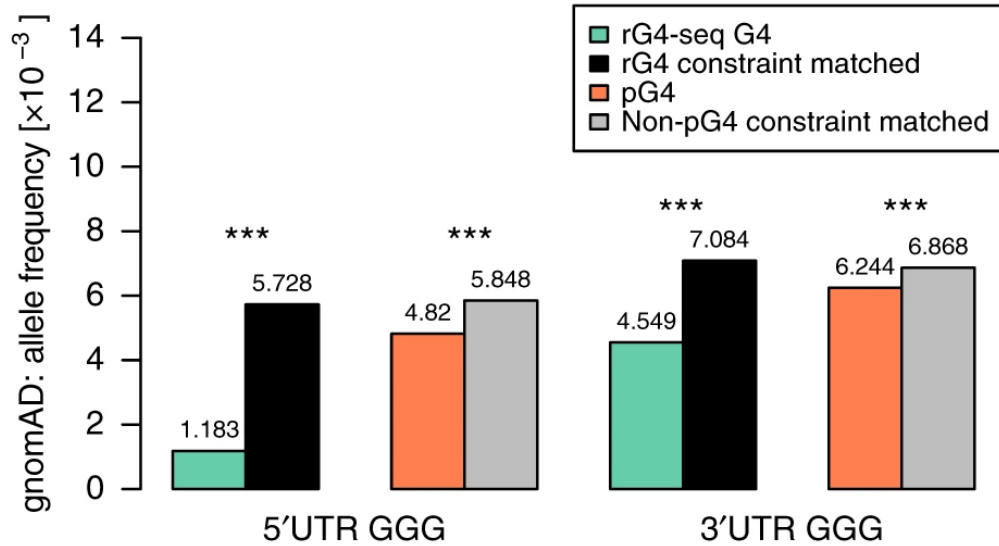


Figure 2.2: Allele frequencies for G-run disrupting variants in gnomAD. Reduction in variant frequencies affecting guanine G-tracts within UTR pG4 forming sequences compared to matched non-pG4 G-tracts by transcript-level constraint. rG4-G-tracts are those within UTR pG4 that have evidence of secondary structure formation by rG4-seq. Asterisks denote P-value $<<2.2 \times 10^{-16}$ by Fisher's exact test.

To provide a complementary measure of sequence constraint, we assessed the number of polymorphic sites within UTR pG4 sequences compared to non-pG4 sequences. We applied a background model of neutral evolution to produce a distribution for the expected number of polymorphic sites in a given region of the genome under the assumption of neutral selection. This model has been shown to explain a median of 81% of the variability in nucleotide substitution probabilities for noncoding regions of the genome based on the local heptamer context of a given position [75]. Using this model, we partitioned UTR pG4 sequences into G-tracts and intervening gap sequences, and compared the ratio of *observed* versus *expected* polymorphic sites in the European sub-population of the 1000 Genomes Project. To additionally control for the possible confounding effects of linked selection driven by nearby constrained coding elements, or differences in sequencing depth across the 1000 Genomes Project, we produced an empirical distribution for observed vs. expected substitutions in constraint-matched 5' and 3'UTR sequences. Consistent with the observed reduction in variant frequencies across UTR pG4s, we

find a significant reduction in the number of observed versus expected polymorphic sites within UTR pG4 sequences compared to non-pG4 forming regions of the UTR. Relative substitution rates in 5' and 3' UTR G-tracts are reduced approximately 30-40% compared to non-pG4 regions of constraint-matched UTRs (permuted $P < 10^{-4}$, for 5' and 3' UTR pG4 and rG4) - Fig. 2.3. In contrast, gap sequences that are not predicted to be important for secondary structure formation in either 5' or 3' UTR pG4 contexts are not significantly different from the background UTR estimates, consistent with a pattern of selective pressure in 3' UTR pG4 sequences that primarily act to maintain the capacity for secondary structure formation across UTR pG4 sequences.

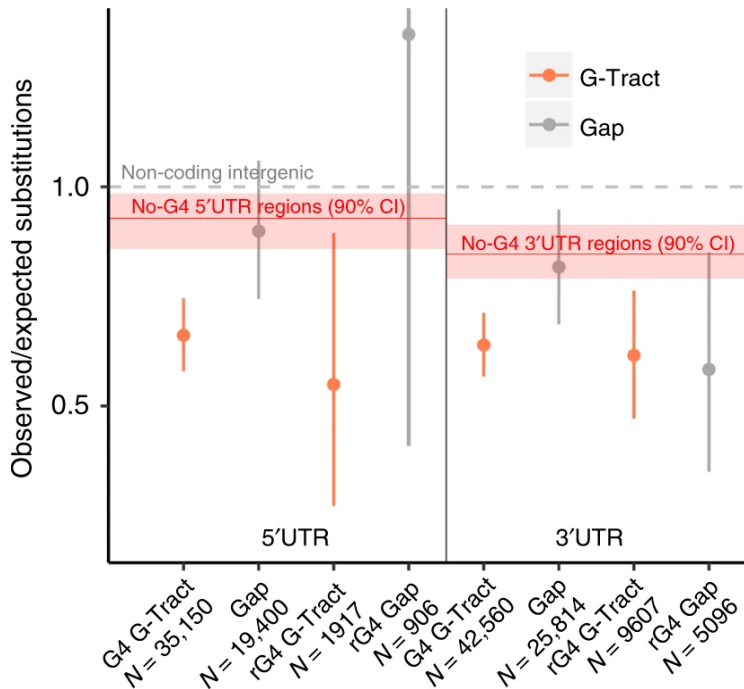


Figure 2.3: pG4-forming G-tracts are depleted of polymorphic sites compared to intervening sequences. Reduction in the number of observed polymorphic sites compared to expectation in 5' and 3' UTR pG4 forming G-tracts using a nucleotide substitution model based on local sequence context (permuted $P < 1 \times 10^{-4}$ in all G-tracts compared to matched non-pG4 UTR sequences). Error bars represent bootstrapped 90% confidence intervals for the ratio of observed vs. expected substitutions within each pG4 region. Red line and shaded regions represent the observed vs. expected number of substitutions in non-pG4 UTR sequences

matched by transcript-level constraint and 90% confidence intervals, respectively. Gray-dashed line represents an expected vs. observed ratio of 1:1.

The reduction in allele frequencies, and in the number of polymorphic sites within UTR pG4 sequences, indicate UTR pG4 are under heightened selective pressures compared to non-pG4 UTR regions. To place the degree of selection on UTR pG4s in context, we applied a mutability-adjusted proportion of singletons (MAPS) metric, which measures the relative enrichment for rare variation within a particular class of variants accounting for differences in mutation rates based on local sequence context [37]. A similar approach has been recently used to assess the degree of selective pressure against upstream open-reading frame-creating variation within 5'UTR variants in the gnomAD database [46]. Within the canonical pG4 motif, we predicted that variants affecting the central guanine of each G-tract should be most constrained, since biophysical studies of G4 stability have shown that mutations affecting the central tetrad (2nd guanine of each trinucleotide guanine repeat) are most detrimental to secondary structure stability [76]. To remove the ambiguity of which specific guanines are involved in secondary structure formation when more than three guanines form a pG4 G-tract, we focused only on single nucleotide variants within trinucleotide G-tracts ($n = 3137$). By examining variation across each pG4 G-tract, we found central guanine positions within UTR pG4 G-tracts are consistently enriched for singletons (one sequenced variant in gnomAD whole genomes) compared to non-pG4 UTR variants (Fig. 2.4, permuted $P < 10^{-4}$). Notably, non-pG4 UTR variants reflected a similar degree of constraint as synonymous coding variants, while central position guanines exhibit a similar degree of selective pressure as missense variation in protein-coding regions of the genome. Interestingly, the most proximal and distal 5' and 3' guanine of each trinucleotide pG4 G-tract demonstrated significantly less enrichment of singleton variants within gnomAD across gene classes compared to central guanine positions as determined by permutation testing (P -values = 0.0237 and 0.0022 respectively - Figure 2.15). This result suggests these positions are under less negative selection

compared to central positions, perhaps because mutations in these positions can preserve the potential for RNA to form non-canonical G4 2-quartets [63].

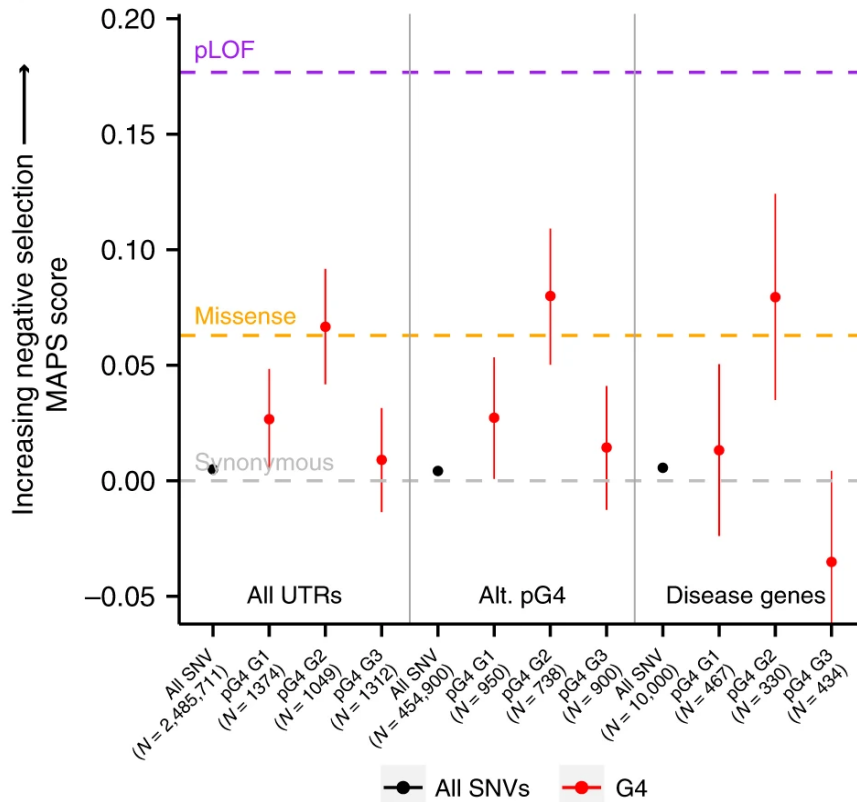


Figure 2.4: Mutability-adjusted proportion of singletons (MAPS) for each set of variants affecting trinucleotide guanines within the meta-pG4 sequence motif. Central position guanines consistently demonstrate the highest MAPS scores (are most constrained) compared to non-pG4 UTR variants (permuted $P < 1 \times 10^{-4}$) across all contexts. Error bars represent the 5% and 95% bootstrap permutations for each variant class. Purple-dashed line, orange dashed line, and gray-dashed line represent MAPS score for Ensembl predicted high-impact coding (predicted loss-of-function), missense, and synonymous mutations respectively.

Finally, to provide additional control for our sequence context-derived mutability rates, we compared the MAPS metric for UTR pG4 G-tracts to UTR trinucleotide G- and C-runs not involved in pG4 formation. Although these non-pG4 G- and C-tracts exhibit modest enrichment for rare variation at the central position, there is a significantly greater enrichment in singletons at the central position of the UTR pG4 G-tract compared to non-pG4-forming contexts (Figure 2.15 -

permuted $P=0.0195$). Thus, the excess rare variation is specific to the guanine within pG4 G-tracts most important for maintaining G4 secondary structure.

2.4: Most pG4 motifs in UTRs are isoform-restricted

Many functional UTR elements, including upstream open reading frames (uORFs), AU-rich elements, and microRNA binding sites are frequently included in alternative 5' or 3' UTR isoforms of the same gene [77,78]. Alternative UTR inclusion is hypothesized to significantly diversify the number of possible post-transcriptional regulatory interactions for a given gene [79]. Given the observed constraint over UTR pG4 sequences, we hypothesized that UTR pG4 sequences should also exhibit patterns of alternative inclusion or exclusion.

To evaluate the extent of alternative UTR pG4 inclusion, we mapped UTR pG4s to protein-coding transcripts for each gene in the Ensembl transcriptome database. Genes were considered to produce constitutive UTR pG4 sequences when all annotated protein-coding transcript isoforms contained at least one pG4, or alternative UTR pG4 sequences if at least one transcript isoform lacked the pG4 sequence. Most constitutive pG4 genes were found to express UTRs with identical pG4s across all transcript isoforms, however 36 of 620 5'UTR and 75 of 1275 3'UTR constitutive pG4 genes produced transcript isoforms with non-identical pG4 sequences. For this subset of non-identical pG4 transcript isoforms, approximately one-third differ by the addition / subtraction of pG4 motifs (13/36 for 5'UTR, 20/75 for 3'UTR). Strikingly, we found that over half of all genes producing UTR pG4 transcripts also encoded for alternative UTRs lacking pG4 motifs (2254 genes with 5'UTR pG4 motifs and 1425 genes with 3'UTR pG4 motifs - Fig. 2.5).

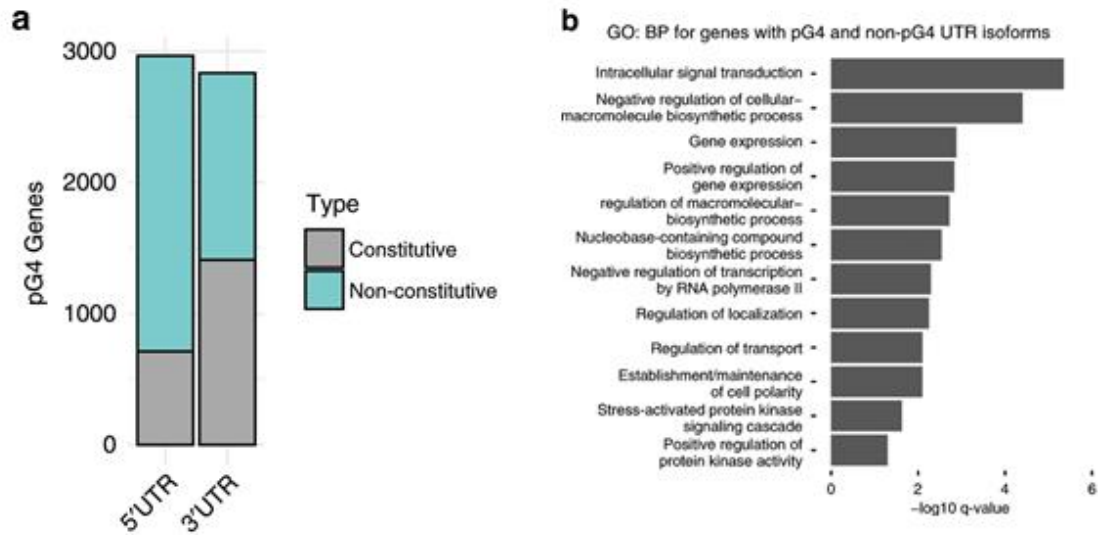


Figure 2.5: Distribution of 5' and 3' UTR transcript isoforms with constitutive or non-constitutive pG4 sequence motifs. (a) Most genes with mRNA transcripts with UTR pG4 sequences also produce alternative isoforms lacking UTR pG4s (non-constitutive). (b) Overrepresented biological processes for protein-coding genes producing both pG4 and non-pG4 5' or 3' UTR isoforms ($n = 3148$). GO-term enrichment was performed using PantherDB⁵⁵ and enrichment was determined by meeting a Benjamini–Hochberg adjusted P value cutoff of 0.05 by Fisher's exact test.

Indeed, of the 5235 total UTR pG4-containing genes, 3395 exhibited either alternative 5' or 3' UTR pG4 inclusion, and 284 produced UTRs with both alternative 5' and 3' pG4s. This distribution of alternative and constitutive pG4 genes for each UTR context was found to be highly significant through permutation testing (P -value < 0.0001 for 5' and 3' UTRs). Moreover, MAPS scores for alternative UTR pG4 indicate that their second guanine position is under a similar degree of constraint as for all UTR pG4 and is comparable to that of missense variations for the set of alternative pG4s found in genes with any disease association in ClinVar (<https://www.ncbi.nlm.nih.gov/clinvar/>) (Fig. 2.4). As is the case for all UTR pG4, this second guanine position was significantly more enriched for rare variation compared to either the 5' or 3' guanine (permuted P -value = 0.0138 and 0.004 respectively). Constitutive UTR pG4 sequences, in contrast, do not show a similar pattern of selective constraint acting on the second G-tract guanine, possibly because these sequences tend to be under less stringent selective pressures,

or because we are underpowered to detect significant enrichment in rare variation. Notably the MAPS metric for the central G-position of alternative pG4 sequence G-tracts remained significantly higher than matched, non-pG4 G-tracts (permuted $P=0.0124$ - see Figure 2.15 for comparison of constitutive pG4 G-tracts and other pG4 gene sets).

We next asked whether the expression of alternative pG4 isoforms tend to be restricted or shared across different tissue contexts. Using transcript-isoform expression data across 45 different tissues from GTEx, we find that many tissues appear to express both pG4 and non-pG4 transcripts simultaneously (Fig. 2.6). Notably, this simultaneous expression of both pG4 and non-pG4 isoforms also occurs in single-cell contexts (lymphocytes, fibroblasts), demonstrating that this effect is not due to cellular heterogeneity in bulk tissue samples. Thus, most UTR pG4-encoding genes express alternative isoforms which lack pG4 sequences, and that the simultaneous expression of both pG4-isoforms and non-pG4 isoforms is widespread across multiple tissue and cellular contexts.

To explore the functional associations of alternative UTR pG4 genes we performed a gene ontology analysis. We find that these genes are frequently involved in dynamic intracellular processes, including signal transduction, cellular responses to stress, and metabolic regulation (Fig. 2.5b). In contrast, constitutive pG4 genes showed enrichment for biological processes associated with the activation of gene expression in discrete temporal stages, including those involved in tissue development, pattern specification, and cellular differentiation. These observations, coupled with our finding that many tissues simultaneously express both pG4 and non-pG4 isoforms of the same gene, suggests that isoform-switching between pG4-containing or non-pG4 transcripts may facilitate dynamic cellular responses to external stimuli. More broadly, our results demonstrate considerable variation in alternative pG4 inclusion within UTRs across multiple tissue contexts, and suggest that the relative abundance of pG4 and non-pG4 UTRs may be dynamically regulated within a given tissue.

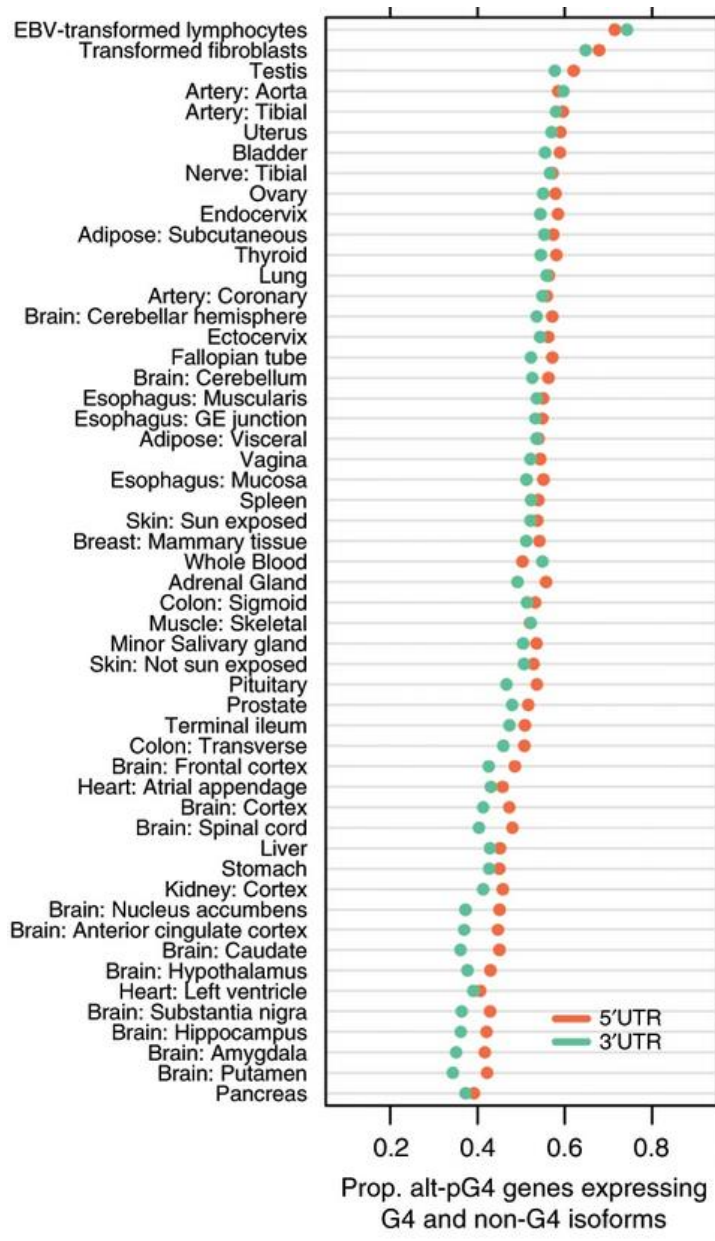


Figure 2.6: Median expression (TPM) of each pG4-transcript or non-pG4 transcript was assessed for each tissue context. For the subset of genes producing UTRs with alternative pG4 inclusion, both pG4-containing and non-pG4 isoforms are frequently expressed simultaneously. Transcripts were considered as expressed if their median TPM measurement exceeded one TPM for each tissue context considered. The proportion of pG4 genes expressing both pG4 isoforms, and non-pG4 isoforms was then compared for each tissue.

2.5: pG4 motifs in the 5' and 3' UTR are enriched for cis-eQTLs

We next evaluated the potential regulatory consequences associated with mutations affecting UTR pG4s, hypothesizing that variants affecting pG4 sequences might also be more likely to be associated with changes in gene expression. To test this hypothesis, we compared the proportion of annotated cis-eQTLs versus non-eQTL SNPs identified by GTEx across pG4 and non-pG4 regions of the UTR, finding significant enrichment for either nominally significant or lead eQTL variants (lowest P-value variant) in 5' and 3' UTR pG4 sequences compared to non-pG4 regions of UTRs (Fig. 2.7). Notably, we continue to observe an enrichment of cis-eQTL variants in UTR pG4 sequences using a reduced set of putatively causal cis-eQTLs [80], suggesting that disruption of UTR pG4 sequences may cause changes in post-transcriptional regulation.

We next explored the direction of gene expression changes for UTR pG4 cis-eQTLs, considering all variant-tissue effects separately for each significant variant-tissue interaction. We hypothesized that variants affecting pG4 G-tracts are more likely to disrupt the structural integrity of the RNA G4s, and thus might influence gene expression differently than variants affecting gap (non-G-tract) sequences within pG4 motifs. Since the magnitude of normalized effect-size estimates in GTEx has no direct biological interpretation, we compared differences in the direction of variant effects across pG4 and non-pG4 sequences. As expected, UTR variants in non-pG4 regions are not significantly biased towards increasing or decreasing gene expression, regardless of whether the mutation affected a G-tract, or non-pG4 G-tract nucleotide. In contrast, mutations affecting structurally important pG4 G-tracts in the 3'UTR tend to increase mRNA expression compared to non-G-tract bases (OR 1.75, 95% CI: 1.34 to 2.30, $P < 3.0 \times 10^{-5}$) - Fig. 2.8. This relationship for the 5'UTR was not observed. Given the role of the 3'UTR in mediating mRNA stability, the tendency for G-tract base mutations to increase gene expression suggests the involvement of 3'UTR G4s in decreasing mRNA stability.

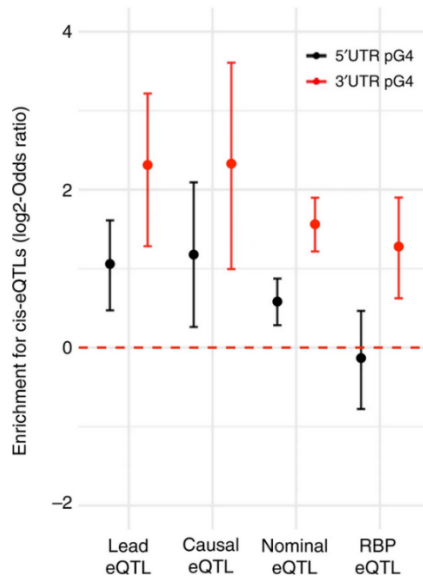


Figure 2.7: Analysis of frequency of variants in UTR pG4 also being GTEx cis-eQTLs compared to non-pG4 UTR variants. GTEx *cis*-eQTLs are enriched within UTR pG4 relative to the number of tested (non-eQTL) SNPs when comparing lead SNPs, high-confidence causal, nominally significant, and nominally significant in RBP-binding sites in matched UTR regions. Error bars represent the 95% confidence interval for the odds ratio.

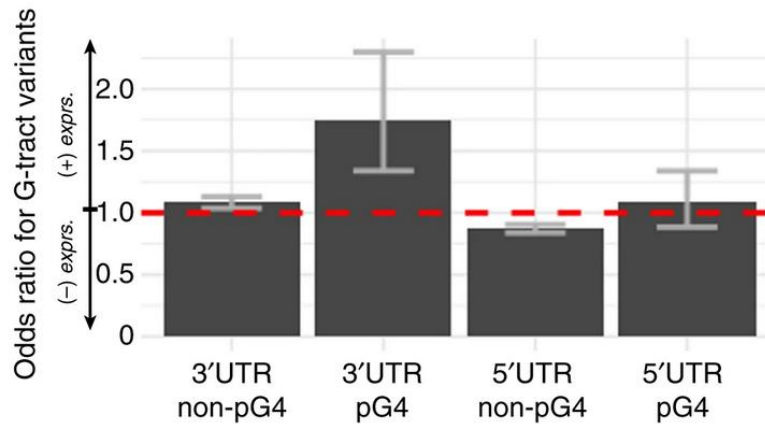


Figure 2.8: Annotated cis-eQTLs affecting 3'UTR pG4 sequences tend to increase gene expression. Odds ratio for a *cis*-eQTL increasing gene expression across all *cis*-eQTL-tissue effects ($n = 379,441$, P value $< 2e^{-16}$, Fisher's exact test), where the variant affects a pG4 G-tract compared to those affecting gap sequences. Error bars represent the 95% confidence interval for the odds ratio.

2.6: RNA-protein binding sites are enriched over UTR pG4 regions

Transcriptome-wide RNA structure mapping studies have suggested that most RNA G4 are unfolded in eukaryotes, but not in prokaryotes, leading to the hypothesis that intracellular factors bind RNA G4s to maintain their unfolded state in cellulo [65]. To gain insights into regulatory mechanisms mediating pG4 effects on gene expression we investigated the propensity of protein-binding sites to overlap UTR pG4s by comparing the proportion of UTR pG4 sequences overlapped by RNA-binding protein (RBP) binding sites published by ENCODE to non-pG4 forming regions of the UTR[81]. This data consists of cross-linking immunoprecipitation sequencing (CLIP-seq) peaks, called from K562 or HepG2 cell lines for over 150 RBPs, containing at least one highly reproducible (IDR = 1000) [82] binding peak within the 5' or 3' UTR. When compared to non-pG4 regions of the UTR, the frequency of overlap between unique (non-overlapping) RBP binding sites and pG4 sequences was almost 6-fold ($P < 2.2 \times 10^{-16}$, Chi-square test) higher compared to non-pG4 sequences in the 5'UTR (Fig. 2.9). Enrichment of RBP binding locations over pG4 sequences within the 3'UTR was markedly higher (14-fold, $P < 2.2 \times 10^{-16}$, Chi-square test). Given the enrichment within UTR pG4s for cis-eQTLs and protein binding sites, we tested for significant colocalization between these two features in pG4s. Taking the subset of pG4 regions overlapped by any protein binding sites, we examined the density of cis-eQTLs in UTR pG4 regions also overlapping CLIP-seq peaks. When all nominally significant cis-eQTLs are considered, we observe a significant enrichment of cis-eQTLs in the 3'UTR that are also protein binding sites (Fig. 2.7), indicating that variation in 3'UTR pG4 sequences may influence gene expression through changing RNA-protein interactions. Given the observed association between protein binding sites and pG4 sequences, we next asked whether specific proteins' binding sites are enriched for pG4s. For each protein, we determined the proportion of protein-specific binding sites containing pG4 sequences, against the total background rate of all CLIP-seq binding sites containing pG4 sequences. To determine a significant overrepresentation of pG4 sequences within a given protein's binding sites, we

performed a hypergeometric test against the null hypothesis that there is no overrepresentation of pG4 binding sites within the set of a protein's binding sites - Fig. 2.9.

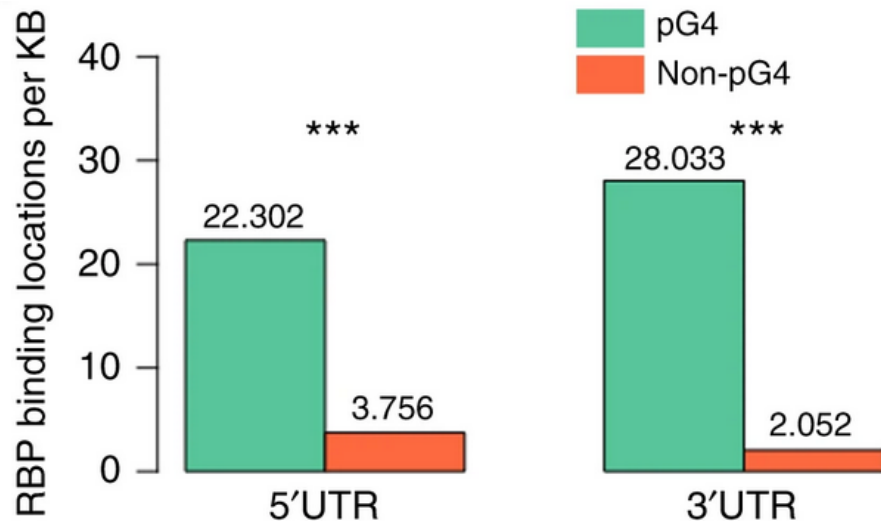


Figure 2.9: Density of RBP-binding sites per kilobase of pG4 sequence compared to non-pG4 regions of the UTR. pG4 sequences are more frequently overlapped by RBP binding sites compared to non-pG4 sequences within the UTR (P value $\ll 2.2 \times 10^{-16}$, chi-square test).

This analysis revealed enrichment for proteins that have been implicated in RNA G4 binding (GRSF1, FUS), and those that, to our knowledge, have not previously been associated with RNA G4 structures (PRPF4, GTF2F1, and CSTF2T). GRSF1 is a cytoplasmic protein involved in viral mRNA translation and has recently been shown to play a role in the degradation of G4-containing RNAs in mitochondria [83,84]. Other proteins with significant enrichment for pG4 binding include those involved in mitochondrial processes (FASTKD2), transcriptional activation (GTF2F1), mRNA transport (FAM120A), mRNA degradation (XRN2, UPF1), in addition to several proteins implicated in RNA polyadenylation and splicing (CSTF2T, PRPF4, RBFOX2), and surprisingly, micro-RNA (miRNA) biogenesis (DCGR8, DROSHA). Interestingly proteins demonstrating a preference for binding UTR pG4 sequences tend to bind both 5' and 3' UTR contexts, with 14 out of 20 proteins' binding peaks showing enrichment for overlap over 5' and 3' pG4 sequences in

HepG2, and 17 out of 25 for K562 independently. Taken together, these data suggest that RBP binding is enriched in UTRs over pG4 sequences, and that RBP-pG4 interactions may regulate gene expression.

An analysis of gene expression changes with sh-RNA knockdown for the majority of pG4-enriched binding proteins showed genes containing pG4 in either the 5' or 3' UTR are much more likely to be significantly differentially expressed compared to non-pG4 genes (Figure 2.16).

Approximately one-third of the proteins exhibiting a binding preference for UTR pG4 change the expression of pG4-containing genes concordantly across K562 and HepG2 cells (GTF2F1, FASTKD2, UPF1, NONO, GRSF1, NCBP2, AKAP8L, DDX6, FKBP4, TAF15, LARP4). Of these 11 RNA-binding proteins, knockdown of eight tends to decrease expression of UTR pG4 genes (GTF2F1, UPF1, NONO, GRSF1, NCBP2, AKAP8L, DDX6, LARP4), while knockdown of three (FASTKD2, FKBP4, TAF15) tends to increase their expression, suggesting that most of the proteins enriched for pG4 binding tend to increase, or stabilize RNA expression rather than facilitate their degradation. This result is consistent with our finding that cis-eQTLs affecting 3'UTR pG4 sequences are more frequently associated with decreasing gene expression.

Finally, to explore the potential existence of post-transcriptional regulatory networks relying on shared RNA G4-protein interactions, we tested for a significant overlap in pG4 containing transcripts targeted by each protein enriched for pG4 binding interactions. Taking the set of 31 proteins with significant overrepresentation for pG4 binding (Bonferroni-corrected $P < 0.001$) and at least 20 unique pG4 binding sites in HepG2 or K562, we assessed overlaps between the various proteins' pG4 gene targets (Fig. 2.10). We found low overlap of targets in helicases that have been hypothesized to bind RNA G4s frequently, such as DDX6, DDX51, and DDX52. In contrast, we find a subset of G4-binding proteins sharing a significant degree of overlap in G4-gene targets, including FASTKD2, FAM120A, CSTF2T, PRPF4 and GTF2F1, none of which have been shown to bind RNA G4 structures previously. These data point to possible mechanisms of

gene control relying on the shared interactions of these proteins with their respective RNA targets. Indeed, assessing the functional associations of 133 pG4 genes sharing at least 3 out of 5 protein-binding interactions from this module revealed enrichment for genes involved in viral process (GO:0016032, FDR-adjusted $P=0.0268$), suggesting that these genes and putative pG4 binding proteins, may be involved in mediating host-viral interactions within the cell.

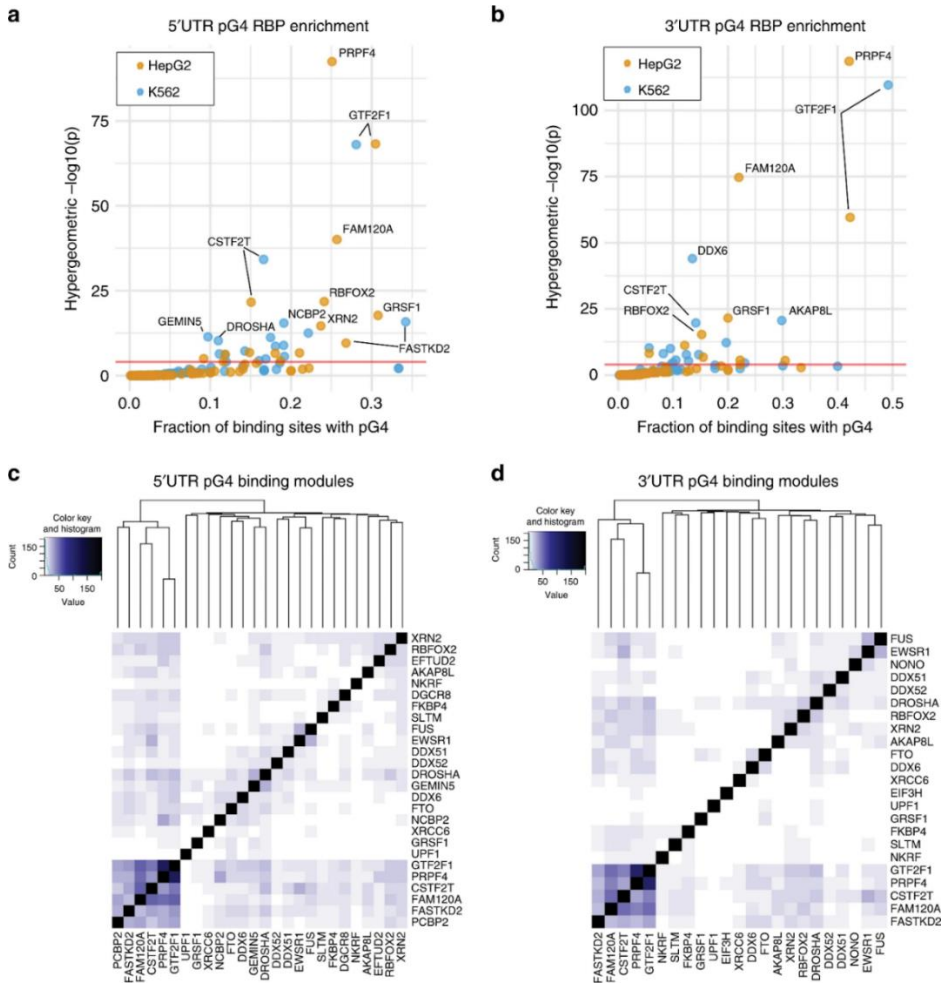


Figure 2.10: Enrichment of specific protein-pG4 binding sites using CLIP-seq data from ENCODE. a, b Enrichment of specific proteins over pG4-binding sites within the 5' UTR (left) and 3' UTR (right)—red line corresponds to $P=0.0001$ (hypergeometric test). c, d Heatmaps depicting the significance of overlap (hypergeometric $-\log P$ value) in pG4 gene targets for proteins found to bind pG4 sequences preferentially.

2.7: 3'UTR pG4 in disease-causing genes are enriched for variants

Multiple studies assessing evolutionary constraints in protein-coding regions of the human genome have shown that regions depleted of genetic variation are also enriched for pathogenic variation [85–87]. Under the principle that purifying selection removes deleterious variants from the genome to produce regions depleted of genetic variation, we expect UTR pG4s should also be enriched for pathogenic variation. Since pathogenic variants in ClinVar are overwhelmingly annotated in protein-coding regions of the genome, we are underpowered to test for a direct association between the set of annotated pathogenic variants and UTR pG4 sequences. Instead, we asked whether potentially pathogenic variation in ClinVar is enriched within UTR pG4 sequences in known disease-associated genes. To test this hypothesis, we mapped all single nucleotide variants annotated in the most recent release of the ClinVar database available at the time of this writing [88] (April, 2019) across UTRs, and compared their relative density in pG4 versus non-pG4 sequences in disease-associated genes. We defined the set of disease-associated genes as any gene with at least one variant having an annotated as Pathogenic or Likely Pathogenic in ClinVar, excluding variants with an annotation of Benign or Likely_benign. To maximize our power for this analysis, we expanded our set of rG4-seq G4s to include all non-canonical G4-forming sequences mapped and reported by rG4-seq in HeLa cells [63]. We found modest enrichment for variation in 3'UTR pG4 sequences, rG4 3'UTR sequences, and a notable enrichment in 3'UTR pG4 sequences within annotated RBP binding sites from ENCODE in disease-associated genes compared to non-pG4 forming regions of the 3'UTR - Fig. 2.11a (All pG4: OR 1.51, 95% CI 1.20-1.88, $P < 0.0005$, rG4-seq pG4: OR 1.18, 95% CI 0.98-1.42, $P = 0.067$, RBP pG4: OR 6.01, 95% CI 3.87-8.91, $P < 5e^{-12}$ - Fisher's Exact Test). In contrast, there was only evidence for enrichment of variants in the 5'UTR rG4 sites (OR 2.32, 95% CI 1.93-2.78, $P < 2.25 \times 10^{-16}$ Fisher's Exact Test), but not pG4 or pG4-RBP overlap regions. It is important to note though that the above statistical test only contrasts relative enrichment of putative pathogenic variants in pG4 vs. non pG4 UTR sequences. Thus, the lack of such relative

enrichment in the 5' UTR may reflect the generally greater density of other functional elements within 5'UTR sequences. In conclusion, these data imply that disease-associated noncoding variation may be enriched in 3'UTR pG4 regions.

Finally, we tested for enrichment of common variants that have been associated with disease phenotypes using annotations available in the NIH GWAS Catalog (April 2019). There were not enough GWAS-associated lead variants within UTR pG4 regions to detect enrichment (7 variants in 5'UTR pG4, 4 in the 3'UTR pG4). However, given the enrichment for cis-eQTLs in UTR pG4, we hypothesized that disruption of UTR pG4 sequences could affect post-transcriptional mechanisms regulating gene expression, thus providing a potential mechanistic link between GWAS variants and their observed phenotypes. To test this hypothesis, we assessed evidence of allelic imbalance at select GWAS SNPs either falling within a UTR pG4 region, or in high LD with a common SNP (r -squared > 0.85 in the GBR population of 1KG) falling within a UTR pG4 in GTEx. Despite being limited by the number of heterozygous individuals in GTEx with matched whole-genome sequencing available, our analysis uncovered several proxy SNPs in high LD with GWAS tag-SNPs (Figure 2.18), and one GWAS lead variant exhibiting evidence of significant allelic imbalance. The lead GWAS variant, rs1048238 is a common SNP within the 3'UTR of *HSPB7*, a chaperone protein that is highly expressed in heart and skeletal muscle and has been associated with hypertension in a recent GWAS [89,90] (Fig. 2.11b). We found that rs1048238 exhibited a substantial imbalance of reads mapping to the alternative allele in 84 heterozygous individuals, even after correcting for read-mapping biases using WASP-filtering [91]. Taken together, these results demonstrate that the predicted pG4-disrupting variant is associated with increased expression of the alternative allele at this locus (Fig. 2.11c-d). This association is consistent with our previous observations from transcriptome-wide mapping of pG4 eQTL showing that 3'UTR pG4 eQTLs tend to increase gene expression (Fig. 2.9) and suggest that the impact of these variants on gene expression are responsible for their respective GWAS associations.

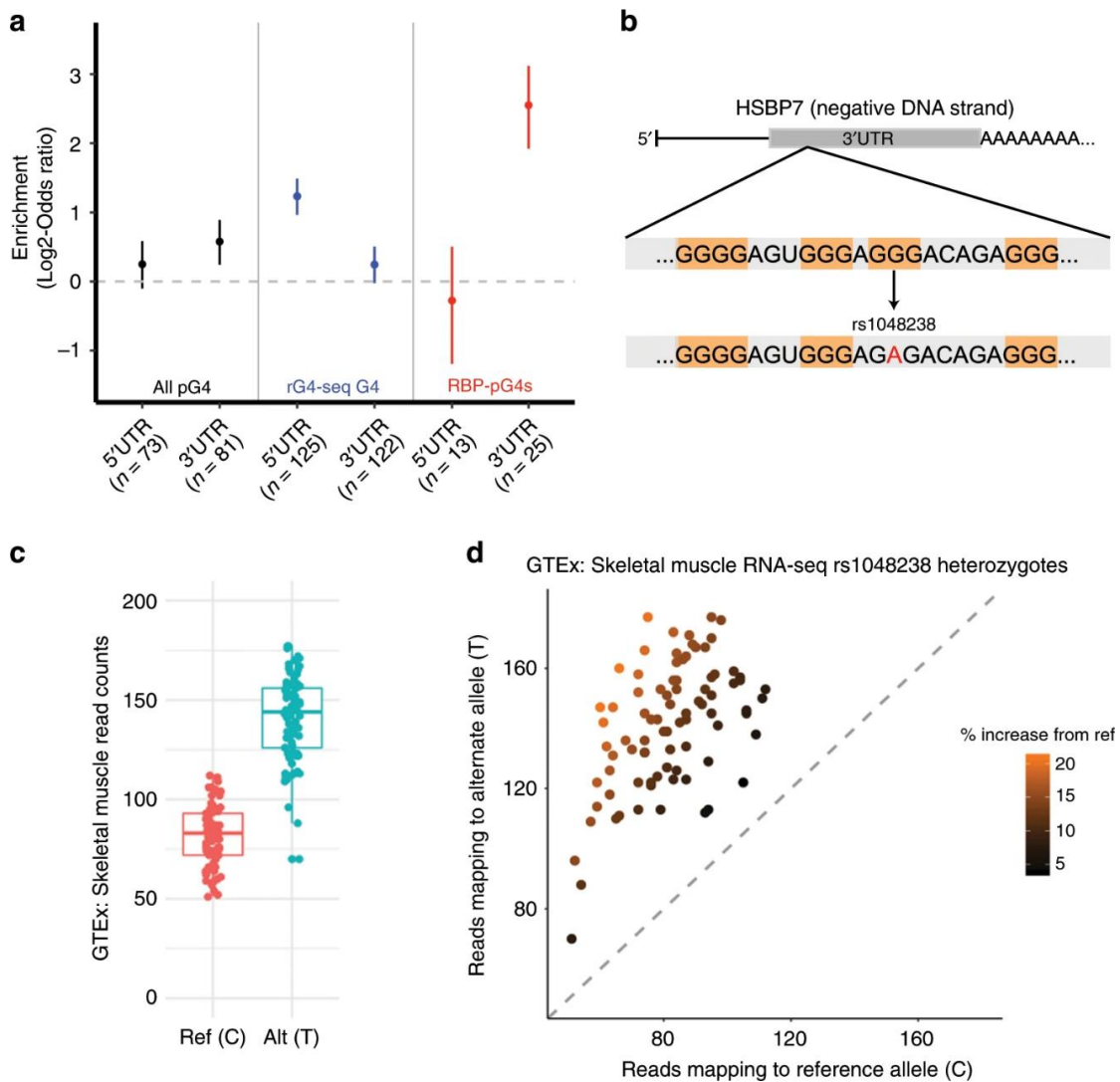


Figure 2.11: UTR pG4 sequences are enriched for known pathogenic, and putative disease-associated genetic variants. (a) Annotated variants within ClinVar disease-associated genes occur with greater frequency in UTR pG4 sequences compared to non-pG4 UTR regions in the 3' UTR across multiple G4 subsets (error bars represent the 95% confidence interval). (b) rs108348 maps to a 3' UTR pG4 G-tract guanine within the primary HSPB7 transcript, which is encoded on the negative DNA strand. The SNP disrupts the canonical G4 sequence motif by causing a G to A mutation in the RNA transcript. (c, d) WASP-mapping of allele-specific reads in 84 GTEx skeletal muscle samples reveals significant allelic imbalance favoring expression of the alternative allele (P value $< 1 \times 10^{-100}$, likelihood ratio test). Boxplot in c represents median and 1.5 times the interquartile range of WASP-aligned RNA-seq reads aligning to the reference (red) or alternative (blue) allele.

2.8: Summary and future directions

We have applied a deep catalog of human genetic variation to assess evolutionary pressures over putative G-quadruplex forming sequences within 5' and 3' UTRs. We hypothesized that if these regions are functionally important they should be depleted of variation. Supporting this hypothesis, we show that variation within UTR pG4 sequences is reduced compared to non-pG4 UTR regions using a local sequence context based substitution model. Moreover, our analysis of positional constraint within the meta-pG4 motif reveals selective pressures acting on central guanines of each trinucleotide G-tract comparable to that of missense mutations in protein-coding regions of the genome. These findings are consistent with in vitro biophysical studies of DNA G-quadruplex stability, which have shown that central position substitutions are most destabilizing, and consequently were predicted to be the most deleterious for native biological functions of G4s [76,92,93]. Interestingly, we find that non-central guanines appear less constrained compared to central positions - possibly because mutations at these positions may preserve the potential for RNA to form non-canonical G4 2-quartets. Indeed, these G4 2-quartets have been estimated to account for 1/4 to 2/3 of all RNA G4 structures observed by transcriptome-wide rG4-seq in HeLa cells [63].

We also uncover a greater proportion of cis-eQTLs mapping to pG4 regions compared to non-pG4 sequences within both 5' and 3' UTRs. Our analysis of nominally-significant cis-eQTL enrichment in UTR pG4 sequences may be confounded by the presence of linked SNPs that reach nominal significance because of their proximity to causal eQTL SNPs, however this likely deflates our estimates of enrichment in UTR pG4 sequences because the relatively smaller size of pG4 motifs (15 - 33nt) makes multiple linked nominally significant cis-eQTLs more likely to occur along the length of non-pG4 UTR regions. Nevertheless, the enrichment of cis-eQTLs within UTR pG4 remain unchanged when we limit each UTR pG4 feature to contain at most 1 nominally-significant cis-eQTL SNP.

Using CLIP-seq data for over 150 proteins published by ENCODE, we find 15 proteins whose binding sites are enriched for pG4 sequences across two cell lines, and identify regulatory modules associating a set of RNA binding proteins, including FAM120A, FASTKD2, and CSTF2T, with pG4 gene targets involved in viral mRNA expression. Indeed, several examples of viral hijacking of eukaryotic RBPs have been reported in the literature [94,95], and G4-forming sequences have been found to occur commonly in multiple viral genomes [96]. This, coupled with the observation that RNA G4s appear to be universally depleted within prokaryotic transcriptomes [65], suggests that viruses might rely on G4s as a mechanism for co-opting host cell machinery involved in gene expression and RNA regulation.

There are three primary limitations to the current study. First, we applied a text-based approach towards identifying regions of putative G-quadruplex formation within RNA UTRs. Although this approach has been commonly employed in previous work [63,72], there exists a considerable literature regarding possible variations to the canonical G-quadruplex forming sequence and methods that capture more variable motif definitions [97,98]; [99]. Given the comparably limited evidence that many of these alternative G-quadruplex sequences form readily in cellulo we used a more stringent motif definition, but alternative G4 sequences will have been missed in our analysis. The modest enrichment in singletons at the central position of trinucleotide G- and C-tracts not matching our canonical pG4 sequence motif is consistent with this possibility. Thus, our assessment of sequence constraint and functional enrichment within UTR G4 forming regions is likely incomplete. Secondly, although we have uncovered evidence suggesting G4 secondary structure formation is constrained, whether these pG4s form secondary structures in vivo remains unclear. Finally, our assessment of selective pressures acting across UTR pG4 sequences using the MAPS metric is limited in power by low variant numbers. Nevertheless, we report multiple lines of evidence supporting the biological importance of putative secondary structure-forming G-

quadruplexes within UTRs. Although RNA UTRs represent only a small fraction of the noncoding genome, they are core components involved in mediating post-transcriptional regulation of gene expression. Ultimately, we hope this work will motivate researchers to consider G4s and other RNA elements in UTRs when assessing the possible impact of genetic variations in human health and disease.

2.9: Supplementary Materials to Integrative analysis reveals RNA G-Quadruplexes are selectively constrained and enriched for functional associations

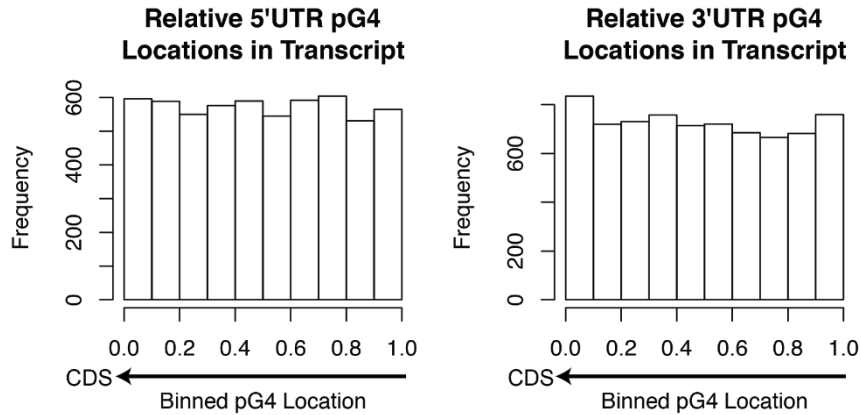


Figure 2.12: Distribution of binned distances for mapped canonical 5' (left) and 3' (right) UTR pG4 sequences with respect to protein-coding sequences across pG4-UTR containing mRNA transcripts. The relative locations of pG4 sequences within UTRs are plotted (x-axis), with 0 being adjacent to the coding sequence, and 1 representing the full-length of the annotated UTR away from the CDS.

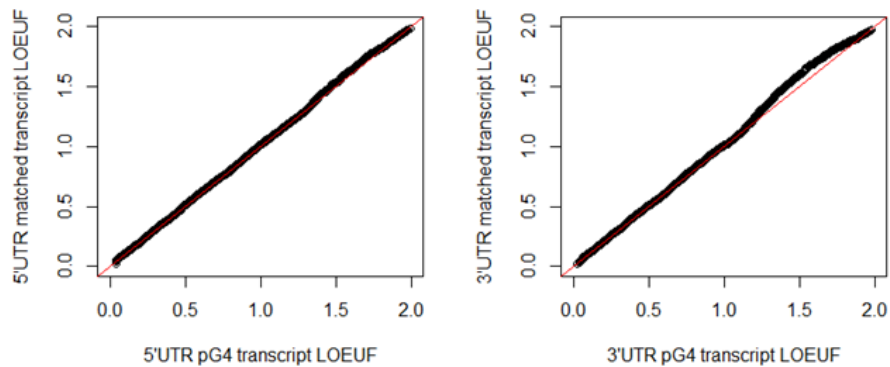


Figure 2.13: Quantile-quantile plot showing matching between pG4 and non-pG4 containing transcripts based on LOEUF scores for 5'UTR (left) and 3'UTR (right) transcripts. Allele frequencies (Figure 1b) and substitutions (Figure 1c) were compared across constraint-matched transcripts using gnomAD's LOEUF metric to control for the possibility that nearby constrained coding sequences might affect local allele frequency estimates. LOEUF scores for non-G4 transcripts plotted on the X-axis and LOUEF scores for G4-containing transcripts are plotted on the Y-axis.

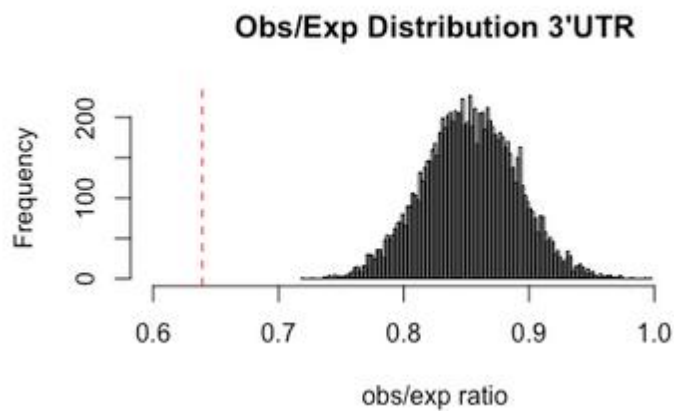
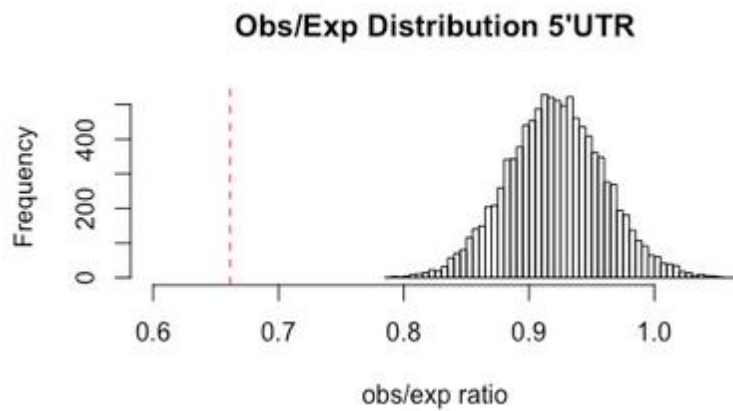


Figure 2.14: The empirical distribution of observed vs. expected number of substitutions across 10,000 bootstrapped 5' and 3' UTR regions in the European subpopulation of the 1000 Genomes Project Phase 1 release. Red dotted line indicates the observed vs. expected ratios estimated by applying the noncoding heptamer mutation model across 5'UTR and 3'UTR pG4 sequences respectively.

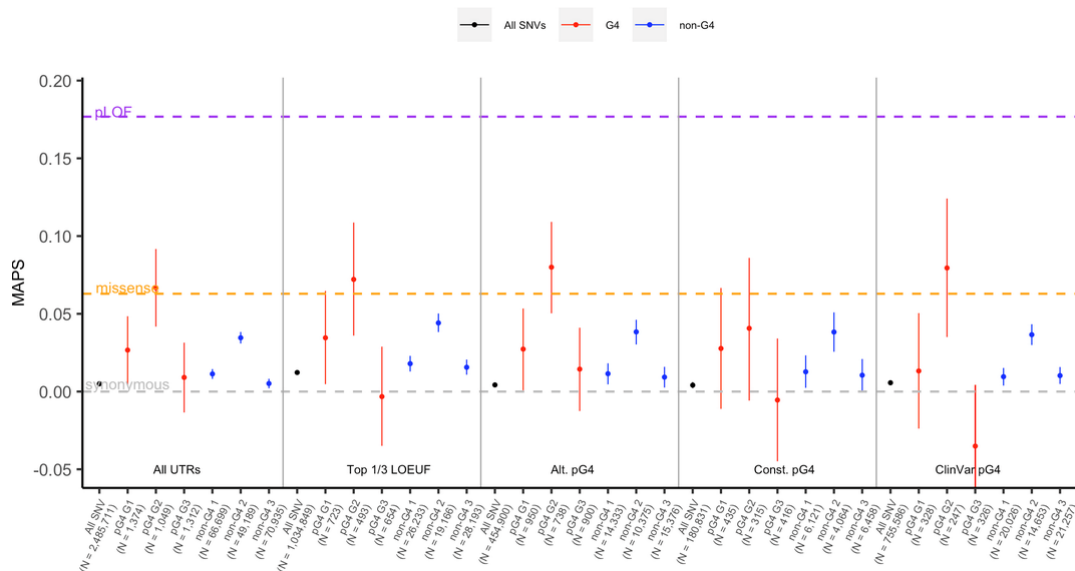


Figure 2.15: MAPS scores for All (black), G4 (red), and non-G4 GGG/CCC (blue) variants across multiple gene sets. Error bars represent 90% CI from 10,000 bootstraps. Top $\frac{1}{3}$ LOEUF represent MAPS scores for pG4 versus non-pG4 variants within the top-1/3rd most constrained genes as estimated by the gnomAD LOEUF metric. Alt. pG4 represent alternatively included pG4 sequences while Const. pG4 represent constitutively included pG4 sequences. ClinVar pG4 sequences are those pG4 within UTRs of disease-associated genes in ClinVar. Permutation P-values for the second guanine of each trinucleotide G-tract context compared to non-pG4 G-tracts in UTRs are: $P=0.0195$ for all UTRs, $P=0.1063$ for Top $\frac{1}{3}$ LOEUF, $P=0.0124$ for Alt. pG4, $P=0.4653$ for Const. pG4, and $P=0.0579$ for ClinVar pG4. Genome-wide MAPS scores for synonymous (grey), missense (orange), and putative loss of function (red) protein-coding variation shown as dotted lines.

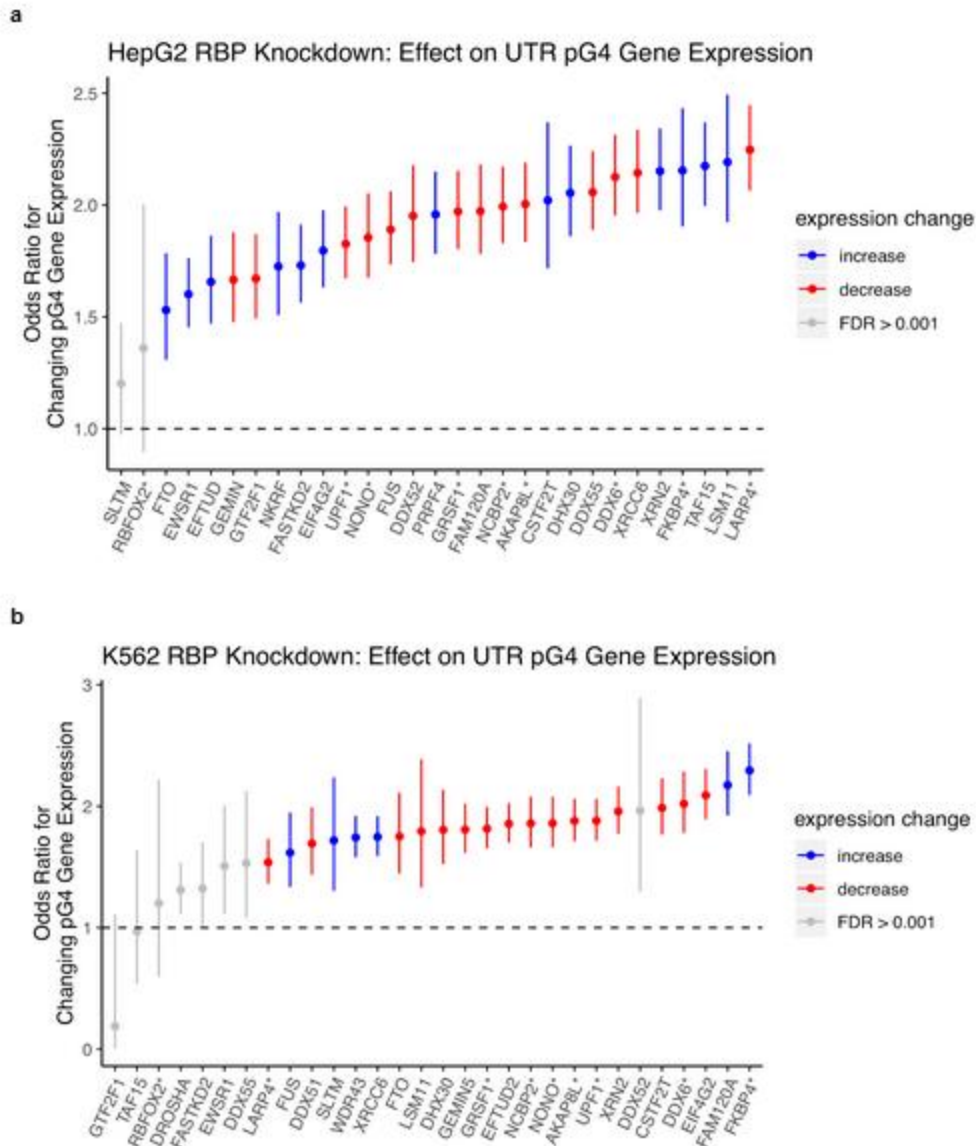


Figure 2.16: Distribution of odds ratio with error bars representing 95% confidence interval for changing gene expression of a pG4 containing gene versus a non-pG4 containing gene with sh-RBP knockdown in ENCODE as determined by Fisher's Exact Test. Results for HepG2 (a) and K562 (b) are shown. Colors represent a tendency for an RBP-knockdown to increase the expression of pG4 containing genes (blue), decrease their expression (red) or have no effect on changing the expression of pG4 genes (grey) at an FDR < 0.001. Proteins having the same direction on changing pG4 gene expression across both HpeG2 and K562 cell lines are marked by an asterisk.

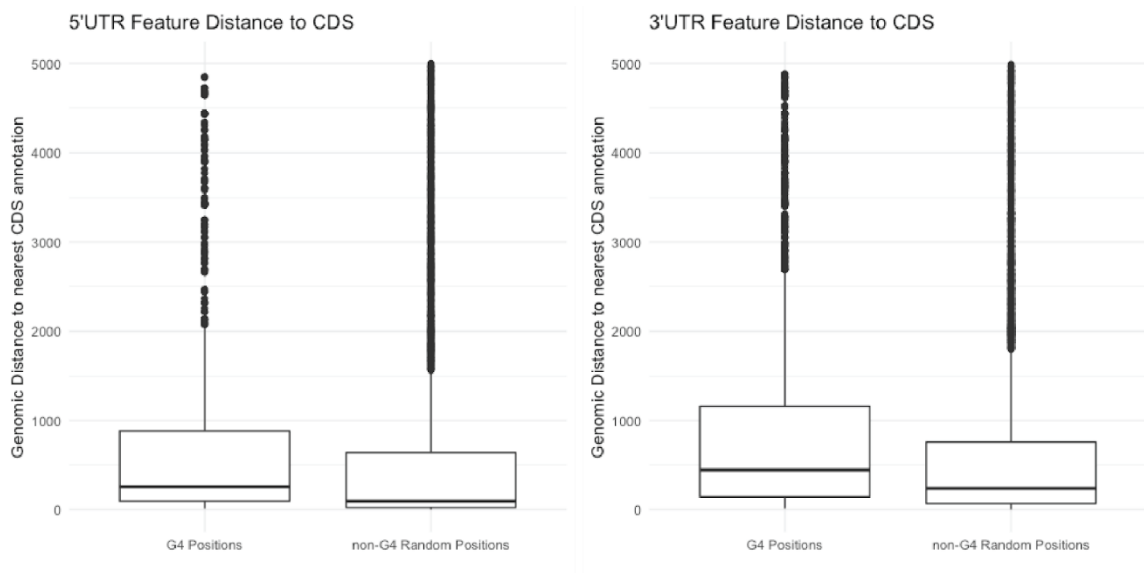


Figure 2.17: Boxplot of distribution of distances between 5' UTR (a) and 3'UTR (b) pG4 sequences and nearest annotated protein-coding exons in ClinVar disease-associated genes showing the 1.5 times the interquartile range and median values. Compared to randomly selected positions within non-pG4 5' and 3' UTRs of ClinVar disease-associated genes, UTR pG4 sequences tend to be located further away from protein-coding exons.

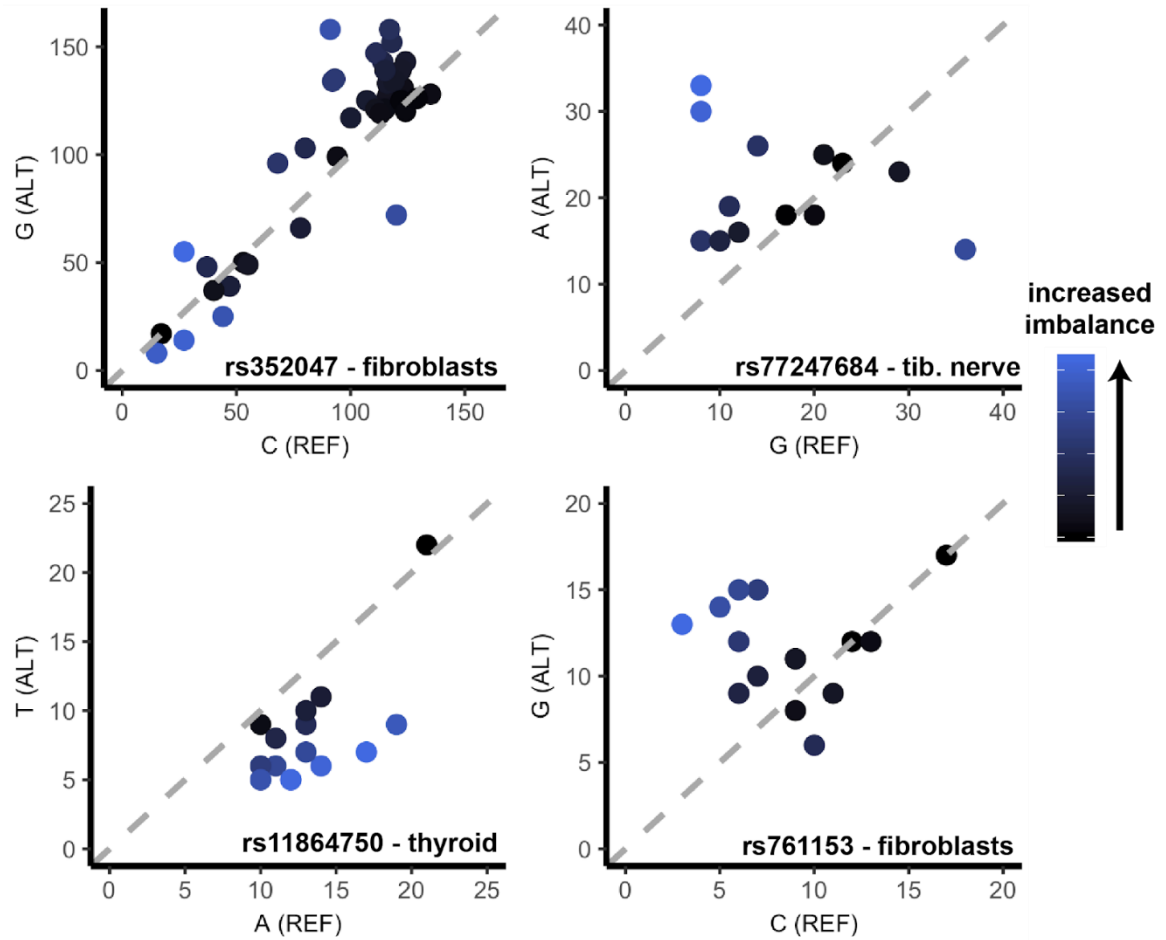


Figure 2.18: Additional common SNPs in high LD ($r^2 > 0.85$ in the 1000 Genomes GBR population) with GWAS tag SNPs exhibiting evidence of allelic imbalance in UTR pG4 sequences. rs352047, rs77247684, and rs761153 affect 3'UTR pG4 sequences. rs11864750 affects a 5'UTR pG4 sequence.

CHAPTER 3: DISRUPTING UPSTREAM TRANSLATION IS ASSOCIATED WITH LOSS-OF-FUNCTION IN HUMAN DISEASE*

3.1: Ribosome profiling and non-canonical open reading frames

With the completion of the human draft genome sequence, elucidating the number of protein-coding genes represented a major challenge. While searching for known homologs using existing gene databases could help map the locations of known protein-coding sequences in the human genome, these strategies depended on the completeness of existing databases and the degree of evolutionary relatedness between humans and other species [100]. Naive approaches to mapping protein coding sequences relied on identifying sequence stretches that could correspond to open reading frames (ORFs) based on the amino acid code. These techniques identified sequence stretches consisting of first identifying an upstream start codon, and scanning along the DNA sequence in search for an in-frame downstream TGA, TAG, or TAA stop codon. Under the assumption that the DNA has a random sequence, and ~50% GC-content, a trinucleotide stop sequence is expected to appear once every 64 base pairs. Putative protein-coding ORFs could therefore be identified if they extended significantly beyond this expected length. Crucially, ORFs smaller than 64 base pairs could not be distinguished from background nucleotide distributions and therefore could not be annotated as functional with certainty. While this ORF-scanning approach could be used to identify almost all the known protein-coding genes in prokaryotes, the increased size and widespread presence of introns of the eukaryotic genome presented significant challenges.

*Published as Lee, D.S.M. et al. BioRxiv 2020. <https://doi.org/10.1101/2020.09.09.287912>.

To address these challenges, several refinements to the basic ORF-scanning approach were proposed. These included using biased codon distributions of known protein-coding regions of the genome to assess new putative ORFs, identifying exon-intron boundaries using known splice-site motifs, and matching putative genes to upstream regulatory sequences including CpG islands that typically mark the beginning of protein-coding genes [100]. Although these strategies significantly improved computational ORF mapping from human DNA sequence alone, low-throughput experimental validation remained indispensable to ascertaining whether predicted ORFs were truly capable of producing endogenous protein products.

Annotation of ORFs using direct experimental evidence of ribosome translation is now possible through high-throughput ribosome profiling. Ribosome profiling is an experimental technique that produces a global, quantitative snapshot of actively translating ribosomes throughout the cell [101]. The typical experimental workflow involves treating cells with compounds that inhibit translational elongation and immobilize translating ribosomes on RNA transcripts. These RNAs are extracted and digested using nucleases that remove RNA fragments unprotected by the presence of immobilized ribosomes. After digestion, ribosomal RNA (rRNA) is depleted, and the remaining ribosome-protected fragments are sequenced and re-aligned to the genome (Fig 3.1). Because translation elongation proceeds with a 3-nucleotide periodicity, this feature can be combined with others to computationally annotate putative ORFs throughout the transcriptome [101,102].

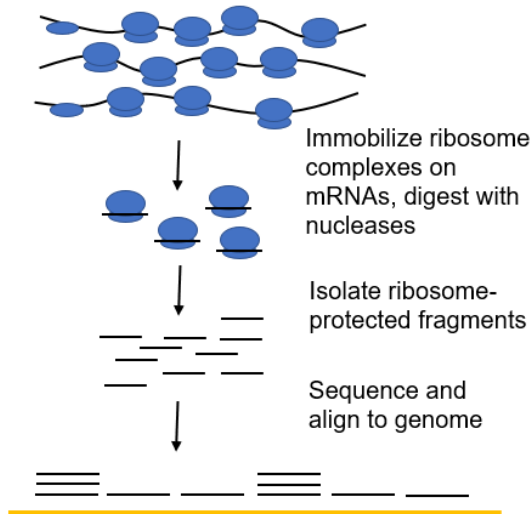


Figure 3.1: Basic experimental workflow for ribosome profiling.

3.2: Pervasive translation in non-canonical open reading frames

The unanticipated abundance of ribosome-protected fragments mapping outside of known annotated protein-coding regions of the genome was a key observation arising from early ribosome-profiling studies [102–104]. When mapped back to the genome, many of these fragments shared similar characteristics to ribosome-protected fragments in known coding regions of the genome, including size distribution and 3-nucleotide periodicity [105]. Together, these studies suggested that cellular RNAs previously not thought to encode for functional proteins or peptides were nevertheless being translated by ribosomes. As many of these newly mapped non-canonical ORFs (ncORFs) were short in length, some suggested that they could encode for functional micropeptides which had been previously overlooked in ORF mapping studies due to biases favoring ORF annotation with longer sequence lengths [104–107].

Early examples of functional micropeptides had been characterized in the literature prior to their being identified in ribosome profiling data. Early evidence of micropeptide functionality was first characterized in yeast, where mutagenesis studies of 247 small ORFs <100 amino acids in length identified 22 ORFs required for haploid growth [108]. Specific functional small ORFs were further characterized from studies in *Drosophila* using polysome profiling - a technique which involves fractionating cellular mRNAs in a sucrose gradient to extract species bound by multiple ribosomes - which identified that select noncoding RNAs without obvious coding potential could be enriched from polysome-bound fractions [109]. Among these early examples of translated noncoding RNAs, the *tarsal-less (tal)* gene transcript was found to encode 4 small ORFs capable of producing 11-amino-acid-long peptides indispensable for early *Drosophila* morphogenesis [109]. Further investigation of polysome-bound noncoding RNAs from *Drosophila* identified a second polycistronic noncoding RNA (pncr003:2L) encoding two ORFs of length 28 and 29 amino acids each producing peptide products involved in regulating cardiac calcium transport [110]. Together these functional micropeptides served as early evidence that length-biases in gene discovery pipelines may have overlooked an entire class of functional elements encoded within sequenced genomes.

To date, several additional examples of functional micropeptides in the human genome have been reported. These include several micropeptides implicated in regulating intracellular calcium levels and muscle contraction [111–113], the inflammatory response [114,115], and cellular metabolism [116–120]. Although modern ribosome-profiling approaches have uncovered widespread evidence translation in thousands of small non-canonical ORFs (ncORFs) in the human genome [102,121,122], what fraction of these ncORFs (ncORFs) can produce functional micropeptides remains an open question.

A second possibility is that translation of ncORFs serve regulatory rather than coding functions. This regulatory hypothesis is most strongly supported by studies of upstream open reading

frames (uORFs). Since uORFs can initiate translation by ribosomes prior to their reaching downstream coding sequences (CDS), they are most frequently associated with repressive effects on downstream CDS translation [122–124]. Moreover, it has been observed that cis-regulatory relationships between uORFs and downstream coding sequences are frequently maintained across species, but the nucleotide content of these ORFs are not [122,125]. Together, this evidence implies that the functional importance of uORF translation is in its regulatory effect on downstream protein expression rather than micropeptide encoding. Nevertheless, a few examples of uORF-encoded micropeptides which are capable of repressing downstream translation initiation at the CDS in a peptide-dependent have also been observed, although this phenomenon appears to be the exception rather than the rule [126,127]. More recently, the striking pervasiveness of translation outside of canonical protein coding ORFs revealed by ribosome profiling has not stopped some from speculating that micropeptides resulting from ncORF translation are broadly functional [107].

3.3: Disrupting upstream translation in mRNAs is associated with loss-of-function in human disease

The classic view of information processing in the cell by gene expression occurs through transcription followed by translation. This basic flow is often complicated by regulatory elements which confer additional stages of processing and control. In particular, upstream open reading frames (uORFs) are segments of 5'UTR mRNA sequences that can initiate and terminate translation upstream of protein-coding start codons. Specific uORFs are known to control protein expression by tuning translation rates of downstream protein-coding sequences, and potential uORFs have been identified in ~50% of all human protein-coding genes [123,128].

Translation initiation is the rate-limiting step controlling post-transcriptional gene expression [129], and rates of translation initiation can significantly impact mRNA stability [130–134]. Cap-

dependent translation initiation begins when the 40s ribosomal subunit encounters a start codon as it scans along the 5'UTR. At the start codon, the 40s subunit acquires the 60s subunit with other translation initiation factors and peptide synthesis begins. Scanning ribosomes encountering uORFs may prematurely initiate translation in the 5'UTR; if this occurs, upon reaching the uORF termination codon the ribosome may dissociate from the mRNA transcript, or the 40s subunit may resume scanning after the 60s subunit is lost. Resumption of scanning leads to translation of downstream reading frames only if the necessary translation initiation factors are reacquired by the 40s subunit before reaching the downstream start codon. Thus, the spatial combination of uORFs and protein-coding start codons can produce different effects on translation of the downstream gene.

Previous analyses of large-scale population data have shown that genetic variants creating new uORFs are rare, suggesting that these variants are subjected to strong negative selection due to their capacity to cause pathogenic loss-of-function of associated proteins [123,135]. Moreover, it has been shown that variants destroying stop codons in translated uORFs are under strong negative selection, presumably because the resultant translational readthrough can decrease start codon recognition and translation initiation at the coding sequence (CDS) [46]. In contrast, less is known about the impact of genetic variation within translated uORFs. Furthermore, recent untargeted ribosome-profiling experiments have revealed striking evidence of active translation at thousands of uORFs throughout the genome, but the biological significance of this phenomenon remains unresolved [123].

Here we use translated uORFs mapped through ribosome-profiling experiments and a deep catalogue of human genetic variation to characterize patterns of selection acting on single nucleotide variants (SNVs) in translated uORF sequences. We assess evidence for the functional importance of translation at uORFs, and explore possible phenotypic consequences associated with genetic variation in these sequences. Using the allele frequency spectrum of SNVs from

71,702 whole genome sequences in gnomAD, we find that SNVs introducing new stop codons, or creating stronger translation termination signals in uORFs are under strong selective constraints within 5'UTRs. We propose that these variants are under selective pressure because they disrupt translation initiation at downstream protein-coding sequences. We then utilize the Penn Medicine Biobank (PMBB) to discover new, robust disease-gene associations using uORF stop-creating and stop-strengthening variants and replicate these associations in the UK Biobank (UKB), and by gene burden tests aggregating rare protein-coding loss-of-function variants. Finally we validate the impact of uORF stop-creating and stop-strengthening variants on protein expression for our top phenome-wide significant associations. These data demonstrate that mutations in translated uORFs creating new stop codons, or strengthening existing stop codons can contribute to disease pathology by changing protein expression.

3.3: Variants introducing new stop codons in uORFs are under strong negative selection

Since elongating ribosomes must translate uORFs before they reinitiate translation at the CDS, we hypothesized that genetic variants introducing new stop codons in translated uORFs could impede downstream translation initiation. Because these variants interrupt translation without affecting the coding sequence directly, we term them upstream termination codons (UTCs) to distinguish them from premature termination codons within protein-coding sequences.

To estimate the deleteriousness of UTC mutations, we assessed their frequency spectrum in gnomAD using the Mutability-Adjusted Proportion of Singletons (MAPS) metric. MAPS compares the strength of selection acting against different classes of functional variation by assessing the relative enrichment for rare singleton (one sequenced allele) variants in gnomAD, adjusted for local mutation rates (see Appendix B). More deleterious groups of SNVs - including premature termination codons and essential splice site mutations - show greater enrichment in singletons in gnomAD, and consequently have higher MAPS scores. MAPS has previously been used to

assess patterns of selective pressures acting on different classes of variation in both protein-coding and non-coding regions of the genome [37,45,135–138].

Using translated uORFs from 4392 genes identified by deep ribosome profiling of two human cell lines (Figure 3.11) [94], we mapped genetic variation from 71,702 whole-genome sequences in gnomAD (version 3)[91]. We identified the subset of UTC mutations by selecting SNVs which mutated uORF codons to either UGA, UAG, or UAA in the mapped uORF reading frame (Figure 3.2a). We calculated MAPS scores for these UTC mutations, finding that they are under strong negative selection within 5'UTRs, comparable to that of missense mutations in canonical protein-coding regions of the genome (Figure 3.2b). Indeed, MAPS scores for these variants are significantly higher than all uORF variants (Figure 3.2b, $P < 0.001$), sets of uORF variants matched by their underlying trinucleotide mutation context (Figure 3.12, $P < 0.001$ - see Appendix B), all 5'UTR variants creating UTCs outside of mapped translated uORFs ($P = 0.0441$), and stop-creating variants in ORFs in 3'UTRs, translated pseudogenes, and lncRNAs also mapped by ribosome-profiling from the same study (Figure 3.2b $P = 0.0041$, Figure 3.11). Intriguingly, MAPS scores were highest for variants predicted to introduce strong (UAA) stop codons that are less susceptible to translational read-through [95–97]. In contrast, variants introducing the weaker UGA stop codon exhibited MAPS scores that are only nominally higher than MAPS scores for all uORF variants ($P = 0.2833$), suggesting that they may be less deleterious by comparison. To account for the possibility that the heightened MAPS scores for UTC mutations resulted from overlap between 5'UTRs and annotated coding sequences in different mRNA isoforms, we repeated this analysis excluding all uORF variants overlapping with any annotated CDS sequence. Re-calculated MAPS scores with all CDS-overlapping variants removed remained essentially unchanged (Table 3.2), ruling out the possibility that the enrichment in rare variation for UTC mutations is driven by selection on coding sequences. Additionally, we previously observed that variants destroying the central guanine of putative G-quadruplex forming sequences exhibit heightened MAPS scores in UTRs. We repeated this analysis with all potential

G-quadruplex disrupting variants ($n = 57$) excluded, seeing a negligible effect on MAPS scores for all UTC mutations (MAPS = 0.0377, 95% CI: 0.0196-0.0557). Overall, the strong selective pressure to remove UTC mutations implies that these variants are also more likely to have functional biological consequences.

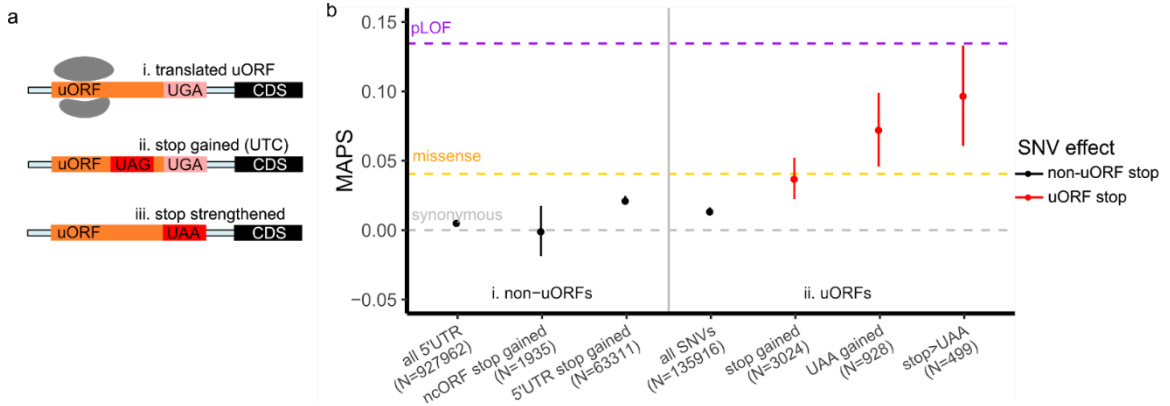


Figure 3.2: Stop-introducing and stop strengthening mutations in translated uORFs are under strong negative selection. (a) Examples of possible stop-gained (UTC) or stop-strengthened mutations in translated uORFs. (b) Mutability-Adjusted Proportion of Singletons (MAPS) scores for different classes of stop-introducing mutations within translated uORFs. Grey, orange, and purple dashed lines represent MAPS scores for synonymous, missense, and predicted loss-of-function (pLOF) SNVs affecting canonical protein-coding sequences in gnomAD. (i) MAPS scores for non-uORF variants including all 5'UTR SNVs, stop gained mutations in ncORFs, and all 5'UTR stop gained mutations (ii) MAPS scores for all uORF SNVs and stop gained mutations in uORFs show that uORF UTC mutations are significantly enriched for singletons. This is also observed for UAA-creating, and stop-strengthening SNVs in translated uORFs. Error bars represent bootstrapped 90% confidence intervals.

3.4: Translated uORFs use weak stop codons

Stop codons have different translation termination efficiencies in both prokaryotes and eukaryotes, with the hierarchy following the general pattern of UAA > UAG > UGA [139,142,143]. Given the observed selection against UTC mutations in translated uORFs, and in particular against UAA-introducing variants, we next asked whether stop codon usage by translated uORFs is distinct from the background distribution of UGA, UAG, and UAA trinucleotides in 5'UTRs. To perform this comparison, we determined the relative frequency that UGA, UAG, or UAA

trinucleotide sequences appeared within non-translated 5'UTR sequences, and compared this frequency to the distribution of stop codons used in translated uORFs. To further control for the possibility that translated-uORF containing UTRs might have significantly different background nucleotide distributions, we also assessed the relative frequency of UGA, UAG, or UAA trinucleotides from uORF-containing UTRs with translated uORF sequences excluded. Strikingly, we find that translated uORF stop codons are significantly depleted of UAAs compared to background UTR distributions (Figure 3.3a), suggesting that weaker stop-codons (UGA, UAG) are preferred (permutation $P < 0.001$ compared to all UTRs, $P < 0.001$ compared to uORF-containing UTRs). Indeed there are approximately 45% less uORF UAA stop codons compared to the relative frequency of UAA trinucleotides in adjacent untranslated UTR sequences (uORF-UAA=19%, matched UTR-UAA=35% - Table 3.3). In contrast, UGA stop codons are enriched within translated uORFs compared to non-translated UTR sequences (permutation $P < 0.001$ compared to all UTRs, $P < 0.001$ compared to uORF-containing UTRs).

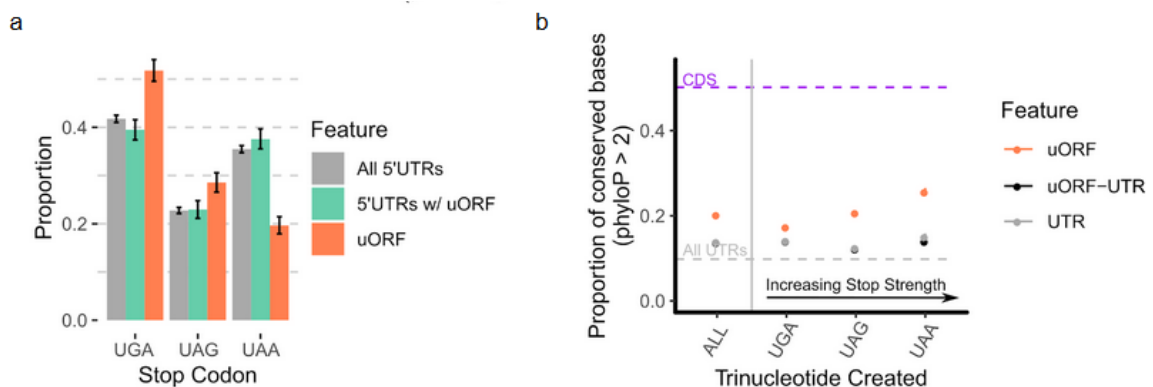


Figure 3.3: Translated uORFs tend to use weak stop codons. (a) Relative frequencies of trinucleotides used as uORF stop codons compared to untranslated regions of uORF-containing 5'UTRs, or all 5'UTRs shows uORFs are significantly enriched for weaker (UGA, UAG) stop codons and depleted of the UAA stop codons compared to control sequences. Error bars represent 95% bootstrapped confidence intervals. (b) Proportion of strongly conserved (phyloP > 2) bases by phyloP scores from 100-way vertebrate alignments for uORF stop-creating, non-uORF stop-creating in uORF-containing UTRs, and non-uORF stop-creating in all UTR genomic positions. Error bars represent 90% bootstrapped confidence intervals.

Given the depletion of UAA-stop codons in translated uORFs, we next asked whether variants changing weaker stop codons (UGA, UAG) to UAA were also enriched for singletons. Compared to synonymous and missense variation within the protein-coding genome, we find that the MAPS metric for stop-strengthening variants is significantly higher (Figure 3.2b-ii). This difference remained significant compared to uORF variants matched by trinucleotide context, indicating that this effect is specific to uORF stop codons ($P=0.012$, Figure 3.12). Given that UAA codons can facilitate greater termination efficiency and more rapid ribosomal dissociation from mRNAs compared to UAG and UGA codons [139,144,145], these results are consistent with the possibility that stronger stop codons in uORFs can also increase the efficiency of translation termination in the 5'UTR. Thus, like UTC mutations, stronger stop codons in uORFs may be disfavored because they decrease the probability that ribosomes reinitiate translation at downstream coding sequences.

3.5 Genomic positions that can create new stop codons in uORFs are conserved

Since the power of MAPS estimates are limited by the number of variants observed in gnomAD, we assessed the evolutionary conservation of each possible uORF stop-creating position as complementary evidence for their functional significance. For this, we compared the distribution of phyloP scores across potential uORF-stop-creating positions derived from the UCSC 100-way phyloP vertebrate alignment [38]. Specifically, for each potential new stop site, we compared the proportion of genomic positions with a phyloP score of > 2 - corresponding to strong conservation across multi-vertebrate alignment - versus those positions that were not strongly conserved ($\text{phyloP} < 2$). A similar approach has been used to show that genomic positions with the potential to produce new uORFs are strongly conserved across vertebrates [135].

We performed several assessments of phyloP scores across 5'UTR contexts. Consistent with our MAPS analysis, potential stop-creating positions in translated uORFs are also more likely to be

conserved compared to UTR positions matched by distance to the downstream coding sequence. This difference remained significant even when compared to potential stop-creating positions in 5'UTR sequences adjacent to (but not within) translated uORFs (Figure 3.3b). Strikingly, conservation at each stop-creating position within mapped translated uORFs mirrored the strength of stop-codon contexts, with a positive correlation between the strength of the potential stop codon introduced and the proportion of uORF genomic positions that are conserved. This trend was not observed for non-translated 5'UTR contexts (Figure 3.3b). In all cases, the proportion of conserved bases for each class of potential stop-creating variant was significantly higher than those positions in all 5'UTRs, and particularly within untranslated regions of translated-uORF containing UTRs ($P < 0.001$, Figure 3.3b). Moreover, the proportion of highly conserved bases at possible stop-creating positions increased in association with increasing gene constraint, as determined by the gnomAD LOEUF score, and remained significantly higher than non-uORF 5'UTR stop-creating positions (Figure 3.13). Together, these complementary analyses support our initial findings that UTC mutations are under strong negative selection within the human genome, and further strengthens the evidence that UTC mutations may functionally disrupt protein expression.

3.6: Upstream open reading frames are not under strong selection to maintain amino acid identity

Multiple transcriptome-wide ribosome profiling studies have proposed that some uORFs can encode functional micropeptides with important cellular roles [106,107,121]. This has fostered significant interest in the possibility that translated, non-canonical ORFs represent an overlooked class of potentially functional micropeptides with biological activity independent of the downstream protein-coding sequences [107,146]. If many uORFs encoded functional micropeptides, the pattern of constraint against UTC mutations might also reflect selection to

preserve micropeptide function rather than downstream translation initiation. To address this possibility, we asked whether uORFs broadly exhibit similar constraints against missense variation, compared to known protein-coding regions of the genome, that could imply peptide functionality. We compared MAPS scores for predicted missense versus synonymous mutations in translated uORFs to those in canonical protein-coding regions of the genome (Figure 3.4). The MAPS scores for missense mutations in uORFs were significantly lower than that of missense mutations in canonical protein-coding regions of the genome, and not significantly higher than MAPS scores for synonymous variants in translated uORFs ($P=0.7118$, Figure 3.4a-iv). These results indicate that selection to maintain amino acid identity in uORF-encoded micropeptides is weak compared to canonical protein-coding sequences. As an additional control, we computed MAPS scores for predicted missense and synonymous mutations in 693, 1188, and 276 translated non-canonical ORFs (ncORFs) mapped by ribosome profiling in 3'UTRs (dORFs), long-noncoding RNAs, and pseudogenes respectively, as these sequences are not thought to broadly encode for functional peptides. Similar to uORFs, predicted missense variants in these additional ncORFs were not significantly higher than predicted synonymous variants by MAPS score (dORFs $P=0.3532$; lncRNAs $P=0.7777$, pseudogenes $P=0.4523$ Figure 3.4a-i-iii).

Since many translated uORFs are short, we asked whether longer uORFs might exhibit greater selection against missense variants compared to shorter uORFs. To test this possibility, we divided uORFs into long sequences >118 codons comprising the top 25% longest mapped uORFs, and short uORFs <118 codons in length. MAPS scores for missense variants in long versus short uORFs yielded no evidence of significant constraint acting on amino-acid changing variants compared to synonymous SNVs (long uORFs $P=0.178$, short uORFs $P=0.9628$, Figure 3.4a-v).

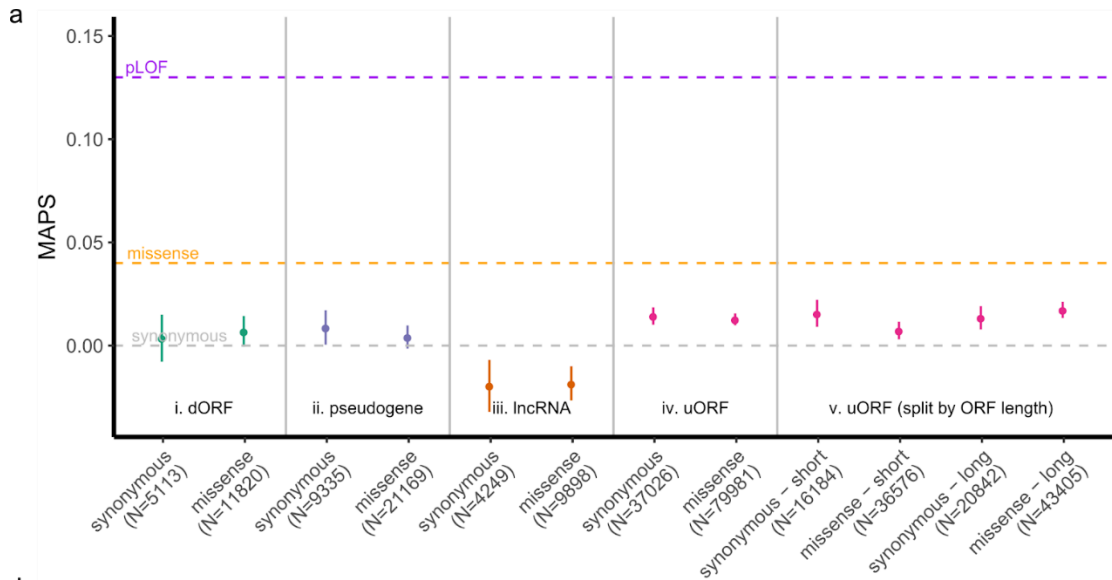


Figure 3.4: ncORFs do not exhibit strong selective pressure to maintain amino acid identity. MAPS scores for single nucleotide variants within each ncORF category separated by predicted consequence (synonymous or missense) in each ORF. (i-iv) Allele frequencies for predicted missense SNVs are not significantly enriched for singletons than those for predicted synonymous SNVs. (v) MAPS scores are no different for long uORFs (> 118 codons) compared to the rest (short). Grey, orange, and purple dashed lines represent MAPS scores for synonymous, missense, and predicted loss-of-function (pLOF) SNVs affecting canonical protein coding sequences in gnomAD. Error bars represent bootstrapped 90% confidence intervals.

Surprisingly, we observed that MAPS scores for both synonymous and missense variants in translated uORFs deviated significantly from all 5'UTR variation (Figure 3.4a-iv). These heightened MAPS scores implied that uORF variants are under increased negative selection compared to all 5'UTR variants. The absence of similar effects for variants in dORFs, lncRNAs, or translated pseudogenes implies that this enrichment in singletons is unique to translated uORFs. One possibility is that synonymous variation in uORFs reflect selective pressures to maintain translational efficiency by preserving codon optimality. Messenger RNAs that are enriched with more optimal codons are both more stable, and more efficiently translated by ribosomes [147]. Like UTC mutations, uORF mutations introducing suboptimal codons could therefore slow translational elongation and impede downstream translation initiation at the CDS. Indeed, mutations introducing suboptimal codons in translated uORFs have been shown to disrupt translation initiation at downstream coding sequences [148–150], and more generally 5'UTRs are

under selective pressures to maintain their capacity for facilitating translation initiation at the CDS [151,152].

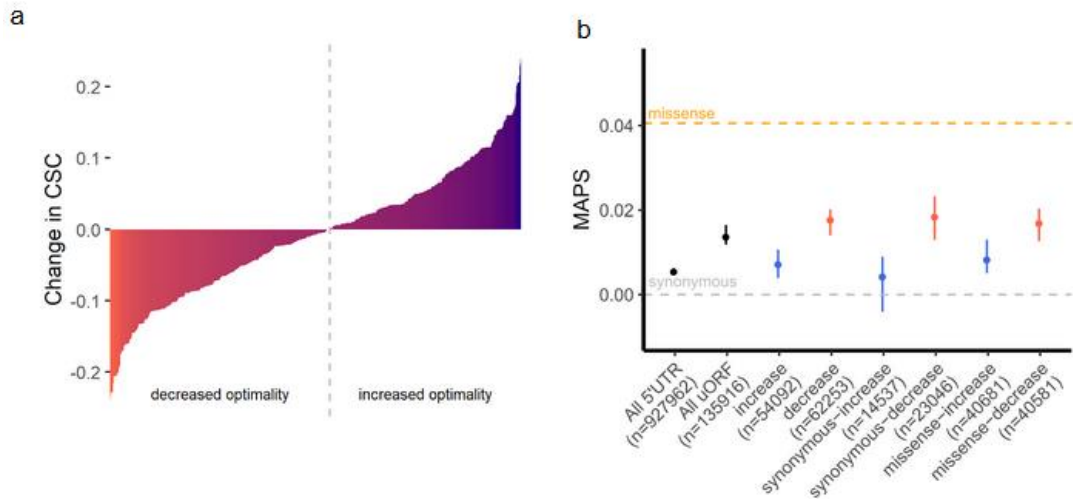


Figure 3.5: Synonymous and missense variants in translated uORFs are under selective pressure to maintain codon optimality. (a) Translated uORF variants ranked by predicted change to codon optimality using codon stability coefficient (CSC) scores from SLAM-seq (red = decreasing, blue = increasing) [110]. Grey dotted line denotes boundary separating optimality increasing versus decreasing SNVs. (b) MAPS scores for SNVs separated by predicted consequence on codon optimality shows heightened constraint against decreasing optimality variants, while variants increasing optimality are indistinguishable from all 5'UTR variants. Error bars represent bootstrapped 90% confidence intervals.

To test whether mutations in translated uORFs are constrained to maintain codon optimality, we asked if MAPS scores for mutations predicted to decrease codon optimality differed from those that increased codon optimality (Figure 3.5a). Using experimentally determined codon-stability coefficients (CSCs) [153], we matched each uORF SNV with its predicted consequence to codon optimality, and compared MAPS scores for optimality-increasing versus optimality-decreasing SNVs. As expected, SNVs increasing codon optimality were indistinguishable from all 5'UTR variants ($P=0.1929$, Figure 3.5b). In contrast, variants predicted to decrease codon optimality had significantly higher MAPS scores ($P<0.001$), although the magnitude of this difference is moderate compared to UTC mutations (Figure 3.3b). This effect remained significant regardless

of whether variants were predicted to cause synonymous or missense mutations ($P=0.0125$ for synonymous; $P=0.009$ for missense), and was notably absent for translated ORFs in 3'UTRs, lncRNAs, and pseudogenes (Figure 3.5b, Figure 3.14). Furthermore, this pattern of increased constraint against optimality-decreasing mutations was robust to the use of CSC scores derived from alternative experimental approaches across several cell lines (Figure 3.15) [153]. Together, these observations further support the hypothesis that natural selection acts to maintain the capacity for translational initiation at downstream coding sequences by preserving translational elongation efficiency in uORFs.

3.7: uORF start codons are conserved and under strong selective pressure

The finding of heightened selection against translation-interrupting variants in uORFs raises the question of why translated uORFs continue to persist in a large fraction of human genes. Evidence that uORF-CDS organization, and the strength of uORF repression is strongly conserved across vertebrates, suggests that translation at uORFs is maintained to regulate downstream translation initiation [122]. Moreover, variants destroying uORF start codons have been implicated in the development of cancer [154]. To provide further genetic evidence that translation at uORFs is maintained by selection, we asked whether allele frequencies for variants affecting uORF start codons also exhibited strong selection to maintain their capacity for translation initiation. Using the MAPS metric, and genome-wide phyloP scores, we evaluated patterns of variation affecting uORF start codons. Since many translated uORFs begin with non-canonical start codons (**Figure 3.6a-i**), we distinguish between variants maintaining the start context by affecting the first position of the NUG trinucleotide from those that disrupt translation initiation by mutating the last two nucleotides in the uORF start codon (**Figure 3.6a-ii**). As expected, start-maintaining variants are no more enriched for singletons in gnomAD compared to synonymous protein coding variants. In contrast, start-disrupting mutations are enriched for singletons at a level comparable to that of protein-coding missense and UTC mutations (**Figure 3.7**). The heightened pressure to maintain translational initiation at uORF start codons is similarly

reflected in phyloP scores for uORF start-disrupting genomic positions compared to distance-matched UTR controls ($P < 0.001$), and uORF-matched controls ($P < 0.001$, **Figure 3.8**). These data show that translation initiation at uORFs is evolutionarily constrained in humans, and are consistent with previous reports that uORF start codons are frequently conserved across species.

Taken together, our analyses of genetic variation in gnomAD show enrichment for rare allele frequencies in the frequency spectra of uORF start-disrupting, stop creating, and stop-strengthening mutations. Results from our analyses indicate that these classes of variation are under a heightened degree of negative selection, and imply that processes of translation initiation, elongation, and termination at translated uORFs are maintained by selective pressure.

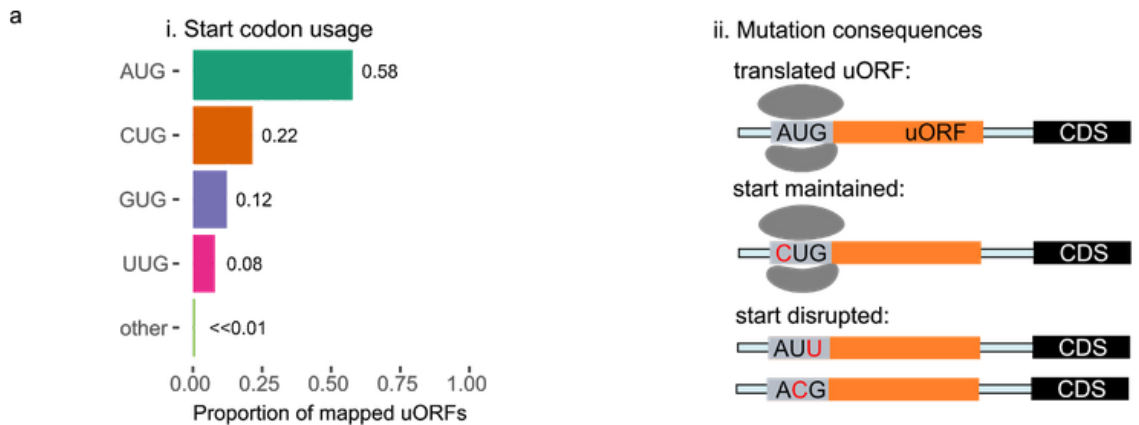


Figure 3.6: Start codon usage for uORFs mapped by ribosome profiling. (i) Distribution of start codon usage for experimentally mapped translated uORFs, and (ii) possible consequences of mutations affecting uORF start codons.

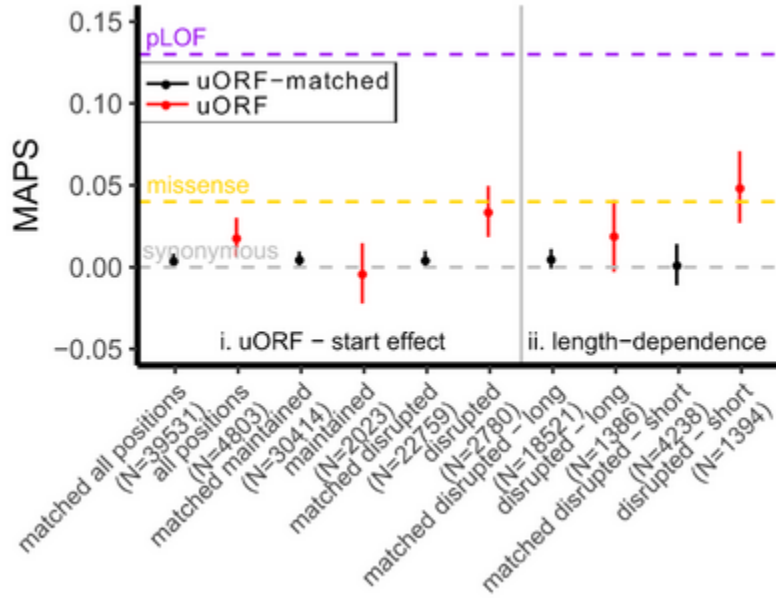


Figure 3.7: MAPS scores for uORF start disrupting variants. (i) MAPS scores for start-disrupting SNVs are compared to uORF variants matched by trinucleotide mutation context. (ii) Start-disrupting SNVs for short (< 20 codons) uORFs are under stronger negative selection compared to start-disrupting variants for long (≥ 20 codons) uORFs. Error bars represent bootstrapped 90% confidence intervals.

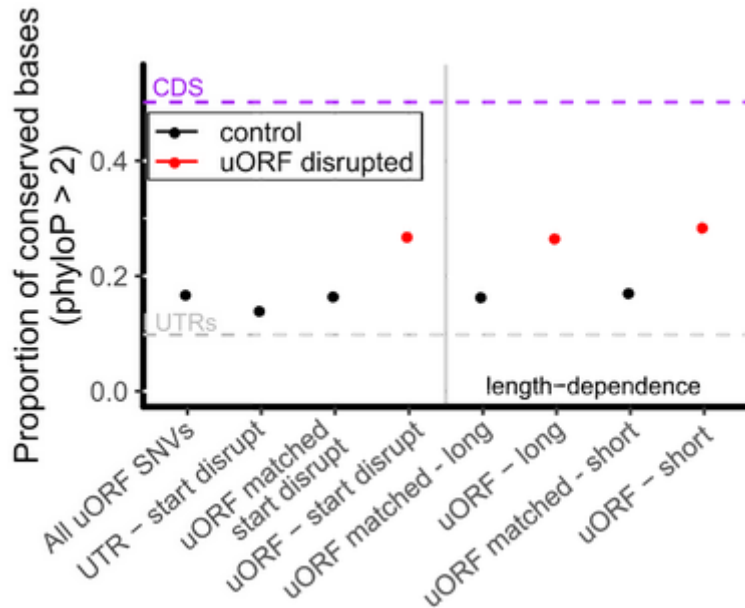


Figure 3.8: PhyloP estimates for possible start codon disrupting positions. uORFs start disrupting positions are compared to all uORF SNVs, UTR-matched start-disrupting positions, and uORF-matched start-disrupting positions in translated uORFs. Start-disrupting genomic positions of short uORFs are more strongly conserved by phyloP scores compared to matched start-disrupting positions within uORFs. Error bars represent bootstrapped 90% confidence intervals.

3.8: uORF-disrupting variants associate genes with new disease phenotypes

The heightened MAPS score for UTC mutations suggests that they are also likely to be functional. To explore the possibility that UTC and uORF stop-strengthening mutations might contribute functionally to human disease susceptibility, we performed a phenome-wide association study (PheWAS) of predicted uORF-disrupting variants using the Penn Medicine Biobank (PMBB) - a large academic biobank with exome sequencing linked to EHR data for 10,900 individuals [155].

Table 3.1: Significant novel associations in PheWAS of Penn Medicine BioBank

Variant	Novel associations					Replication		
Gene (SNP, uORF effect)	Phenotype (Phecode)	OR (95% CI)	P value	Cases	Controls	UKB	PMBB LOF	UKBB LOF
<i>PMVK*</i> (rs181302437) UAG>UAA	250.13 (T1D - ophthalmic manifestations)	27.29 (6.88-108.29)	2.58E-06	23	5189	No	No	Yes (250.13, P = 7.27e-03)
	250.14 (T1D - neurological manifestations)	22.71 (5.86-87.97)	6.20E-06	25	5189	No	No	No
	250.22 (T2D - renal manifestations)	7.79 (2.87-21.17)	5.73E-05	136	5189	No	No	No
<i>VPS53</i> (rs35915949) UGA>UAA	300.10 (Anxiety disorder)	0.64 (0.53-0.77)	4.23E-06	1060	6939	No	No	No
	300.00 (Anxiety disorders)	0.69 (0.58-0.82)	2.00E-05	1249	6939	No	No	No
<i>NALCN</i> (rs139848407) CAA>UAA	270.33 (Amyloidosis)	38.92 (7.49-202.36)	1.34E-05	30	7727	No	No	Yes (270.00, P=0.0264)
<i>BCL2L13†</i> (rs140799351) UGA>UAA	610.00 (Benign mammary dysplasias)	270.57 (19.69-3718.08)	2.80E-05	55	7689	No	Insufficient variants	Insufficient variants
	187.20 (Malignant neoplasm of the testes)	331.41 (21.68-5065.35)	3.03E-05	26	7700	Yes (187.20, P = 2.09e-4)	Insufficient variants	Insufficient variants
	187.00 (Cancer of other male genital organs)	220.01 (15.67-3089.83)	6.31E-05	34	7700	Yes (187.00, P = 3.33e-4)	Insufficient variants	Insufficient variants
<i>SHMT2</i> (rs28365863) UAG>UAA	527.00 (Diseases of the salivary glands)	6.37 (2.60-15.65)	5.27E-05	90	9774	Insufficient cases	Yes (527.00, P = 5.515e-03)	Insufficient cases
<i>MOAP1</i> (rs116450723) UAC>UAA	350.00 (Abnormal movement)	4.99 (2.20-11.33)	1.22E-04	362	9414	No (variant not present in UKB)	No	No (variant not present in UKB)

*As of the Gencode 32 release the 5'UTR *PMVK* annotation (September 2019) was shortened to exclude this uORF; however inspection of the raw ribosome profiling reads from Ji et al. [121] in conjunction with nearby transcription start sites annotated in FANTOM5 confirm the presence of a longer *PMVK* 5'UTR isoform (Figure 3.17). †The stop-strengthening variant in *BCL2L13* affects a minor transcript isoform, and is also annotated as a synonymous mutation on the primary *BCL2L13* transcript.

Using exome sequencing from the PMBB, we identified heterozygous and homozygous individuals carrying UTC and stop-strengthening mutations. For the former class, we focused on variants introducing UAA stop codons, as the heightened MAPS score for such variants implied these mutations would be most deleterious. Filtering for variants with at least 5 heterozygous carriers with high-quality genotype, we identified 10 variants matching the above criteria (6 stop-strengthening mutations, 4 UAA-UTC mutations). For each of these mutations we performed a single-variant PheWAS across 800 EHR phenotypes. Of those 10 candidates, 6 passed an FDR threshold of 0.1 ($P < 1.25e-4$) used in previous PheWAS studies [156,157], including 5/6 of the stop-strengthening variants and 1/4 of the UAA UTCs. Even more strikingly, two of these six variants passed a highly conservative Bonferroni correction ($P < 6.25e-6$), both being uORF stop-strengthening variants. The stop-strengthening variant in *PMVK* was associated with increased risk of Type 1 diabetes while the stop-strengthening variant in *VPS53* was associated with a protective effect against anxiety disorders (Figure 3.9, Figure 3.16, Table 1, Table 3.4). Notably, of the identified phenotype-associated variants, only *VPS53* is annotated as a cis-eQTL in the latest GTEx release (version 8).

To replicate associations from our exploratory analysis in the PMBB, we performed additional single-variant association analyses for each of the 6 significant variant-phenotype associations in the UK Biobank (UKB). Direct replication using the original significant 4- or 5-digit ICD-9 code from the PMBB was tested for each variant-phenotype association. Where there were insufficient case numbers in the UKB, we used the broader 3-digit ICD-9 code. Out of six novel associations

reaching FDR < 0.1, one (*rs140799351*) showed $P < 0.05$ in the UKB at the 5-digit ICD code level, reaching study-wide significance (Table 3.1, Table 3.5). For the remaining putative novel associations, the *VPS53* uORF stop-strengthening variant did not replicate, although the direction of effect is consistent with results from the PMBB. Finally variants in *SHMT2* could not be replicated because there were fewer than 20 cases in the UKB cohort, and *MOAP1* could not be replicated because this variant was absent from the UKB.

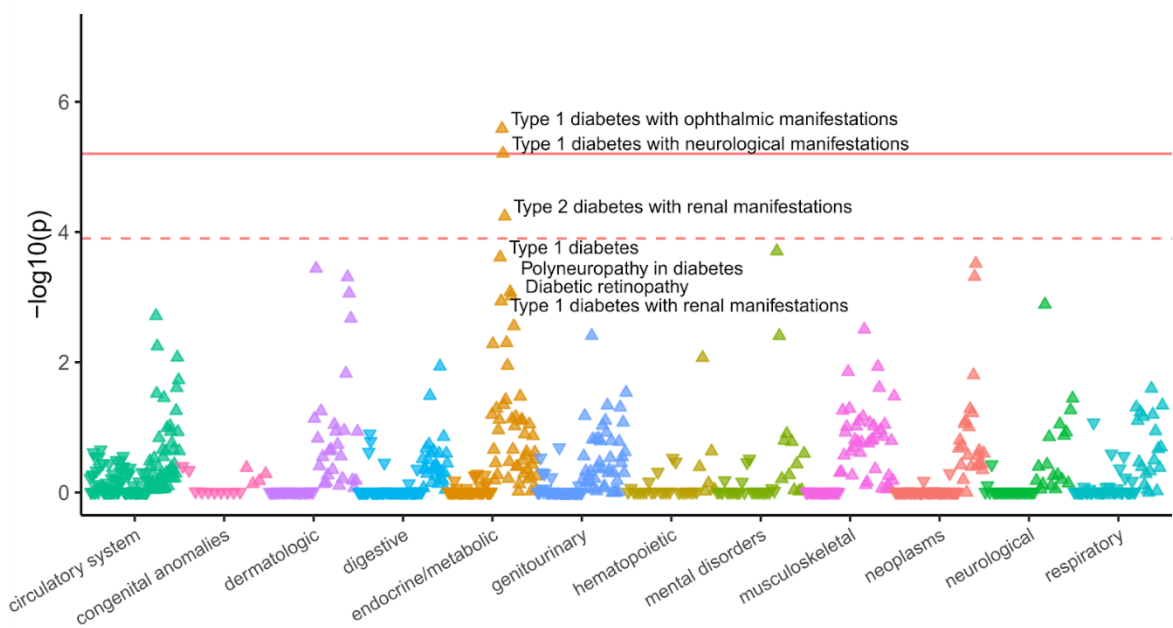


Figure 3.9: Phenome-wide association study (PheWAS) of predicted stop-strengthening variant in a translated uORF in PMVK. PheWAS plot of translated uORF stop-strengthening variant in the 5'UTR of PMVK (N = 65 carriers) in the Penn Medicine BioBank. ICD-9 and ICD-10 Phecodes are organized and plotted by category on the X-axis. The solid red line represents the threshold for Bonferroni-adjusted significance ($P=6.25e-6$) and the red dashed line represents the FDR threshold ($P=1.25e-4$). The direction of each arrowhead corresponds to increased risk (up) or decreased risk (down).

3.9: Disease-associated uORF variants change protein expression

To elucidate the possible biological consequences of UTC and stop-strengthening mutations, we selected three PheWAS association signals in the discovery analysis for functional assessment. To determine if these variants could affect protein expression, we measured the expression of a set of dual-luciferase reporters in HEK293T cells for *PMVK*, *VPS53*, and the *BCL2L13* uORF variants. We compared the expression of the wild-type 5'UTR sequence for *PMVK*, *VPS53*, and *BCL2L13* cloned upstream of a Firefly Luciferase ORF to two variant sequences - one with the predicted uORF start codon removed, and a second sequence with the PheWAS-significant stop-strengthening mutation inserted. For *VPS53*, we also tested the effect of a mutation changing a tryptophan UGG codon to a UAG UTC (Figure 3.10b). Across all constructs, we observed a significant reduction in expression of the downstream ORF when the PheWAS-significant stop-strengthening mutation was introduced (Figure 3.10). Introducing a new UTC in the 5'UTR of *VPS53* also significantly reduced reporter protein expression relative to the wild-type sequence. Similar results were obtained from assays performed in HeLa cells (Figure 3.22).

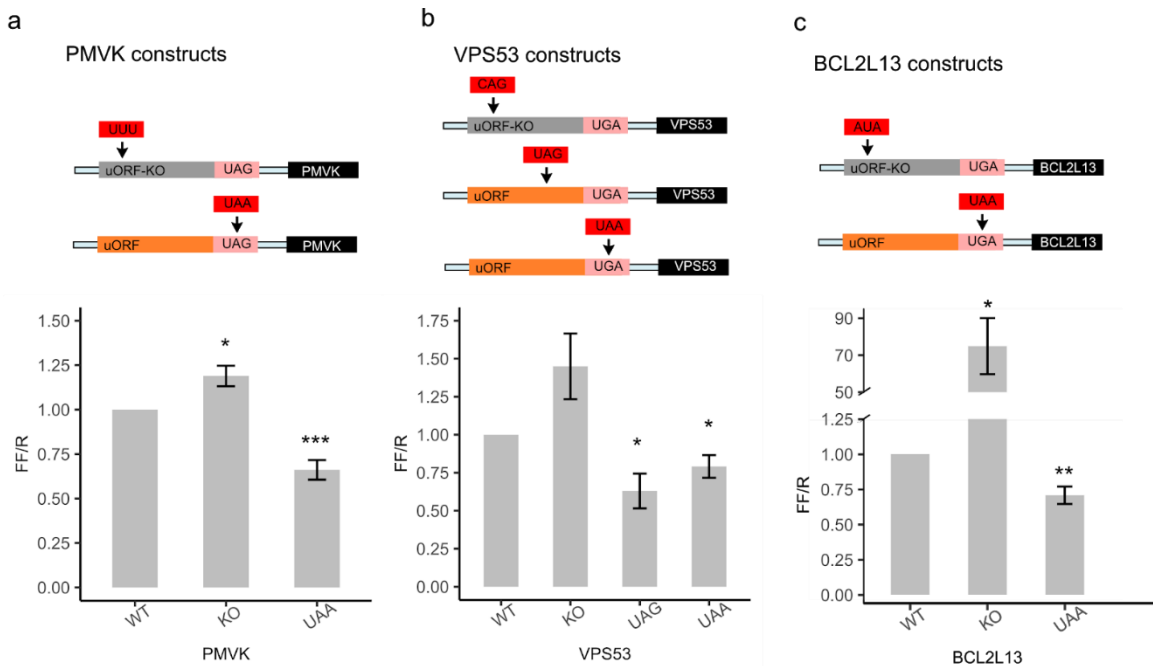


Figure 3.10: Reporter gene assays for translated uORF stop-introducing and stop-strengthening variants. Dual-luciferase reporter assay quantifies relative expression for uORFs with UTC and stop-strengthening mutations associated with EHR phenotypes by PheWAS. Experimental 5'UTRs for (a) *PMVK*, (b) *VPS53*, and (c) *BCL2L13* for uORF KO, stop-strengthened, or stop-introduced variants are shown. Bars represent co-transfected Firefly to Renilla Luciferase luminescence ratios normalized to associated wild-type 5'UTRs in HEK293T cells measured 48 hours post-transfection. Significant P-values from one-sample T-test for each condition denoted by * (>0.05), ** (>0.01), and *** (>>0.001). Error bars represent mean + S.E.M. of at least 3 independent experiments.

In all the tested constructs, UTC and stop-strengthening mutations decreased relative Firefly expression. These data are consistent with the hypothesis that UTC or stop-strengthening mutations are under negative selection because they decrease the probability of translation initiation at downstream coding sequences. These results are congruous with our genetic analysis, and imply that UTC and stop-strengthening mutations represent a new class of functional variation in 5'UTRs capable of causing loss-of-function of downstream coding genes.

3.10: Replication of novel associations by loss-of-function gene-burden studies

Results from reporter-gene experiments showed that UTC and stop-strengthening mutations could decrease expression of the downstream protein for *PMVK*, *VPS53*, and *BCL2L13*. Our

findings implied that uORF UTC and stop-strengthening mutations cause phenotypic consequences through potential loss-of-function of the downstream protein-coding gene. To further validate this hypothesis, we performed a gene burden test by aggregating rare loss-of-function protein-coding variants in the PMBB and UKB for each significant uORF-PheWAS association. These studies could confirm that predicted loss-of-function in the protein coding sequence of the uORF-regulated gene causes the same phenotype as the uORF UTC or stop-strengthening mutations. Indeed, similar loss-of-function gene burden approaches using rare protein-coding variants have successfully been applied to identify both known and new gene-disease associations in the PMBB and UKB [155,158].

Of six PheWAS-significant associations uncovered in our discovery analysis (FDR<0.1), two associations were replicated by an independent loss-of-function gene burden test in either the UKB or PMBB. The associations between *PMVK* and diabetes, and *SHMT2* and diseases of the salivary gland, were replicated in the UKB and PMBB respectively (*PMVK* P=0.00727, *SHMT2* P=0.005515, Table 1, Table 3.5). Although no significant LOF-burden association for *PMVK* was replicated in the PMBB, predicted loss-of-function of *PMVK* was nominally associated with impaired fasting glucose (P=0.0235). A second uORF-disease association was replicated for *NALCN* and the parent 3-digit parent PheCode of disorders of plasma protein metabolism in the UKB (P=0.0264). Gene-disease associations for *BCL2L13* could not be replicated in either the PMBB or UKB due to lack of carriers for predicted loss-of-function variants. Ultimately this analysis confirmed that loss-of-function gene burden tests using protein-coding variants are associated with the same phenotype for two uORF stop-strengthening mutations. This evidence of allelic heterogeneity for these phenotypes further strengthens the likelihood that uORF stop-strengthening variants can cause loss-of-function of downstream protein-coding genes.

3.11: Summary and future directions

By combining large databases of human genetic variation with ribosome profiling, we identified two new categories of mutations in 5'UTRs capable of causing loss-of-function in downstream coding genes. These mutations either introduce upstream termination codons in uORFs or strengthen uORF stop sites. Given that ~50% of human protein-coding genes are estimated to be under translational control by uORFs, these findings provide a novel framework for interpreting the functional significance of 5'UTR variation for a large fraction of human genes.

Using these mutations, we additionally identified new gene-disease associations in the PMBB and replicated one of these associations in independent single-variant association tests in the UKB. Two associations involving stop-strengthening variants in *PMVK* and *SHMT2* and one involving a UTC in *NALCN* were also replicated using protein-coding mutations in loss-of-function gene burden tests. These results provide independent validation of uORF variant-phenotype associations uncovered through the PMBB discovery analysis and demonstrate that uORF stop-strengthening and UTC mutations associate with the same phenotype as predicted loss-of-function coding mutations in downstream coding sequences. In support of these conclusions, we have shown that introducing UTCs and stop-strengthening mutations in translated uORFs decreases protein expression of downstream genes in reporter assays. These findings establish that uORF UTC and stop-strengthening mutations can have functional consequences on protein expression and are associated with disease in humans. If we assume that pathogenic UTC or stop-strengthening mutations are under similar selective pressures as pathogenic loss-of-function variants in protein-coding regions of the genome, we estimate that approximately 24% (90% CI 21-28%) of uORF-containing genes may be affected by UTC and stop-strengthening mutations with severe pathogenic consequences (3.13: Suppl. Note). Moreover, if we assume that missense mutations in functional uORF micropeptides are similarly enriched in singletons as in protein-coding regions of the genome, we estimate that ~5-15% of translated uORFs are under constraint for amino acid function (3.13: Suppl. Note). This latter estimate is consistent with

recent CRISPR screens reporting a statistically significant decrease in growth phenotypes for ~14% (157/1098) of uORF-specific knockouts across two cell lines when the CDS was preserved [107]. Finally, of the 4392 genes with translated uORFs used for this analysis, 1121 (26%) are also annotated as having pathogenic coding sequence variants in ClinVar, suggesting that UTC and stop-strengthening mutations in these genes may have additional utility for the diagnosis of rare disease.

Our results suggest uORF translation has broad roles in regulating CDS translation. Translation initiation is rate-limiting for protein production and selection against mutations disrupting translation elongation (UTCs) or termination at uORFs (stop-strengthening variants) may reflect the importance of preserving translation initiation efficiency at the CDS. This suggested mode of regulation is in-line with observations that cis-regulatory relationships between uORFs and downstream coding sequences are frequently conserved across vertebrates while features conferring strong uORF repression are less maintained [122,124]. For stop-strengthening variants, the increased translation termination efficiency could accelerate ribosomal release from the mRNA transcript, thus decreasing downstream CDS translation. This mechanism is consistent with previous data in human cell lines showing that decreased translation termination efficiency by global knockdown of eRF3A increases translation of genes under uORF-repression [159, 160]. For UTC mutations, the introduction of stop codons in the uORF may lead to either ribosome stalling and subsequent collisions that further repress CDS expression [161,162]. This early translation termination in uORFs might also facilitate greater rates of premature ribosome release from the mRNA transcript, or can lead to nonsense mediated decay (NMD). While a handful of translated uORFs that activate NMD have been described in the literature [163–165], whether uORF-activated NMD broadly regulates protein expression remains an open question. Indeed, depletion of UPF1, a central component of the canonical NMD pathway, produced only minimal changes in uORF-containing mRNAs abundance in human cell lines [159].

The capacity for translated uORFs to produce functional micropeptides independent of regulating CDS expression remains an area of active investigation. In canonical protein-coding regions of the genome, amino acid substitutions in critical protein domains can be highly deleterious for cellular functioning and fitness. Previous studies have found that uORF-encoded peptides show evidence of amino acid conservation using statistical tests relying on a null hypothesis of neutral selection [121]. It is unclear if the conclusions drawn from these approaches account for the possibility that codon-optimality constrains variation within uORFs rather than amino acid identity. In contrast, we do not observe similar constraints on missense mutations in translated uORFs, suggesting that amino acid substitutions within most uORF-encoded micropeptides are well-tolerated in humans. This was also the case for other non-canonical translated ORFs, including 3'UTR ORFs, pseudogenes, and lncRNAs, that are not thought to widely encode for functional micropeptides. Although a handful of functional micropeptides have been identified previously, our analysis implies that most ncORFs do not produce peptide products whose function depends on their amino acid composition. It is also important to note that ribosomes are among the most abundant proteins within cells, occupying approximately 5% of the entire intracellular volume [166]. As improvements in ribosome profiling facilitate deeper characterization of the translome, observations of widespread translation in non-canonical ORFs should be interpreted cautiously in light of potential functionality.

Interestingly, bi-allelic loss-of-function mutations in *SHMT2* have recently been described in a novel brain and heart developmental syndrome involving spastic paraparesis and ataxias [167]. Indeed, in addition to the genome-wide significant association with diseases of the salivary gland uncovered in our study, the SHMT2 uORF stop-strengthening variant was nominally associated with several Phecodes related to cardiac and movement disorders in the PMBB (Table 3.7), including Congenital anomalies of the great vessels (ICD 747.13, $P = 0.0117$), Abnormal involuntary movements (350.1, $P = 0.0238$), Abnormality of gait (350.2, $P = 0.02575$), Mobitz II AV block (426.22, $P = 0.03432$), and Arrhythmia (cardiac) NOS (427.5, $P = 0.04977$).

These additional nominal associations suggest that *SHMT2* uORF variants may be capable of contributing to similar phenotypic consequences as described in loss-of-function mutation carriers, however further studies are needed to investigate this possibility. The novel association between stop-strengthening and pLOF variants in *PMVK* with diabetes further strengthens existing genetic and epidemiological evidence linking the mevalonate pathway to diabetes. *PMVK* encodes for phosphomevalonate kinase, an enzyme in the mevalonate pathway catalyzing the conversion of mevalonate-5-phosphate to mevalonate-pyrophosphate downstream of HMG-CoA reductase. Multiple randomized clinical trials have shown that inhibiting HMG-CoA reductase with statins increases the risk of developing new-onset type 2 diabetes in a dose-dependent manner, although the mechanism driving this association has remained elusive [168–170]. Moreover, genetic variants in and near the *HMGCR* gene that are associated with lowered LDL cholesterol levels have been similarly shown to confer an increased risk of developing diabetes [171,172], suggesting that decreased *HMGCR* activity contributes to diabetes pathogenesis. Our data is the first to establish a putative link between *PMVK* and diabetes. Given the shared involvement of *PMVK* and *HMGCR* genes in the mevalonate pathway, it is possible that variants in both these genes confer an increased risk of diabetes through a similar mechanism, however additional studies will be needed to further elucidate the precise relationship between *PMVK* and diabetes.

A limitation of our analysis is that we cannot directly assess the impact of additional factors on uORF-mediated translational regulation. As an example, a pathogenic UTC mutation in the *U2HR* gene has previously been reported to confer gain-of-function in Marie Unna hereditary hypotrichosis [173]. However, missense variants in this uORF also confer gain-of-function effects, suggesting that these mutations contribute to pathology through disrupting a functional micropeptide. Indeed, previous studies have shown that a multitude of factors may impact uORF regulatory function and dissecting these effects remains a challenge for future studies.

Finally, we note that being a hospital-based biobank, participants in the PMBB are generally less healthy than the general population. As phenotypes within broader disease Phecode families are often highly correlated, we sought to replicate associations uncovered in the discovery analysis by first testing for a specific hypothesis-driven phenotype association in addition to related phenotypes in the corresponding Phecode families. We recognize that controlling for Type 1 error in this framework remains challenging. However, to remedy this we sought additional confidence by further replicating significant uORF-variant associations through loss-of-function gene-burden analyses. Moreover, the relative enrichment in diseased individuals in the PMBB may account for why few associations discovered in our analysis of the PMBB are replicated in the UKB which contains a healthy volunteer selection bias [174]. Indeed, we were unable to test for an association for two of the six PMBB associations due to an inadequate number of individuals having the phenotype in UKB. As hospital-based biobanks become more prevalent these unreplicated associations should be revisited and confirmed.

Understanding and interpreting the impact of noncoding genetic variation is a fundamental challenge in biology. Many mutations affecting uORFs are known to cause disease [175–178], but until now, most studies have focused on mutations which abolish start codons, stop codons of existing uORFs, or those that create new inhibitory uORFs. By examining patterns of genetic variation within translated uORFs, we have uncovered two new categories of variation affecting 5'UTRs that may lead to loss-of-function in associated genes. We have used these variants to identify new gene-disease associations, and provide evidence for their ability to impact downstream gene expression. Our approach demonstrates the power of integrating population-scale databases of human genetic variation with cellular-scale -omics data to identify new patterns of how variation impacts regulatory elements. Taken together, our data broadens the scope of functional translational regulation by uORFs in the transcriptome and establishes new approaches for interpreting functional genetic variation in 5'UTRs.

3.12: Supplementary Figures and Tables for Disrupting upstream translation in mRNAs is associated with human disease

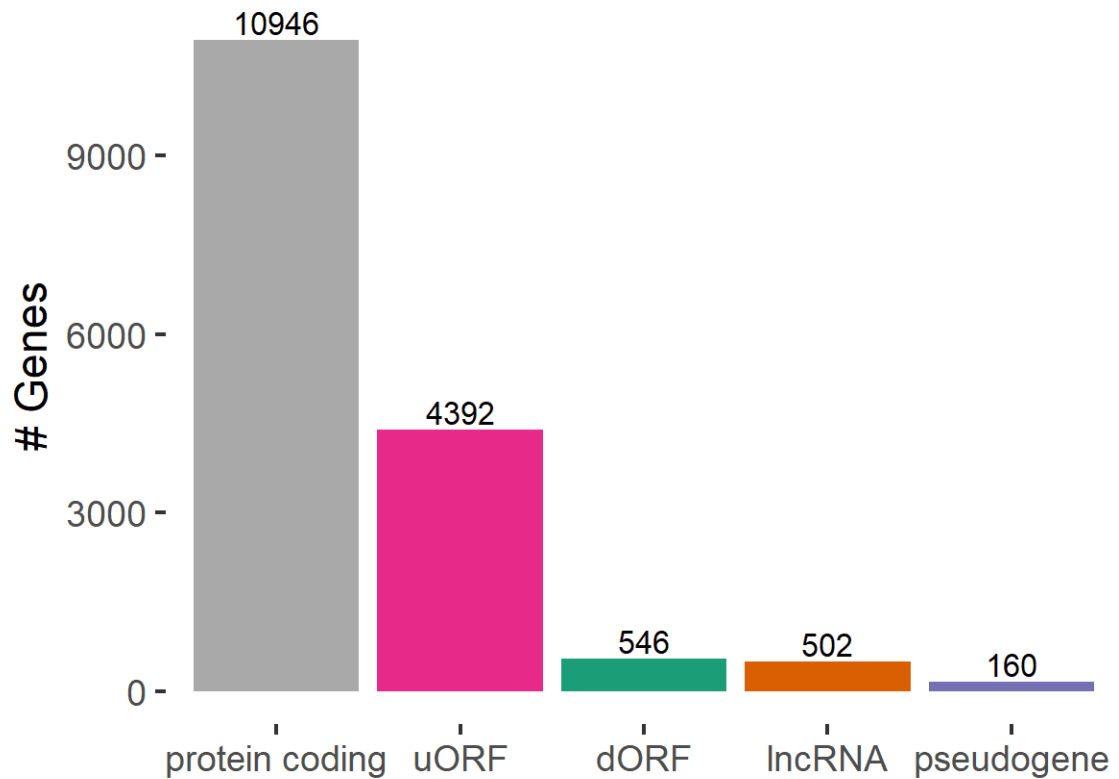


Figure 3.11: Distribution of protein coding ORFs, uORFs, and other non-canonical ORFs mapped by ribosome profiling from Ji et. al paper [121]. dORFs represent ORFs mapped in 3'UTRs, lncRNAs represent ORFs mapped in long-noncoding RNAs, and pseudogenes represent translated pseudogenes respectively.

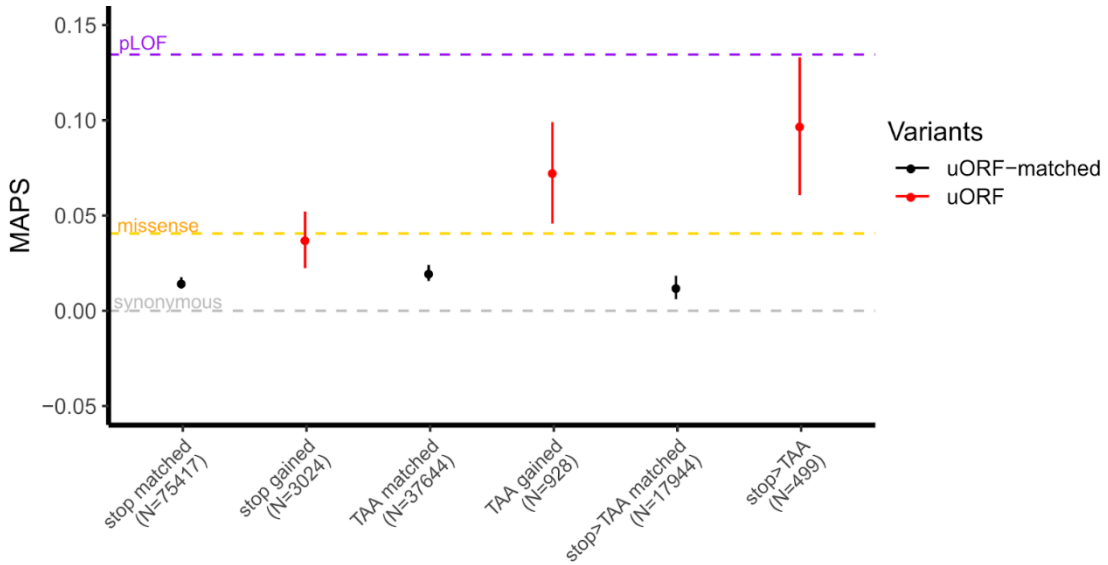


Figure 3.12: MAPS scores for uORF UTC-creating and stop-strengthening variants compared to non UTC-creating or stop-strengthening uORF variants matched by trinucleotide mutation context. Error bars represent bootstrapped 90% confidence intervals.

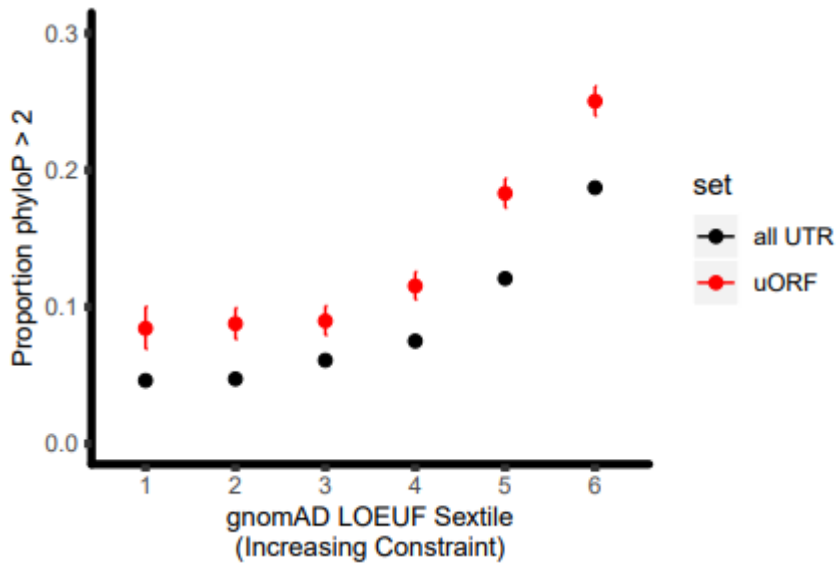


Figure 3.13: PhyloP scores for possible UTC-creating positions and gnomAD protein-coding constraint. PhyloP scores in translated uORFs (red) compared to 5'UTR sequences (black) across all sextiles of gene constraint as determined by gnomAD LOEUF scores (1 being least constrained, 6 being most constrained).

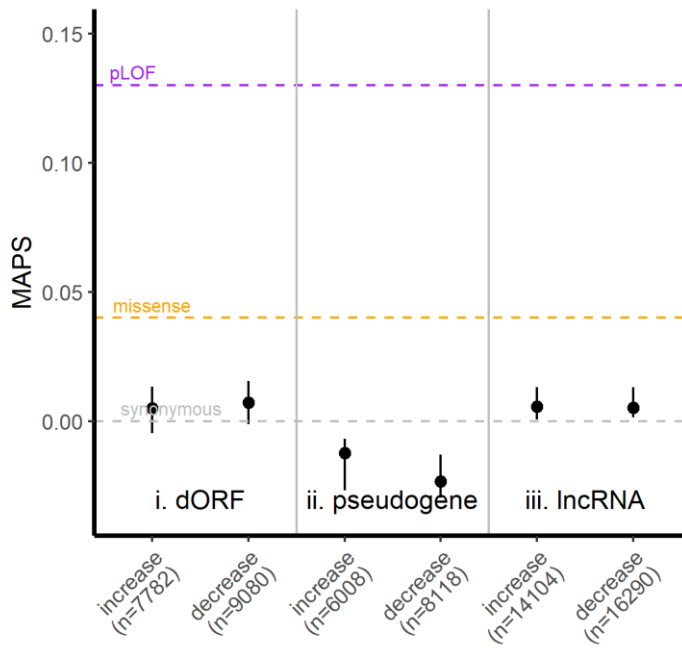
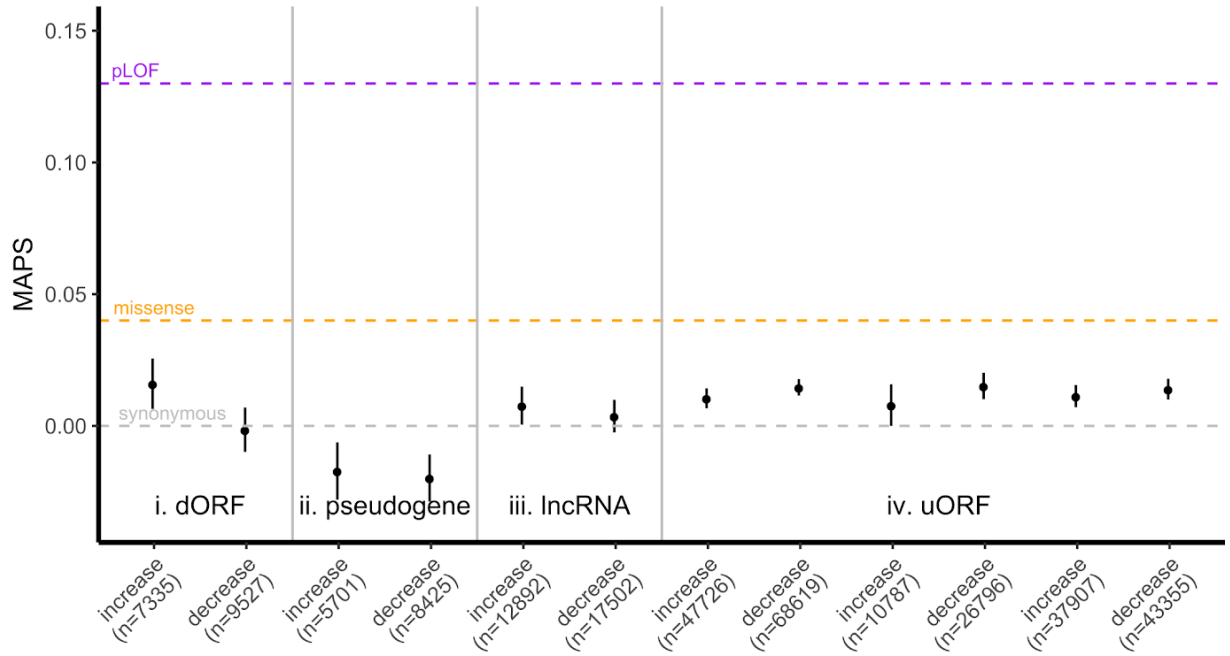


Figure 3.14: Optimality changing MAPS scores for SNVs in dORFs (3'UTRs), pseudogenes, and long-noncoding RNAs (lncRNAs). Error bars represent bootstrapped 90% confidence intervals.

a. CSC from 293 orfome library



b. CSC from RPE endogenous mRNAs

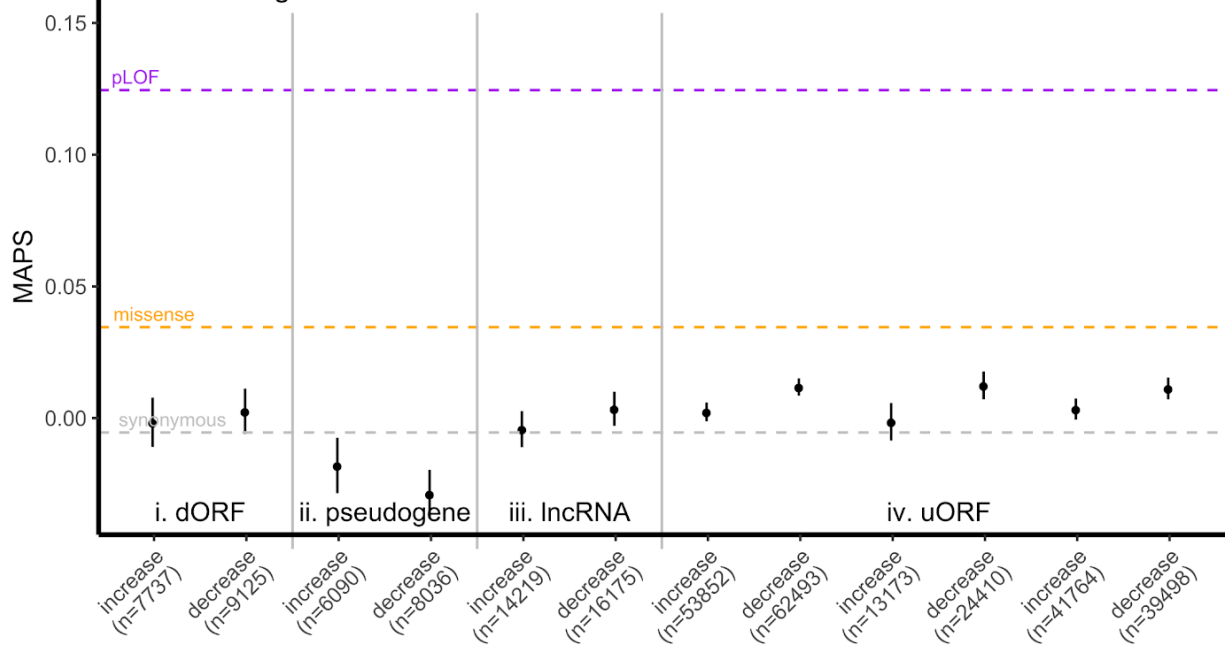


Figure 3.15: MAPS score for optimality changing variants using different optimality scores. MAPS dependence on whether a SNV increases or decreases codon optimality in uORFs is robust to changing CSC-scores used to calculate codon optimality across 293 cell lines using the orfome approach, and from retinal pigment epithelium cells with CSC-scores calculated using endogenous mRNAs [153].

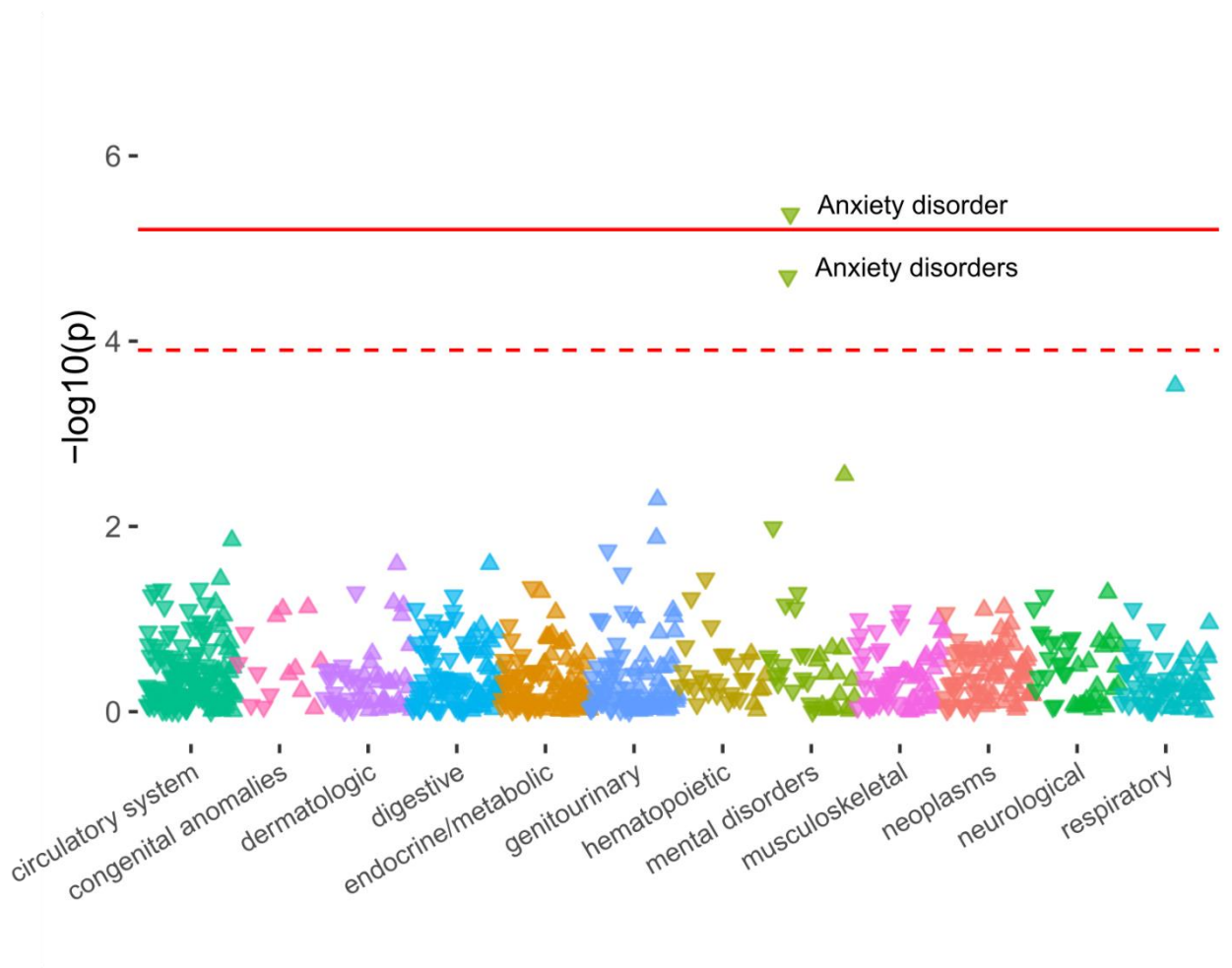


Figure 3.16: PheWAS plot of *VPS53* stop-strengthening variant. Red solid line indicates Bonferroni significance threshold ($P=6.25e-05$). Red dashed line represents the FDR < 0.1 threshold.

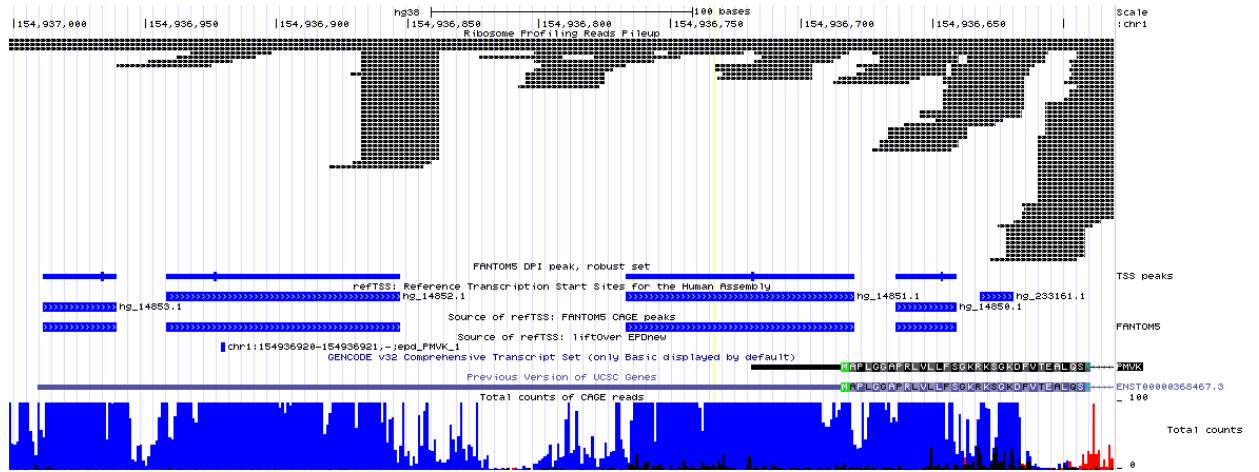


Figure 3.17: Change in PMVK 5'UTR annotation as of September 2019 Gencode 32 release. The longer 5'UTR isoform for PMVK is supported by transcription start-site mapping from FANTOM5, and by the remapped ribosome-profiling reads (top, black) from [GSE65885](#).

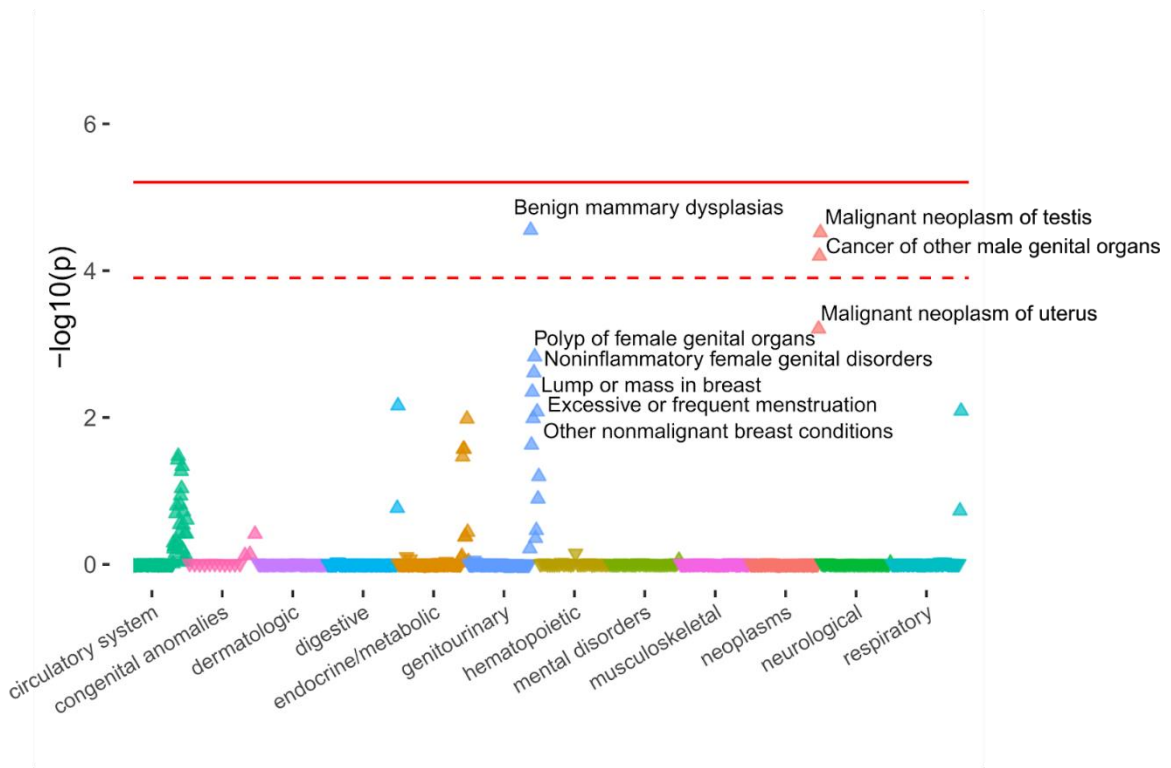


Figure 3.18: PheWAS plot of *BCL2L13* stop-strengthening variant. Red solid line indicates Bonferroni significance threshold ($P=6.25e-05$). Red dashed line represents the FDR < 0.1 threshold.

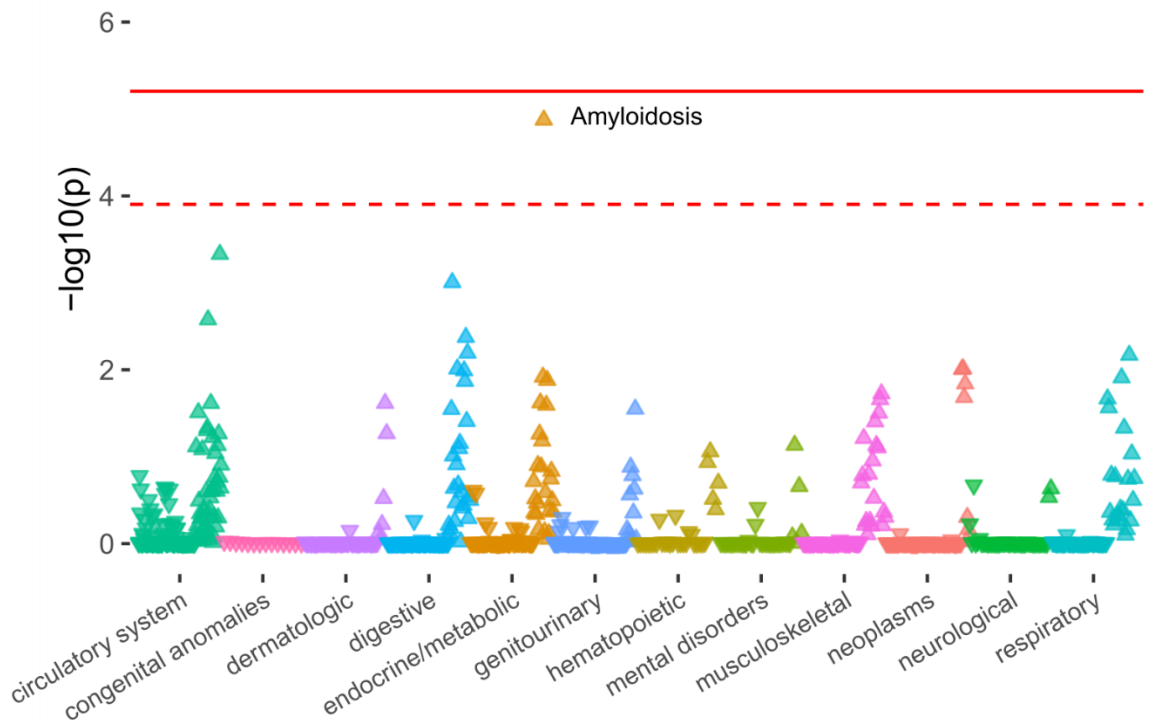


Figure 3.19: PheWAS plot of *NALCN* UAA UTC variant. Red solid line indicates Bonferroni significance threshold ($P=6.25 \times 10^{-5}$). Red dashed line represents the FDR < 0.1 threshold.

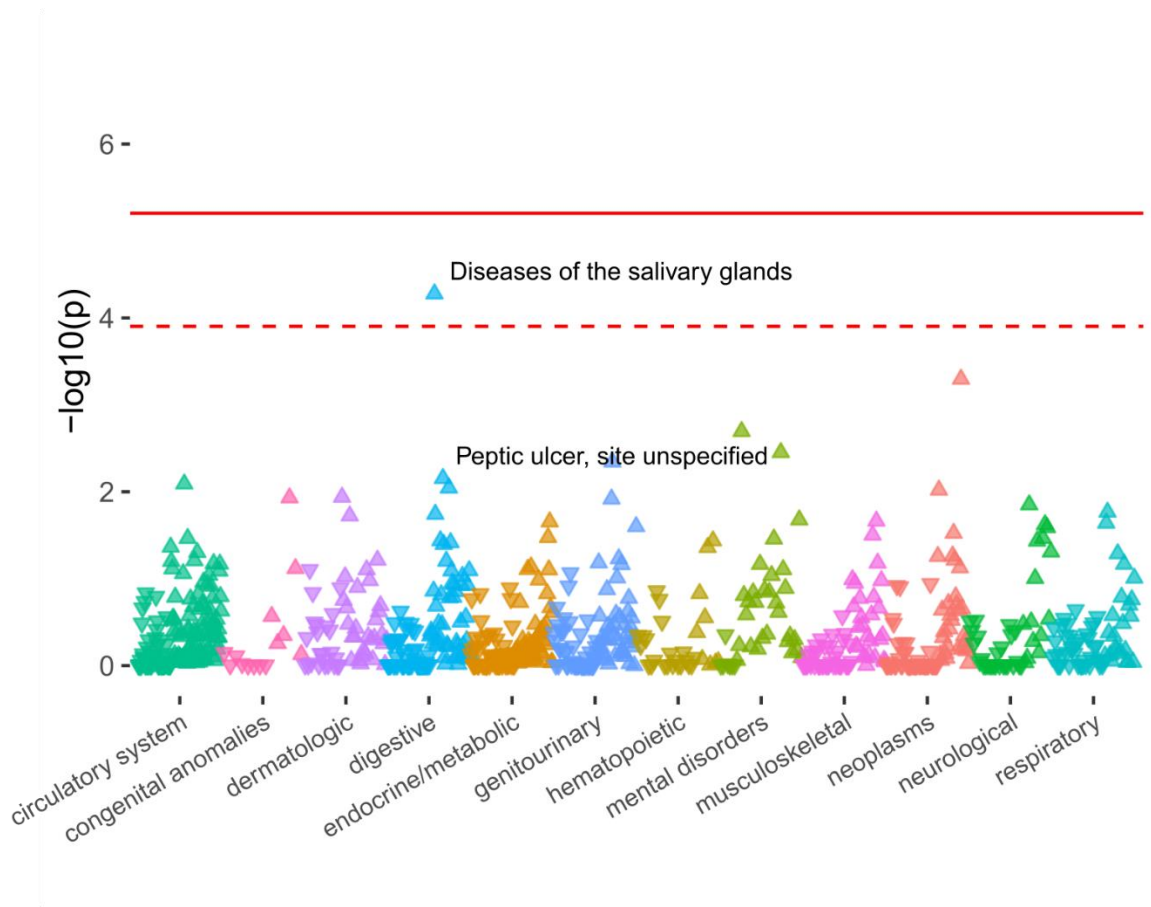


Figure 3.20: PheWAS plot of *SHMT2* stop-strengthening variant. Red solid line indicates Bonferroni significance threshold ($P=6.25e-05$). Red dashed line represents the FDR < 0.1 threshold.

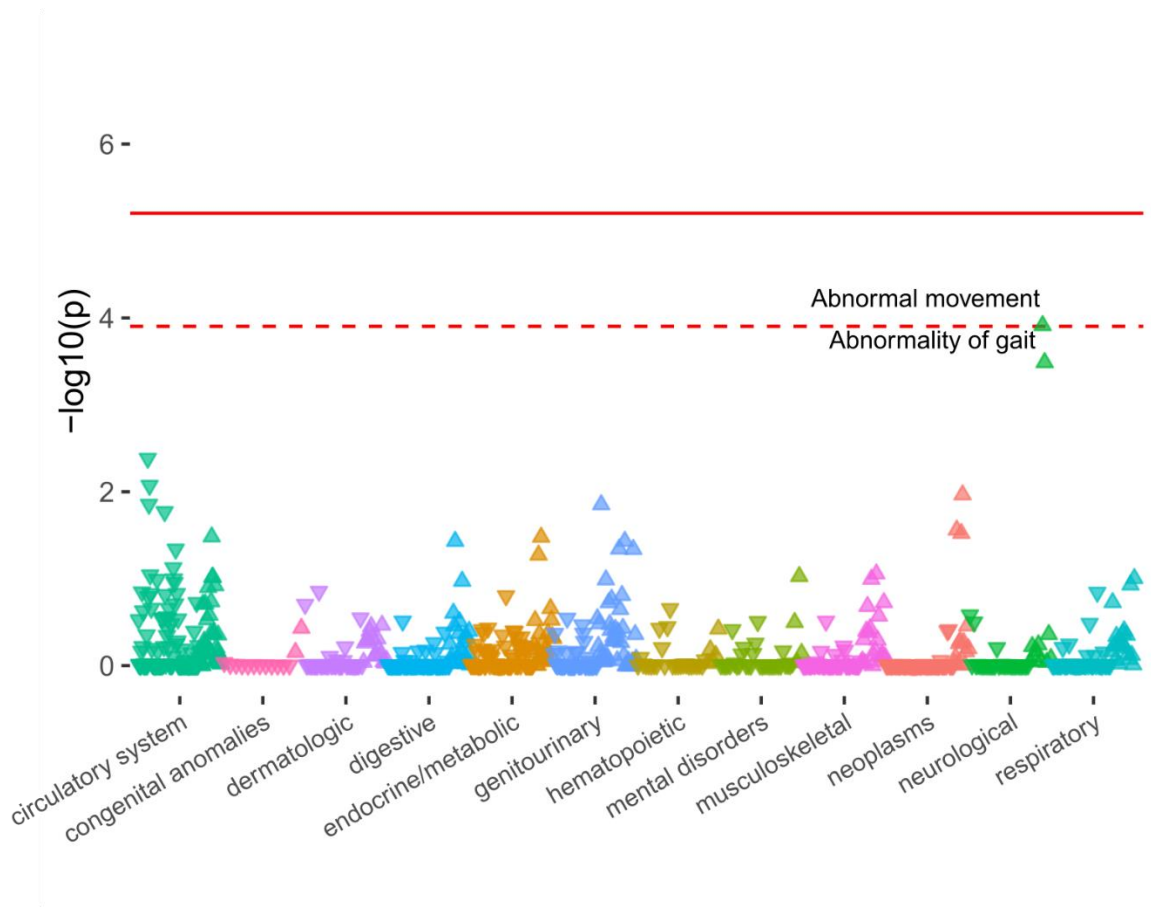


Figure 3.21: PheWAS plot of *MOAP1* UAA UTC variant. Red solid line indicates Bonferroni significance threshold ($P=6.25 \times 10^{-5}$). Red dashed line represents the FDR < 0.1 threshold.

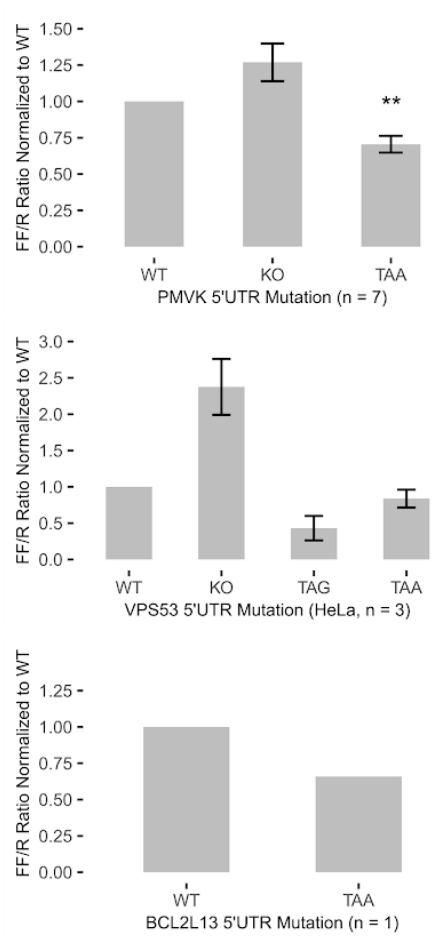


Figure 3.22: Luciferase experiments for *PMVK* and *VPS53* plasmid Constructs showing similar direction of effect for UTC and stop-strengthening variants using HeLa cells for transfection.

Table 3.2: uORF UTC / Stop-strengthening MAPS analysis with all CDS-overlapping variants removed.

uORF Variant Type	Number of SNVs	SNVs overlapping any CDS (%)	Original MAPS (95% CI)	MAPS with CDS-overlap removed (95% CI)
UTC	3024	163 (5.39%)	0.0373 (0.0194-0.0552)	0.0362 (0.0172-0.0552)
TAA gained	928	93 (10.02%)	0.0726 (0.0407-0.1045)	0.0765 (0.0430-0.1101)
Stop > TAA	499	51 (10.22%)	0.0971 (0.0537-0.1404)	0.1051 (0.0595-0.1507)

Table 3.3: Relative frequencies of TGA, TAG, and UAA trinucleotides across different 5'UTR sequence contexts

Trinucleotide	Context	Proportion	2.5% quantile	97.5% quantile
UAA	All 5'UTRs	0.354813	0.347512	0.362177
UAA	5'UTRs w/ uORF	0.375554	0.355256	0.396686
UAA	uORF	0.196608	0.17912	0.214626
UAG	All 5'UTRs	0.227442	0.221035	0.233975
UAG	5'UTRs w/ uORF	0.229501	0.211289	0.248058
UAG	uORF	0.285639	0.265501	0.305776
UGA	All 5'UTRs	0.417745	0.410151	0.425083
UGA	5'UTRs w/ uORF	0.394945	0.3739	0.415847
UGA	uORF	0.517753	0.495495	0.540541

Table 3.4: Minor allele frequencies for all PheWAS-significant variants tested in discovery and replication analyses

Gene	Carrier Freq (PMBB)	PMBB EUR	PMBB AFR	Carrier Freq (UKB)	LOF Allele Freq. (PMBB)	LOF Allele Freq. (UKB)
PMVK	0.00298	0.00336	0.00000	0.00324	0.00086	0.00094
VP553	0.09468	0.10954	0.01940	0.14370	0.00131	0.00109
NALCN	0.00147	0.00129	0.00106	0.00084	0.00122	0.00162
BCL2L13	0.00028	0.00039	0.00190	0.00029	0.00028	0.00025
SHMT2	0.01266	0.00540	0.03715	0.00749	0.00182	0.00033
MOAP1	0.00188	0.00006	0.00793	N/A	0.00026	N/A

Table 3.5: PheWAS replication analyses phenotypes tested

Variant	Gene	PheCode	Cases (UKB)	Controls (UKB)	Cases (PMBB)	Controls (PMBB)	Single Variant		pLoF Gene Burden			
							OR (UKB)	Rep. P (UKB)	OR (UKB)	Rep. P (UKB)	OR PMBB	Rep. P (PMBB)
rs181302437	PMVK	250.13	33	32689	23	5198	3.3E-06	0.986	15.82	0.0073	2.46E-05	0.9887
rs181302437	PMVK	250.14	13	32689	25	5198	N/A	N/A	N/A	N/A	2.48E-05	0.989
rs181302437	PMVK	250.22	11	32689	315	6134	N/A	N/A	N/A	N/A	5.96E-05	0.9672
rs35915949	VPS53	300.1	627	31247	1060	6939	0.88	0.1631	1.35	0.6751	1	0.8052
rs35915949	VPS53	300	684	31247	1249	6939	0.88	0.1757	1.48	0.5833	1	0.5808
rs139848407	NALCN	270.33	11	34565	30	7727	N/A	N/A	N/A	N/A	7.25E-06	0.9892
rs139848407	NALCN	270	34	34554	134	9594	9.35E-06	0.9887	9.74	0.0264	4.42	0.1518
rs140799351	BCL2L13	610	277	33848	55	7689	1.77E-04	0.973	N/A	N/A	N/A	N/A
rs140799351	BCL2L13	187	51	33957	26	7700	56.02	0.0003	N/A	N/A	N/A	N/A
rs140799351	BCL2L13	187.2	30	33957	34	7700	79.78	0.0002	N/A	N/A	N/A	N/A
rs28365863	SHMT2	527	N/A	N/A	90	9774	N/A	N/A	N/A	N/A	2	0.0055
rs116450723	MOAP1	350	N/A	N/A	362	9415	N/A	N/A	N/A	N/A	2.35E-05	0.9659

Table 3.6: 5'UTR Fragments used in expression constructs

Gene (Transcript)	Mutation	Sequence
PMVK (ENST00000368467)	WT	TATAGGGAGACCCAAGCTGGCTAGTTAAGCTTAGATCTTG ATATCCTCGAGAGAAGGTTCTGGGCGGGGCTGGACTGTT CTAAGTGAGTTCGGGTGGGGGAGCTTACAGAGGGGAGG CTGCTCTGTGAAGGAACCGCCTTTCTCTCCGCGTGTCTCA CCCTTTTCTCCCCATATCTGTTTGGACATGAGCTGAGGGC ACGGTCGCGGGCGGTGAGCCCTGTTTCGCAGCTACGGCG AGGAGGGGCGCGATTGTTCTTGTGCGGCTCCGCTTAG TGGCCGCGTCCATTCCGCGCGGTGTCCCGATTTAGGGG TAGGGAGAAGTGTCAGCTTCAGGCATCGCGAGGCGTGGC GGCCCATGGAAGATGCCAAAACATTAAGAAGGGCCCA GCGCCATTCTACCCACTCGAAGACGGGACCGCCGGCGAG CAGCTGCACAAAGCCATGA
PMVK (ENST00000368467)	KO	TATAGGGAGACCCAAGCTGGCTAGTTAAGCTTAGATCTTG ATATCCTCGAGAGAAGGTTCTGGGCGGGGCTGGACTGTT CTAAGTGAGTTCGGGTGGGGGAGCTTACAGAGGGGAGG CTGCTCTGTGAAGGAACCGCCTTTCTCTCCGCGTGTCTCA CCCTTTTCTCCCCATATCTGTTTGGACATGAGCTGAGGGC ACGGTCGCGGGCGGTGAGCCCTGTTTCGCAGCTACGGCG AGGAGGGGCGCGATTGTTCTTGTGCGGCTCCGCTTAG TGGCCGCGTCCATTCCGCGCGGTTTCCCGATTTAGGGG AGGGAGAAGTGTCAGCTTCAGGCATCGCGAGGCGTGGC GCCCCATGGAAGATGCCAAAACATTAAGAAGGGCCAG CGCCATTCTACCCACTCGAAGACGGGACCGCCGGCGAGC AGCTGCACAAAGCCATGA
PMVK (ENST00000368467)	TAG>TAA	TATAGGGAGACCCAAGCTGGCTAGTTAAGCTTAGATCTTG ATATCCTCGAGAGAAGGTTCTGGGCGGGGCTGGACTGTT CTAAGTGAGTTCGGGTGGGGGAGCTTACAGAGGGGAGG CTGCTCTGTGAAGGAACCGCCTTTCTCTCCGCGTGTCTCA CCCTTTTCTCCCCATATCTGTTTGGACATGAGCTGAGGGC

		ACGGTCGCGGGCGGTACGCCCTGTTGCGAGCTACGGCG AGGAGGGGCGCGATTGTTCTTGTGGCGCTCCGCTTAG TGGCCGCGTCCATTCCGCGCGGTGCCGATTGTAAGGG TAGGGAGAAGTGTACGCTTCAGGCATCGCGAGGCGTGGC GGCCCCATGGAAGATGCCAAAAACATTAAGAAGGGCCCA GCGCCATTCTACCCACTCGAAGACGGGACCGCCGGCGAG CAGCTGCACAAAGCCATGA
VPS53 (ENST0000 0437048)	WT	TATAGGGAGACCCAAGCTGGCTAGTTAAGCTTAGATCTTG ATATCCTCGAGACTGGGGCCTGGGTGGCGGCTGGAGGC CTGAGTTGGGCTCGCGCGGGGGTCCGCAGGGGGCCGG GTGGCGGAATGGAAGATGCCAAAAACATTAAGAAGGGCC CAGCGCCATTCTACCCACTCGAAGACGGGACCGCCGGCG AGCAGCTGCACAAAGCCATGA
VPS53 (ENST0000 0437048)	KO	TATAGGGAGACCCAAGCTGGCTAGTTAAGCTTAGATCTTG ATATCCTCGAGACAGGGGCCTGGGTGGCGGCTGGAGGC CTGAGTTGGGCTCGCGCGGGGGTCCGCAGGGGGCCGG GTGGCGGAATGGAAGATGCCAAAAACATTAAGAAGGGCC CAGCGCCATTCTACCCACTCGAAGACGGGACCGCCGGCG AGCAGCTGCACAAAGCCATGA
VPS53 (ENST0000 0437048)	TGA>T AA	TATAGGGAGACCCAAGCTGGCTAGTTAAGCTTAGATCTTG ATATCCTCGAGACTGGGGCCTGGGTGGCGGCTGGAGGC CTAAGTTGGGCTCGCGCGGGGGTCCGCAGGGGGCCGG GTGGCGGAATGGAAGATGCCAAAAACATTAAGAAGGGCC CAGCGCCATTCTACCCACTCGAAGACGGGACCGCCGGCG AGCAGCTGCACAAAGCCATGA
VPS53 (ENST0000 0437048)	TGG>T AG	TATAGGGAGACCCAAGCTGGCTAGTTAAGCTTAGATCTTG ATATCCTCGAGACTGGGGCCTGGGTAGCGGCTGGAGGCC TGAGTTGGGCTCGCGCGGGGGTCCGCAGGGGGCCGG GTGGCGGAATGGAAGATGCCAAAAACATTAAGAAGGGCC CAGCGCCATTCTACCCACTCGAAGACGGGACCGCCGGCG AGCAGCTGCACAAAGCCATGA
BCL2L13 (ENST0000 0543133)	WT	TCGGAGCACTCACCGCCGCTGGGGGACCCTGTCCGAAG CAACTGCCGCCGCCGCTCTTTTATCTCTTCTGGGGCAG GGGCCAGGGCCAGTTTTACACATCCATAAGTAGACCTTT TTGGAGCCTCACCAGCCAATTCAATGGCGTCTCTTCTAC TGTGCCT CTGGGATTTCACTATGAAACAAAGTATGTTGTTCTCAGCTA CTTGGGACTCCTCTCTCAAGAGAAGCTGCAAGAGCAACAT CTTTCCTCACCCCAAGGGTTCAACTAGATATAGCTTAC AATCTCTGGATCAAGAAATTTTATTAAGTTAAACTGAA ATTGAAGAAGAGCTAAAATCTCTGGACAAAGAAATTTCTGA AGGCCAGTGACATATCAGGCATTTCCGGGAATGTACTCTGG AGACCACAGTTCATGCCAGCGGCTGGAATAAGATTTTGGT GCCTCTGGTTTTGCTACGACAA
BCL2L13 (ENST0000 0543133)	ATG>A TA	TCGGAGCACTCACCGCCGCTGGGGGACCCTGTCCGAAG CAACTGCCGCCGCCGCTCTTTTATCTCTTCTGGGGCAG GGGCCAGGGCCAGTTTTACACATCCATAAGTAGACCTTT TTGGAGCCTCACCAGCCAATTCAATAGCGTCTCTTCTAC TGTGCCT CTGGGATTTCACTATGAAACAAAGTATGTTGTTCTCAGCTA

		<p>CTTGGGACTCCTCTCTCAAGAGAAGCTGCAAGAGCAACAT CTTTCCTCACCCCAAGGGTTCAACTAGATATAGCTTCAC AATCTCTGGATCAAGAAATTTTATTTAAAAGTTAAAAGTAA ATTGAAGAAGAGCTAAAATCTCTGGACAAAGAAATTTCTGA AGGCCAGTGACATATCAGGCATTTCTGGGAATGTACTACTGG AGACCACAGTTCATGCCAGCGGCTGGAATAAGATTTTGGT GCCTCTGGTTTTGCTACGACAA</p>
<p><i>BCL2L13</i> (ENST00000543133)</p>	<p>TGA>T AA</p>	<p>TCGGAGCACTCACCGCCGCTGGGGGACCCTGTCTGGGAAG CAACTGCCGCCGCCGCTCTTTTCATCTCTTCTGGGGCAG GGGCCAGGGCCAGGTTTTACACATCCATAAGTAGACCTTT TTGGAGCCTCACCAGCCAATTCAATGGCGTCTCTTCTAC TGTGCCT CTGGGATTTCACTATGAAACAAAGTATGTTGTTCTCAGCTA CTTGGGACTCCTCTCTCAAGAGAAGCTGCAAGAGCAACAT CTTTCCTCACCCCAAGGGTTCAACTAGATATAGCTTCAC AATCTCTGGATCAAGAAATTTTATTTAAAAGTTAAAAGTAA ATTGAAGAAGAGCTAAAATCTCTGGACAAAGAAATTTCTGA AGGCCAGTAAACATATCAGGCATTTCTGGGAATGTACTACTGG AGACCACAGTTCATGCCAGCGGCTGGAATAAGATTTTGGT GCCTCTGGTTTTGCTACGACAA</p>

Table 3.7: Nominal cardiac and movement disorder associations with SHMT2 stop-strengthening variant uncovered through PheWAS in Penn Medicine Biobank.

phenotype	beta	OR	SE	p	n_total	n_cases	n_controls	allele_freq	description
747.13	1.518	4.561	0.602	0.0117	7300	62	7238	0.005410959	Congenital anomalies of great vessels
350.1	1.124	3.077	0.497	0.0238	9519	104	9415	0.011503309	Abnormal involuntary movements
350.2	0.660	1.935	0.296	0.0258	9626	211	9415	0.011790983	Abnormality of gait
426.22	1.635	5.129	0.773	0.0343	2481	50	2431	0.005441354	Mobitz II AV block
972.1	1.434	4.197	0.703	0.0412	6285	40	6245	0.005966587	Cardiac rhythm regulators causing adverse effects in therapeutic use
427.5	0.859	2.362	0.438	0.0498	3412	157	3255	0.013335287	Arrhythmia (cardiac) NOS

3.13: Supplementary Note: Estimating the proportion of uORFs that may cause pathogenic loss-of-function equivalent consequences in ClinVar disease genes

We assume that MAPS score for UTC and stop-strengthening variants in uORFs represent the combination of variants from uORFs where loss-of-function is well-tolerated, and those from uORFs where loss-of-function is not tolerated (has severe impact on fitness). In other words, the MAPS score is a mixture of variants that are not under selective pressure, and variants that are under comparable selective pressure to predicted loss-of-function variants in protein-coding regions of known disease-associated genes.

Under this assumption we design a simulation study to model the association between the proportion of pathogenic uORFs and MAPS scores for all uORF-disrupting (UTC and stop-strengthening) variants using variants from protein-coding sequences. We calculate a distribution of MAPS scores for each simulated proportion of uORFs capable of harboring pathogenic loss-of-function variants as described in Figure 3.23.

Specifically the approach is to:

1. Randomly select ~4000 genes (approx. number of genes with uORFs) from genes in ClinVar with annotated pathogenic consequences
2. Partition these genes to a sub-fraction of "TRUE LOF" (10%, 20%, 30%, ...) and a sub-fraction of "FALSE LOF". The "TRUE LOF" genes will be contributing protein-coding LOF variants to the MAPS score in gnomAD
3. "FALSE LOF" genes will be contributing synonymous variants annotated in gnomAD
4. Calculate MAPS score for this mixed set of TRUE LOF and synonymous variants
5. Repeat 10,000 times to build confidence intervals for each proportion of "TRUE LOF" genes

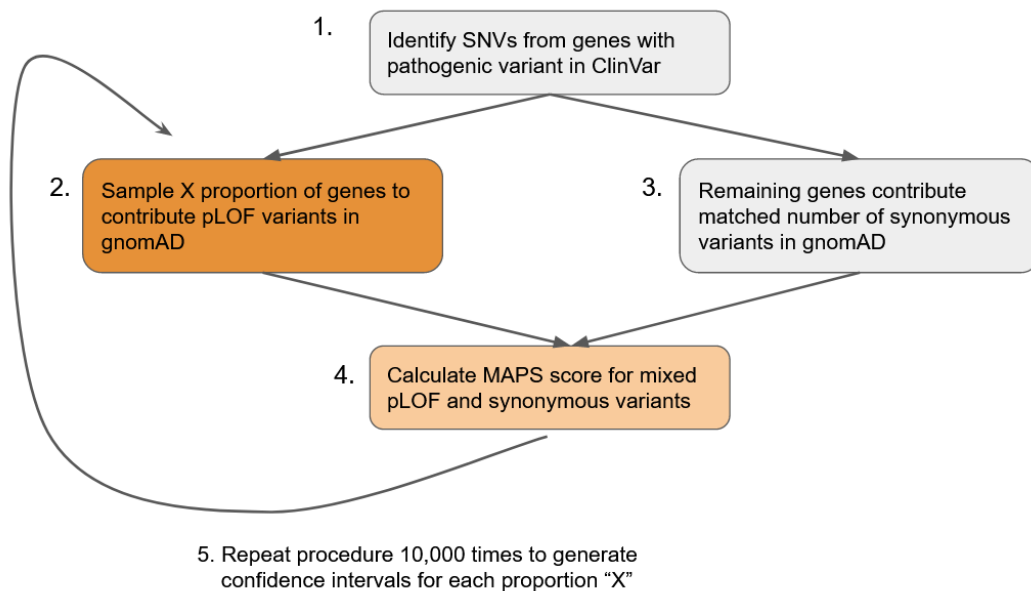


Figure 3.23: Sampling procedure to generate MAPS scores to model the proportion of uORFs where UTC or stop-strengthening variants are capable of having pathogenic consequences.

We can repeat this procedure for several fractions of "TRUE" pathogenic genes (10-80% in increments of 5%) and determine the range of possible MAPS scores given a particular proportion of genes contributing true pLOF variants. Since the baseline estimate for uORF variants is higher than that for synonymous coding variants, we adjust these MAPS scores by adding the baseline estimate for uORF variants to all simulated MAPS scores. We then plot the relationship between fraction of TRUE LOF genes and MAPS scores (Figure 3.24):

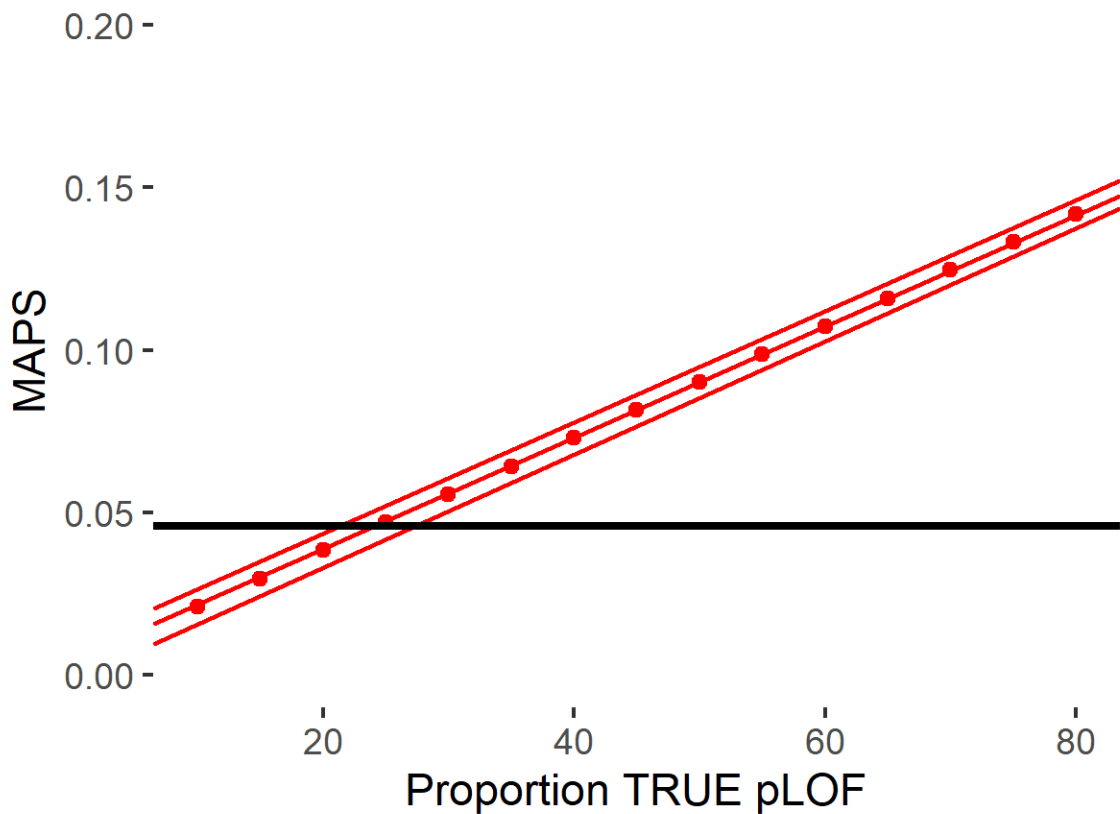


Figure 3.24: Relationship between the fraction of true LOF variants in ClinVar pathogenic genes and MAPS scores adjusted for uORF-synonymous variants baseline. Black line represents MAPS scores for UTC and stop-strengthening variants combined in all uORFs. Orange points represent the mean of MAPS scores over 10,000 bootstraps. Orange band represents 90% confidence interval of 10,000 bootstraps MAPS scores at each simulated proportion of genes contributing true pLOF variants.

We fit a simple regression model to these simulated MAPS scores as a function of the proportion of genes contributing LOF variants, and use the fitted model to determine the proportion of LOF features corresponding to the MAPS score for uORF-UTC and stop-strengthening variants. This corresponds to the proportion of uORFs contributing UTC or stop-strengthening variants with loss-of-function consequences capable of causing pathogenicity in humans. Solving this linear equation gives an estimate of the proportion of uORFs where UTC or stop-strengthening variants have comparable consequences as LOF protein coding variants in ClinVar pathogenic genes. This gives us the estimate of 24.15354%. We fit separate lines to the 5% and 95% bootstrapped

MAPS estimates at each proportion to determine the 90% confidence interval for this estimate (21.35285 - 27.33626).

To derive an estimate of the proportion of uORFs under similar constraint to maintain amino acid identity as protein-coding regions of the genome, we repeat the above procedure substituting missense variants in all protein coding genes for pLOF variants in ClinVar disease-associated genes. This procedure gives us an estimate of 9.916032% (5.216744 - 14.69604).

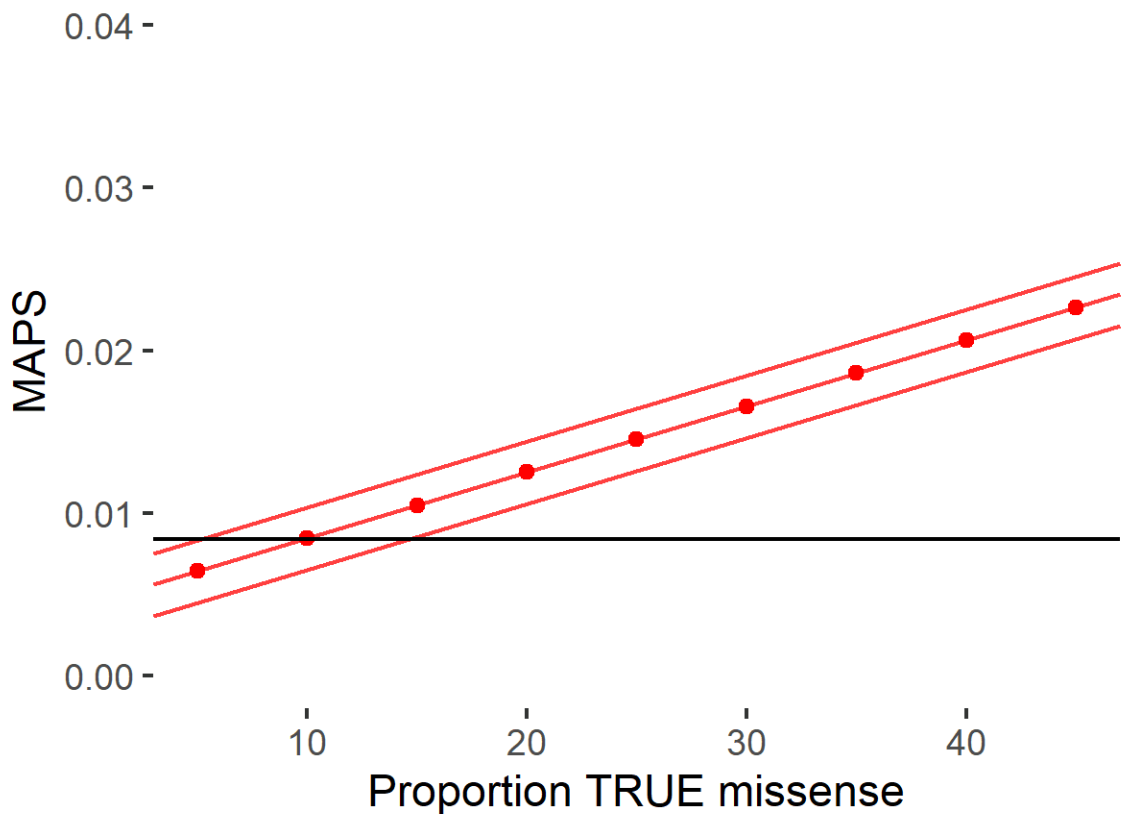


Figure 3.25: Relationship between the fraction of true missense variants in all protein-coding genes and MAPS scores adjusted for uORF-synonymous variants baseline. Black line represents MAPS scores for predicted missense variants for all uORFs. Red points represent the mean of MAPS scores over 10,000 bootstraps. Red lines above and below points represent a simple regression line fitted to the 90% confidence interval of 10,000 bootstrap MAPS scores at each point.

CHAPTER 4: TRANSLATING REGULATORY INSIGHTS INTO THERAPIES

4.1: G-quadruplexes and upstream open reading frames in UTRs

Much of the human genome is noncoding, but knowledge of how and when genetic variation can impact biology and disease has been primarily limited to coding sequences. Interpreting genetic variation in noncoding DNA remains challenging because the universe of functional noncoding elements in the human genome remains incompletely mapped and because the potential impact of variants affecting known noncoding regulatory elements is difficult to predict. Here, we have adopted an integrated omics approach towards addressing these dual challenges: first, by examining the genetic and genomic evidence that putative G-quadruplex forming sequences in UTRs of mRNAs are functional, and second by elucidating new patterns of functional variation affecting translated upstream open reading frames. For putative G-quadruplex forming sequences in UTRs, we have shown that central guanines in the canonical G-quadruplex motif are significantly depleted of genetic variation in concordance with their importance for facilitating stable secondary structure formation. We further find that these sequences are enriched for cis-eQTLs annotated in GTEx, and overrepresented among RNA-protein binding interactions mapped by ENCODE. Taken together, this evidence suggests that UTR G-quadruplexes are functional noncoding elements in UTRs.

Although these studies imply that G4 forming sequences in UTRs are constrained by selection, dissecting the precise functional importance of these sequences will be an important step to understand their roles more clearly in gene regulation. Our finding that both 5' and 3'UTR G4 sequences are enriched for cis-eQTLs implies that these sequences are broadly capable of influencing mRNA abundance. In particular, the enrichment for cis-eQTLs in 3'UTR protein suggests that G4s in 3'UTRs are involved in mediating mRNA stability [32], and that this is accomplished by facilitating RNA-protein binding interactions. More generally, the observed

enrichment for RNA protein binding interactions over both 5' and 3' UTR G4s implies that the capacity for these secondary structures to bind proteins is central to their functionality.

Whether some subsets of G4 binding proteins bind differentially to folded versus unfolded G4 secondary structures remains an interesting question for further investigation. Indeed, subsequent studies have replicated our findings that helicase RNA-binding proteins including DDX6 tend to bind G-rich sequences capable of forming G4 structures [179], suggesting that dynamic folding and unfolding UTR G4s plays a role in regulation. G4 folding could possibly mask other regulatory protein binding sites within mRNAs, or create a new substrate for protein binding, depending on cell state. Our finding that many UTR sequences with potential G4 structures tend to be overrepresented among genes involved in cellular stress response pathways further implies that differential G4 binding interactions may be involved in post-transcriptional responses to cellular stress. Future targeted functional studies should explore this possibility further.

We have additionally examined patterns of selective pressure in non-canonical open reading frames mapped through ribosome profiling experiments. Our analysis of allele frequencies in gnomAD identified uORF stop-introducing and stop-strengthening variants as categories of variation in the 5'UTR that exhibit a strong signature of negative selection, comparable to that of missense mutations in protein-coding regions of the genome. We further demonstrate that certain uORF stop-strengthening and stop-introducing variants associated with human disease phenotypes in two EHR-based biobanks can decrease downstream protein expression in reporter gene assays. Together, these data demonstrate that variants introducing new stop codons or strengthening existing stop codons in uORFs can impact protein expression and imply that they may contribute to human disease.

Previous studies of translational regulation by uORFs have revealed that several additional factors, including the strength of a uORF start codon [180], intercistronic distance between the uORF stop codon and downstream coding gene [181,182], and potential secondary structures in

the 5'UTR can impact translational regulation by uORFs [183,184]. Although these factors were not addressed in the current study, they are likely to influence the ability for uORF stop-introducing or stop-strengthening variants to decrease downstream protein expression. Notably, mutagenesis studies have reported that shortening repressive uORFs significantly may result in increased expression of downstream proteins [181]. This raises the possibility that some uORF stop-introducing variants can confer gain-of-function rather than loss-of-function depending on the length of the uORF they interrupt, the specific uORF that is affected, and the amount by which the uORF is shortened. In the scanning model of translation initiation, uORF-mediated translational repression occurs by preventing ribosomes from reacquiring the necessary translation initiation factors in time to begin translation at the downstream CDS. Thus, uORF stop-introducing variants may sufficiently lengthen the intercistronic distance between uORF and CDS to release downstream coding sequences from uORF-mediated translational repression. Further characterizing the precise relationship between intercistronic distance and uORF repressiveness will help improve our ability to accurately predict when uORF stop-introducing variants can increase protein expression.

The observed allele frequencies of variants affecting translated uORFs also suggests that these regulatory elements largely act at the level of translation rather than through encoding functional micropeptides. In support of this interpretation, we do not observe widespread evidence that amino acid encoding is constrained in most non-canonical ORFs, but we do observe that uORF translation initiation in 5'UTRs is maintained by selection. While there has been growing interest in the possibility functional micropeptides are widely encoded in ncORFs throughout the genome, the absence of significant selection to maintain amino acid encoding in ncORFs more broadly suggests that the functionality of micropeptides is either not associated with their amino acid encoding, or that perhaps that the fraction of functional micropeptides is much smaller than implied through ribosome profiling studies. Indeed, as translation in 5'UTRs has long been observed to have regulatory functions, it is possible that widespread translation in ncORFs may

also have regulatory rather than peptide-encoding roles in biology. Nevertheless, it remains possible that patterns of selection in ncORF encoded micropeptides do not reflect similar selective forces as those acting in canonical protein coding regions of the genome. In this scenario the function of ncORF micropeptides may alternatively depend on their length, or perhaps the presence of a few key amino acids in their sequence.

4.2: From regulation to therapy

Together, our studies help expand the understanding of regulatory elements in UTRs, and the interpretation of genetic variation in 5' and 3' UTRs – core components of all protein-coding messenger RNAs. As databases of human genetic variation grow larger, so too will the resolution by which we can observe and understand patterns of functional genetic variation in noncoding DNA. Further work establishing detailed maps of functional noncoding RNA regulatory elements will help inform the future development of new medicines which manipulate these regulatory elements for therapy. Indeed, drugging RNAs has received growing interest as a therapeutic mechanism in recent years [185]. The molecular tools of RNA manipulation – including small interfering RNAs (siRNAs) and antisense oligonucleotides (ASOs) – have existed for decades, however strategies for modulating protein activity or direct genome editing have traditionally received greater attention for therapeutic development. Compared to small molecules, oligonucleotides do not widely distribute throughout the body with the same efficiency, thus developing improved targeted drug delivery strategies is critical to realizing the promise of oligonucleotide therapies.

Nevertheless, interest in RNA targeting drugs has grown in recent years as advances in our understanding of RNA regulation and disease has improved. Early oligonucleotide-based therapies were developed for pharmacologically convenient tissues in the human body, beginning with the eye. Approved by the FDA over 20 years ago as a therapy for cytomegalovirus retinitis,

the first antisense oligonucleotide to find use in humans, Fomivirsen, acted through blocking translation of viral mRNAs and was delivered by intravitreal injection [187]. As advances in oligonucleotide chemistry conferred greater resistance to degradation and improved biodistribution, other routes of administration and mechanisms of action became feasible [186]. Subsequent RNA-based therapies have focused on modulating alternative splicing, or inducing RNA degradation relying on intravenous administration for muscle targeting or intrathecal delivery for targeting neurons [185]. Despite these advances, oligonucleotide delivery remains much less efficient than for small molecules, and devising improved methods for targeted delivery remains a key bottleneck to the translational application of oligonucleotides for therapy [188].

Informed by our growing appreciation for the diversity of RNA regulatory mechanisms, there is now increasing interest in new strategies for changing protein expression at the RNA level. Indeed, approaches for targeting G-quadruplex elements and uORFs to modulate protein expression have been proposed previously in the literature [189–191], and several preclinical studies have been published with promising results across a number of human disease contexts [192–196]. As catalogues of RNA regulatory elements encoded in the transcriptome expand, so too will the possibilities of therapeutic manipulation.

APPENDIX A: Methods for Integrative analysis reveals RNA G-quadruplexes in UTRs are selectively constrained and enriched for functional associations

Identification of UTR G-quadruplex Sequences

Annotated UTR sequences and genomic coordinates were downloaded using biomaRt[189,190] for Ensembl Transcript Database version 75 for all protein-coding transcripts. Putative G-quadruplex forming sequences (pG4) were identified using this set of annotated UTRs by performing a pattern-matching text search to identify regions of UTRs matching the canonical G-quadruplex pattern as a regular expression: $(G:3+)-\{N:1-7\}(3)-(G:3+)$. Genomic coordinates for pG4 sequences within UTRs were obtained using a custom python script, and cross-referenced with annotated protein-coding regions of the genome from the Ensembl Transcript Database to remove overlaps between annotated UTRs and coding sequences using human genome hg19 coordinates. The identified set of 5' and 3' pG4 sequences, and corresponding genomic coordinates were used for all downstream analysis. This approach yielded a total of 5235 protein-coding genes harboring a UTR pG4 sequence, with 2967 genes having a 5'UTR pG4, and 2835 genes having a 3'UTR pG4. Of the 5235 genes with either a 5' or 3' UTR pG4, 567 have both.

A second set of pG4 sequences with evidence of secondary structure formation was defined by overlapping pG4 motifs identified transcriptome-wide with published *in vitro* rG4-seq annotations by using the K+PDS conditions [63]. Only the subset of rG4-seq G4s matching the canonical pG4 motif were used in this analysis. For the set of rG4 sequences, we find 243 protein-coding genes encoding a 5'UTR rG4, and 803 genes encoding a 3'UTR rG4. Of these rG4 genes, 16 have 5' and 3' UTR pG4 encoding transcript isoforms.

Constraint Analysis

Variants from gnomAD release 2.2.1 were obtained from (URL here: <https://gnomad.broadinstitute.org/downloads>) and filtered to exclude those marked with segmental duplication, low complexity regions (LCR), and decoy flags, in addition to those variants whose True Positive probability as determined by a random forest model trained in gnomAD did not exceed 40% [44]. As an additional requirement, only those variants where the total observed allele number was at least 80% of the maximum number of sequenced alleles was considered to control for differences in sequencing depth in the gnomAD WGS dataset. The remaining set of high-confidence variants was overlapped with genomic coordinates for UTR pG4, non-pG4, and CDS regions, using bedtools2 (version 2.27.1) intersect with the -u and -b flags.

The transcript constraint-table from gnomAD release 2.2.1 (*URL* <https://gnomad.broadinstitute.org/downloads>) was used to randomly select a matching set of transcript-level constraint-matched non-pG4 UTR sequences based on the gnomAD observed / expected metric for the 5' UTR and 3' UTR separately. Specifically, transcript constraint was matched between pG4 and non-pG4 forming sequences using the observed versus expected ratio of loss of function variants metric (LOEUF) provided for each transcript by gnomAD.

The fraction of variants per sequenced allele across UTR regions were computed as the fraction of the observed allele count versus observed allele number. The distribution of frequencies for variants mapping to each UTR region was extracted from the gnomAD summary variant call files directly. P-values for difference between the expected number of variants per sequenced allele across genomic regions were calculated using a two-sided Fisher exact test. Only variants that did not overlap annotated coding regions of the transcriptome were compared to ensure that UTRs overlapping coding regions of other transcript isoforms were excluded. All statistical tests were conducted for 5'UTR and 3'UTR features separately.

For positional constraint analysis, we applied the mutability-adjusted proportion of singletons (MAPS) metric [37] for each nucleotide position across all trinucleotide G-tracts with G4-forming capacity, as defined by our bioinformatic analysis. We developed a MAPS model using custom code based on a previously published MAPS model (<https://github.com/pjshort/dddMAPS>) with the addition of adjusting for methylation levels at variant positions. We divide variants by median estimated methylation levels across 37 tissues at CpG sites into None/Low (<0.2), Intermediate (0.2-0.6), or High (0.6<) bins for which separate methylation-adjusted mutation rates were available. Our model was trained by regressing the observed proportion of singleton synonymous variants for each trinucleotide context within protein-coding regions of the genome on mutation rates for each trinucleotide context (methylation-adjustment was performed only at CpG dinucleotides) derived from intergenic noncoding regions of the genome [137]; [44]. All variants used in this analysis, including synonymous variants used for training the model, were subject to the same filtering requirements as used in the analysis of allele frequencies (random forest True Positive probability exceeding 40%, and total observed allele number was at least 80% of the maximum number of sequenced alleles). To control for ambiguity regarding which specific guanines within each G-tract are involved in pG4 formation for G-tracts having more than 3 guanines, we considered only variants within trinucleotide G-tracts. MAPS values were also determined for the set of variants with a VEP consequence of missense, or those variants predicted to cause a loss of function (pLoF) in gnomAD to provide context for the different degrees of purifying selection acting over a set of variants. pLoF variants were defined as those annotated with Ensembl predictions for having a high impact and includes transcript_ablation, splice_acceptor_variant, splice_donor_variant, stop_gained, frameshift_variant, stop_lost, and start_lost terms. In our assessment of positional constraint within the meta-pG4 sequence consisting of only trinucleotide G-tracts, we calculated MAPS for four categories of pG4 variants: 1) all genes, 2) genes with at least one transcript falling in the upper one-third of transcripts that are most intolerant to loss of function mutations (as determined by the gnomAD o/e metric), 3)

alternative pG4 genes, and 4) alternative pG4 disease-associated genes extracted from ClinVar database (April, 2019 release). Permutation P-values were obtained by performing 10,000 bootstraps for each set of pG4 variants in gnomAD with replacement to produce a distribution of MAPS score for each variant context; and then comparing these distributions to a matched set of resampled MAPS scores using either all UTR variants, or position-matched non-pG4 GGG/CCC trinucleotide variants to determine the proportion of bootstrapped samples whose MAPS score of pG4 regions exceeded the matched non-pG4 variant set.

Posterior substitution probabilities for noncoding regions of the genome based on local heptameric sequence contexts were obtained from a published model [75] based on the Phase 1 release of the 1000 Genomes Project. Cumulative substitution probabilities for each of the possible mutations within a heptamer context (e.g. A→C, A→T, A→G) were calculated by summing over all nucleotide substitution probabilities for a given heptamer context. To produce a null distribution of observed / expected number of substitutions for non-pG4 regions of the UTR, we randomly sampled 5000, 25 nucleotide UTR regions from constraint-matched transcripts, 10,000 times to generate a null distribution. Specifically, constraint-matched non-pG4 transcripts were divided into heptamers using a sliding window across the entire region, and substitution probabilities based on heptameric context alone was summed for each nucleotide position of each region to estimate the expected substitution frequency across each region of interest. The number of expected substitutions as derived from the heptamer substitution model for a given region was compared to observed substitutions for the European subpopulation within the Phase 1 of the 1000 Genomes Release. We performed comparisons across UTR pG4 G-tracts, pG4 non-G-tract Gap sequences, and constraint-matched non-pG4 UTRs for all pG4s, and the subset of rG4-seq supported pG4 motifs. Because the model does not adjust for methylation at CpG dinucleotides, all CpG dinucleotide positions within UTRs were removed from consideration. Statistical significance was determined by randomly sampling a set of genomic positions from each region of interest with replacement, matching the original combined size of each region,

over 10,000 iterations to produce a distribution of observed / expected ratios for each region of interest. P-values for each region (pG4, rG4, gap sequences) were calculated as the proportion of the observed / expected ratios obtained from the above bootstrapping procedure that were less than a matched set of observed / expected ratios using all 5' and 3'UTR genomic positions obtained by the same procedure.

pG4 Isoform Expression Across Tissues in GTEx

The median expression of each annotated RNA transcript (as measured in units of TPMs) in each tissue context was downloaded from GTEx v7. Median TPMs for each transcript were extracted for all pG4- or non-pG4 containing transcript for each pG4 gene, the highest expressed pG4 or non-pG4 isoform was selected, and a threshold of 1 TPM was used to determine expression within a specific tissue context. pG4 transcripts were deemed constitutive if only one of the pG4, or none of the non-pG4 transcripts exceeded this threshold, and labeled alternative if both the pG4, and non-pG4 transcripts exceeded this threshold. Significance of the distribution of alternative versus constitutive UTR pG4-encoding genes was assessed by randomly assigning pG4 and non-pG4 transcripts each gene, maintaining the number of transcript isoforms encoded by each gene constant with the condition that each gene should contain at least 1 pG4-encoding transcript. The distribution of the ratio of alternative to constitutive pG4 genes from the randomly distributed pG4 transcripts was then computed over 10,000 iterations to obtain a P-value for the true ratio of alternative to constitutive UTR pG4 encoding genes.

cis-eQTL and protein-binding enrichment

Significant variant-gene pairs were obtained from GTEx release version 7 (URL: <https://gtexportal.org/home/datasets>) constituting the set of nominally significant cis-eQTLs. Lead

cis-eQTL variants for each gene were defined as the variant with the lowest P-value for each gene, from the set of all significant variants in each tissue context separately. The set of lead and nominally significant variants was overlapped with UTR pG4 and non-pG4 regions of the UTR, and the number of significant cis-eQTL variants per region was compared to the number of non-significant tested SNPs occupying the same region to determine the proportion of cis-eQTLs compared to non-cis-eQTL SNPs. UTRs with cis-eQTLs not associated with changing the expression of the parent gene were excluded this analysis. Enrichment of cis-eQTLs was computed using a two-sided Fisher Exact Test. The set of causal eQTL candidates were obtained directly from the supplemental material of Brown et. al [80], and enrichment statistics were computed using GTEx v6p tested SNPs instead of v7 to match the data used in that study.

The direction bias of nominally significant cis-eQTLs within UTR pG4 G-tracts versus non-G-tract variants was computed by binarizing the normalized effect size pre-computed for each QTL by GTEx, and comparing the proportions of QTLs in each feature with either a positive effect, or negative effect on gene expression for each possible cis-eQTL annotation across all tissue contexts combined. Statistical significance was determined by a two-sided Fisher Exact Test.

High-confidence protein-binding sites were obtained from ENCODE CLIP-seq summaries and only peaks called with an Irreproducible Discovery Rate = 1000 were used for downstream enrichment analyses as determined by ENCODE [82]. Overlapping binding sites for multiple proteins were collapsed into a single protein-binding site, and the density of unique binding sites overlapping UTR pG4 regions compared to non-pG4 regions of the UTR was compared by dividing the number of CLIP-seq peaks overlapping each feature by the total number of nucleotides in each region. Significance was assessed using a chi-square test with 2 degrees of freedom.

Proteins whose binding sites are enriched for pG4 overlaps were computed using a hypergeometric test, by comparing the proportion of set of pG4 containing versus non-pG4 binding sites for a given RBP compared against the background proportion of of all UTR CLIP-seq peaks containing a pG4 sequence. The significance of pairwise overlaps between protein-gene targets was also computed using a hypergeometric test to assess the degree that one protein's pG4 binding genes were also targets for another protein.

Gene Expression with RBP Knockdown in ENCODE

Processed differential gene expression tables for K562 and HepG2 were obtained directly from ENCODE (<https://www.encodeproject.org/>) for each of the pG4-enriched binding proteins and their respective knockdown experiments. For each experiment, a gene was considered differentially expressed at an FDR threshold of < 0.05 . Genes from ENCODE differential expression tables were annotated as either a pG4 gene or non-pG4 gene on the basis of whether they encoded for a transcript isoform possessing a UTR pG4 sequence in either the 5'UTR or 3'UTR. The odds ratio for being significantly differentially expressed was calculated by comparing the ratio of pG4 to non-pG4 genes reaching statistical significance for differential expression between shRNA knockdown of the RBP, and the control for each protein separately. Statistical significance was determined by Fisher's Exact Test and FDR was controlled at 0.001 by applying the Benjimini-Hochberg procedure to the resultant P-values for each cell line. The direction of effect on pG4 gene expression for protein knockdown to cause an increase or decrease in pG4 gene expression was determined by taking the median value for log₂-fold change in expression for all pG4-containing genes measured in a given experiment.

Variants in ClinVar and the NIH-GWAS Catalogue

The April 2019 release of ClinVar was obtained from <ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/>. Using these variant annotations, we identified a subset of disease associated genes as any gene with at least one variant having a Pathogenic or Likely Pathogenic annotation. These genes were used to subset the ClinVar database and all variants spanning 40 nucleotides or less were overlapped with UTR regions to assess for enrichment in pG4 sequences. Insertions or deletions spanning greater than 40 nucleotides were not considered in this analysis, nor were any variants with an annotation of Benign or likely benign in Clinvar. The number of variants across each region was then divided by the total number of bases in each respective region to estimate of the density of variation in a given region. The odds ratios for the number of single nucleotide variants compared to the number of bases in a given region were then compared using a two-sided Fisher Exact Test.

For identification of GWAS-implicated SNPs affecting annotated UTR pG4 sequences, publicly available phenotype-associated SNPs from were obtained from the NIH-EBI GWAS Catalogue. Genomic coordinates for GWAS SNPs were converted from hg38 to hg19 coordinates using the NCBI Genome Remapping Service (<https://www.ncbi.nlm.nih.gov/genome/tools/remap>). This set of lead GWAS SNPs was used to identify nearby linked SNPs in high LD using the Linkage Disequilibrium Calculator tool from the Ensembl GRCh37 website using a 50KB window surrounding each lead GWAS SNP and selecting the set of SNPs with $r^2 > 0.85$ using the GBR population of the 1000 Genomes Project.

Allele-specific expression for GWAS-variants

RNA-seq libraries were trimmed using TrimGalore [191]. Reads were aligned to the GRCh37 human genome using STAR (version 2.7.0c) with the WASP-filtering option, and matched whole-genome sequencing variant files obtained from GTEx for Skeletal Muscle, Thyroid, Fibroblast,

Esophagus, and Tibial Nerve tissue samples. Reads that did not pass WASP-filtering were removed from the resulting aligned bam files. PCR duplicates were removed using the python script `remove_duplicates.py` included in the WASP version 0.3.3 pipeline (<https://github.com/bmvdgeijn/WASP>). Read counts matching the reference and alternate alleles in the resultant WASP-filtered bam files were compiled using `bcftools mpileup` across UTR pG4 variants. A beta-binomial model was fitted using the *R* *VGAM* package [192] for each variant across all heterozygous samples identified using matched whole-genome sequencing from GTEx to estimate the ratio of reference reads to alternate reads. Estimates of statistical significance were obtained by using a likelihood ratio test comparing the log-likelihood of the observed count distribution for each variant using the beta-binomial estimate for ρ versus the null hypothesis of no bias ($\rho = 0.5$).

Code Availability

All analysis scripts used to generate the primary results and figures reported in this study are publicly available from: <https://bitbucket.org/biociphers/g4-paper-2019/src/master/>.

Data Availability

Pre-processed data, and instructions for how to access public data resources used in this study that can be used to regenerate the primary figures of this analysis have been uploaded to <https://bitbucket.org/biociphers/g4-paper-2019/src/master/>. A subset of the processed publicly available data underlying Figs 3, 4, and 5 are included in this repository, with associated instructions on how to access other data as required to regenerate these figures where necessary. This repository also contains a link to a Source Data file which contains raw data underlying Figs 1b-d, 3a, b, 4a-d, 5a, c, d, and Supplementary Figs 5 and 7. Genetic variation

data from The Genome Aggregation Database version 2 release are available from: <https://gnomad.broadinstitute.org/downloads>. Gene expression and cis-eQTL mapping data from the Genotype Tissue Expression Project version 7 release are available from the GTEx Portal website: <https://gtexportal.org/home/>. RNA-seq data used for allelic imbalance analysis are available from dbGaP (phs000424.v7.p2 [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v7.p2]). RNA-protein binding interaction data can be retrieved from the ENCODE Consortium: <https://www.encodeproject.org/>. GWAS-associated variation data can be accessed from the NIH-EBI GWAS Catalog: <https://www.ebi.ac.uk/gwas/>. A copy of the filtered ClinVar database used to generate Fig. 5a is included in the code repository from April 2019. The most updated version of disease-associated genetic variant annotations are also available from ClinVar: <https://www.ncbi.nlm.nih.gov/clinvar/>.

Statistics

Data were analyzed and statistics performed using *R* (version 3.5.0) and *Python* (version 3.7 and 3.6.1).

APPENDIX B: Methods for Disrupting upstream translation in mRNAs is associated with loss-of-function in human disease

Annotation of translated non-canonical open reading frames

Non-canonical ORF (ncORF) annotations encompassing 5'UTR ORFs (uORFs), 3'UTR ORFs (dORFs), long-noncoding RNA ORFs (lncRNA) and pseudogene ORFs were retrieved from Supplementary File 1 from Ji et al. [121]. These ncORFs were mapped by ribosome-profiling in human BJ fibroblasts and MCF10A breast epithelial cells using the RibORF algorithm. Using the final set of genomic coordinates for ncORFs identified in this study, we converted these coordinates to match hg38 annotations using the UCSC LiftOver executable (<https://genome.ucsc.edu/cgi-bin/hgLiftOver>). Out of 10,007 distinct non-canonical uORFs mapped in the original study, 27 whose length changed after conversion (N = 5 uORFs, 4 dORFs, 16 lncRNA ORFs, 2 pseudogenes) were excluded from subsequent analyses. Each Refseq mRNA ID for each ORF-associated RNA transcript was annotated to its associated Ensembl transcript ID using the BioMart database v86 annotations. The first three nucleotides of each ORF were used as start codons for downstream analyses. The final three nucleotides of each ORF were used as stop codons for downstream analyses. 5' and 3' UTR definitions used in this study are derived from the Ensembl v86 annotations.

Quality filtering and annotation of variants from gnomAD version 3

Variants from gnomAD 3 release were downloaded from the gnomAD browser website (<https://gnomad.broadinstitute.org/downloads>). A set of high-confidence variants were obtained by removing those failing the Filter column (Filter != PASS) from the gnomAD version 3 vcf files using bcftools (version 1.9), and those falling in low complexity regions (lcr != 1). This set of variants was used for all downstream analyses. We additionally removed variants where the total

observed allele number was at less than 80% of the maximum number of sequenced alleles to control for differences in sequencing depth in the gnomAD WGS dataset. The remaining set of high-confidence variants was overlapped with genomic coordinates for annotated ncORFs, 5'UTR sequences, and annotated protein-coding sequences using bedtools (version 2.27.1) intersect with the -u and -b flags. The predicted consequence of each variant was obtained using the Ensembl Variant Effect Predictor (VEP, version 98.2) based on hg38 gene models obtained from Ensembl. VEP consequences were further filtered to only include the predicted consequence for the canonical Ensembl transcript as determined in [136].

Positional constraint analysis using variants from gnomAD

For the positional constraint analysis we applied the MAPS metric to each variant set. We developed a MAPS model following previous methods [136]. The set of synonymous protein-coding variants are used as a baseline measurement for neutral selection, and the proportion of singletons in a variant class are adjusted for differences in mutation rates due to local sequence context [37,136]. We trained our model by regressing the observed proportion of singleton-synonymous variants for each trinucleotide context within protein-coding regions of the genome using previously published context-dependent mutation rates derived from intergenic noncoding regions of the genome [136]. Since negative selection prevents deleterious mutations from becoming common in human populations, more deleterious mutations - including those disrupting essential splice sites or introducing premature termination codons - are also more enriched for singletons compared to neutral variants.

MAPS scores for a given set of variants are calculated as described previously [37,135,136].

Briefly, for a given set of variants, we use the MAPS model to determine the expected number of singletons that should be observed, based on the transformed mutation rates which account for

trinucleotide context and methylation levels. To calculate the MAPS score, we take the observed number of singletons for this set of variants, and subtract the expected number of singletons calculated using the MAPS model. We then divide this value by the number of variants total to obtain the proportion of singleton variants adjusted for mutation context.

To estimate of MAPS scores for missense-causing mutations in canonical protein-coding sequences within the genome, we selected the subset of SNVs in gnomAD with an annotated VEP consequence of missense, and removed SNVs from this set of variants if they had additional VEP annotations that could be considered predicted loss-of-function (pLoF). The set of variants used to calculate MAPS scores for pLoF variants relied on aggregating variants with a VEP annotation of transcript_ablation, splice_acceptor_variant, splice_donor_variant, stop_gained, frameshift_variant, stop_lost, and start_lost terms. The set of synonymous variants used to train the MAPS model was filtered to remove variants with any of the previous predicted high impact annotations, and those with a possible missense consequence.

We computed MAPS scores for each set of variants based on uORF annotations, or 5'UTR annotations from Gencode (GRCh38.p13; https://www.gencodegenes.org/human/release_32.html). Using the set of filtered variants we matched them to uORF positions annotated by their relative position within the uORF reading frame, strand, and codon. We determined how the mutation affected the codon within the translated uORF sequence, and annotated each variant with its consequence on the encoded amino acid. We used these annotations to select variants that could introduce new stop codons (UTC-introducing variants) and those that strengthened existing stop codons within uORFs. For UAA-introducing variants we selected any variant that produced an in-frame UAA stop codon. For each set of stop-introducing or stop-strengthening variants, we selected a set of uORF variants matching the underlying trinucleotide context of each experimental set of variants. MAPS scores

for these variant sets were computed and confidence intervals were determined by resampling from each variant set with replacement over 10,000 iterations.

For codon optimality analysis, we used the set of codon stability coefficients (CSC) scores derived from SLAM-seq in 562 cells obtained from <https://doi.org/10.7554/eLife.45396.006> [153]. Optimality decreasing variants were defined as any variant which decreased the CSC score for the encoded codon, and optimality increasing variants were defined as any variant which increased the CSC score for the encoded codon.

Confidence intervals for MAPS scores were calculated using bootstrapping as described [135]. For each set of n variants used to compute a MAPS score, we select n variants randomly with replacement and recalculate MAPS scores. This is repeated over 10,000 permutations and the 5th and 95th percentiles of the MAPS scores distribution are used as confidence intervals. P-values for differences in MAPS scores were determined by calculating the proportion of bootstrapped MAPS scores from an experimental group of variants that were larger than those from the control group [135].

Determining the distribution of stop codons used by upstream open reading frames

Stop codons from each uORF were extracted based on genomic coordinates and the uORF reading frame. Confidence intervals were determined by sampling with replacement from the set of uORF stop codons over 10,000 iterations. For 5'UTR sequences, all stop-codon matching trinucleotides (UGA, UAG, UAA) were extracted from annotated canonical 5'UTR sequences of protein-coding genes in the BioMart Ensembl database (version 86). The set of canonical transcripts annotated in the gnomAD flagship release paper were used to define 5'UTR sequences for this analysis [136]. For each iteration, one stop codon was randomly selected from each 5'UTR and the proportion of UGA, UAG, and UAA trinucleotides selected from all 5'UTR

sequences were calculated. This procedure was repeated 10,000 times to form a distribution of TGA, TAG, and UAA trinucleotides in all 5'UTR sequences. This procedure was also repeated for uORF-matched UTR sequence segments that did not overlap known translated uORFs. P-values for the depletion of UAA stop codons used in translated uORFs were calculated by determining the number of bootstrap iterations where the frequency of UAA codons from uORFs was higher compared to non-uORF sequences. P-values for enrichment of UGA and UAG sequences were calculated by determining the fraction of sampled iterations where fewer UGA and UAG sequences were selected from uORF stop codons compared to all 5'UTRs and uORF-matched 5'UTR sequences respectively.

Assessing variant conservation using genome-wide phyloP scores

PhyloP scores for each base were downloaded from the UCSC genome browser (<http://hgdownload.cse.ucsc.edu/goldenpath/hg38/phyloP100way/>). 1-indexed bigwig files were converted to bed file format using the wig2bed tool from bedops (version 2.4.36; <https://bedops.readthedocs.io/en/latest/index.html>). These base-level annotations were matched to each uORF base and used to determine the proportion of bases that were significantly conserved (proportion of bases with phyloP score > 2). Possible inframe stop-codon creating positions were identified based on mapped reading frames for each uORF. These sites were extracted and further categorized by whether or not a mutation could create a UGA, UAG, or UAA stop codon. Some positions could be mutated to either a UGA, UAG or UAA codon and these were considered separately from potential UGA, UAG or UAA-creating positions. We have included all potential stop-introducing positions in Suppl. File 1. As a control we used phyloP scores for genomic positions with the potential to create non-uORF UGA, UAG, or UAA trinucleotides by mutation, but matched by distance to CDS in 10-base pair windows.

Start-disrupting genomic positions were annotated as those mutating the second or third position in the first codon of each translated uORF. Conservation based on phyloP scores were assessed for start-disrupting positions similar to potential stop-introducing positions. As a control we compared phyloP scores for uORF start-disrupting positions to out-of-frame start-disrupting positions within annotated uORFs, and a set of NTG start-disrupting variants that were not part of translated uORFs but matched by distance to the CDS as determined by 10-bp windows.

P-values were determined by sampling with replacement from each set of variants 10,000 times and re-calculating the proportion of significantly conserved bases (phyloP score > 2). The distribution of the fraction of conserved base positions were then compared against different sets of variants, and the P-value was defined as the fraction of samples where one group was higher than the other.

Setting and study participants

All individuals who were recruited for the Penn Medicine Biobank (PMBB) are patients of clinical practice sites of the University of Pennsylvania Health System. Appropriate consent was obtained from each participant regarding storage of biological specimens, genetic sequencing, access to all available electronic health record (EHR) data, and permission to recontact for future studies. The study was approved by the Institutional Review Board of the University of Pennsylvania and complied with the principles set out in the Declaration of Helsinki. Replication analyses were conducted using the whole exome sequencing (WES) dataset from the UK Biobank (UKB).

Genetic sequencing

This PMBB study dataset included a subset of 11,451 individuals in the PMBB who have undergone WES. For each individual, we extracted DNA from stored buffy coats and then obtained exome sequences generated by the Regeneron Genetics Center (Tarrytown, NY). These sequences were mapped to GRCh37 as previously described [155]. Furthermore, for subsequent phenotypic analyses, we removed samples with low exome sequencing coverage (*i.e.* less than 75% of targeted bases achieving 20x coverage), high missingness (*i.e.* greater than 5% of targeted bases), high heterozygosity, dissimilar reported and genetically determined sex, genetic evidence of sample duplication, and cryptic relatedness (*i.e.* closer than 3rd degree relatives), leading to a total of 10,900 individuals.

For replication studies in UKB, we interrogated the 32,268 individuals of European ancestry (based on UKB's reported genetic ancestry grouping) with ICD-10 diagnosis codes available among the 49,960 individuals who had WES data as generated by the Functional Equivalence (FE) pipeline. We focused our replication efforts on 32,268 individuals after removing samples with poor genotype quality, individuals closer than 3rd degree relatives, and those with dissimilar reported and genetically determined sex. The PLINK files for exome sequencing provided by UKB were based on mappings to GRCh38. Access to the UK Biobank for this project was from Application 32133.

Variant annotation and selection for association testing

For both PMBB and UKB, genetic variants were annotated using ANNOVAR [193] as 5' untranslated region (5' UTR), predicted loss-of-function (pLOF), or missense variants according to the NCBI Reference Sequence (RefSeq) database [193,194]. Rare (MAF \leq 0.1%) pLOF variants were defined as frameshift insertions/deletions, gain/loss of stop codon, or disruption of canonical splice site dinucleotides. Predicted deleterious rare (MAF \leq 0.1%) missense variants

were defined as those with Rare Exonic Variant Ensemble Learner (REVEL) [195] scores ≥ 0.5 . pLOF and REVEL-informed missense variants were selected for gene burden testing to validate the robustness of significant uORF variants' corresponding gene-disease associations.

Clinical data collection

International Classification of Diseases Ninth Revision (ICD-9) and Tenth Revision (ICD-10) disease diagnosis codes and procedural billing codes, medications, and clinical imaging and laboratory measurements were extracted from the patients' EHR for PMBB. ICD-10 encounter diagnoses were mapped to ICD-9 via the Center for Medicare and Medicaid Services 2017 General Equivalency Mappings (<https://www.cms.gov/Medicare/Coding/ICD10/2017-ICD-10-CM-and-GEMs.html>) and manual curation. Phenotypes for each individual were then determined by mapping ICD-9 codes to distinct disease entities (*i.e.* Phecodes) via Phecode Map 1.2 using the R package "PheWAS" [156,196]. Patients were determined to have a certain disease phenotype if they had the corresponding ICD diagnosis on two or more dates, while phenotypic controls consisted of individuals who never had the ICD code. Individuals with an ICD diagnosis on only one date as well as individuals under control exclusion criteria based on PheWAS phenotype mapping protocols were not considered in statistical analyses.

For UKB, we used the provided ICD-10 disease diagnosis codes for replication studies, and individuals were determined to have a certain disease phenotype if they had one or more encounters for the corresponding ICD diagnosis given the lack of individuals with more than two encounters per diagnosis, while phenotypic controls consisted of individuals who never had the ICD code. Individuals under control exclusion criteria based on PheWAS phenotype mapping protocols were not considered in statistical analyses.

Association studies

A phenome-wide association study (PheWAS) approach was used to determine the phenotypes associated with 5' UTR variants predicted to create new UAA UTCs, or strengthen existing uORF stop sites and carried by individuals in PMBB for the discovery experiment [156]. Each disease phenotype was tested for association with each uORF variant using a logistic regression model adjusted for age, age², sex, and the first ten principal components (PCs) of genetic ancestry. We used an additive genetic model to collapse variants per gene via an extension of the fixed threshold approach [197]. Given the high percentage of individuals of African ancestry present in the discovery PMBB cohort, association analyses were performed separately in European (N=8198) and African (N=2172) genetic ancestries and combined with inverse variance weighted meta-analysis. Only 5' UTR variants with at least five total alternate alleles in PMBB were selected for univariate PheWAS analyses in the discovery phase while variants with greater than half of the genotypes annotated as missing due to low quality were excluded. This resulted in a final set of N=10 variants. Our association analyses considered only disease phenotypes with at least 20 cases, leading to the interrogation of 800 total Phecodes. All association analyses were completed using R version 3.3.1 (Vienna, Austria).

We evaluated the robustness of significant uORF-phenotype associations in the same PMBB discovery cohort by aggregating pLOF and predicted deleterious missense variants in each uORF's corresponding gene into a 'gene burden' for hypothesis-driven association with the significant phenotype from discovery. Only gene burdens with at least five total alternate alleles in PMBB were selected for replication studies. All gene burden association studies in PMBB were based on a logistic regression model adjusted for age, age², sex, and the first 10 PCs of genetic ancestry.

Additionally, we replicated our findings in UKB for significant uORF associations in the PMBB discovery using 1) hypothesis-driven univariate association studies for the same uORF variants

and 2) hypothesis-driven gene burden collapsing pLOF and predicted missense variants for the corresponding genes. Only uORF variants and gene burdens with at least five total alternate alleles in PMBB were selected for replication studies. Association statistics were calculated similarly to PMBB, such that each disease phenotype was tested for association with each gene burden or single variant using a logistic regression model adjusted for age, age², sex, and the first 10 PCs of genetic ancestry. Replication significance was defined using a P-value threshold of 0.05. All association analyses for PMBB and UK Biobank completed using R version 3.6.1.

Construction of expression vectors

The test plasmids used a modified pGL4.12[luc2CP] (Promega) vector backbone where the control of expression of the Firefly ORF was modified by the addition of an upstream CMV promoter. The modified pGL4.12 vector was linearized using Bgl-II and MreI restriction sites. Hybrid 5'UTR fragments containing the entire 5'UTR sequence and the first 91 nucleotides of the Luc2 Firefly ORF were produced by gBlock synthesis and received from Integrated DNA Technologies using sequences in Suppl. Table 3. Test plasmids were constructed by sub-cloning these hybrid 5'UTR sequences for PMVK, VPS53, and BCL2L13 into the modified pGL4.12 vector to preserve the uORF-CDS relationship for each construct. Correct fragment insertion was verified for each engineered construct by sanger sequencing. For PMVK and BCL2L13, the entire annotated 5'UTR sequence was used. For VPS53, because of a G-rich sequence in the 5'UTR upstream of the uORF complicated synthesis of the gene's entire 5'UTR fragment, we removed the first 75 nucleotides of the annotated 5'UTR sequence. Construct assembly was accomplished using the NEB Hi-Fi assembly protocol following manufacturer's instructions.

Cell culture and transfections

HEK293T cells were used for conditional expression of reporter genes. For transient transfections, HEK293T cells were split 1 day before transfection and seeded in 24-well plates at a density of 100,000 cells per well. 2 ug of the test Firefly reporter plasmid was transfected into each well using Lipofectamine 3000 following the manufacturer's protocol using 1.5 uL of transfection reagent and 0.5 uL of the P3000 reagent for each well. As a control for transfection efficiency, 0.02 ug of the pRL-CMV Renilla Luciferase plasmid (Promega Accession No. [AF025843](#)) was co-transfected with firefly luciferase plasmids. Biological replicates were obtained by transfecting cells from separate passages on separate days using newly prepared reagents. All transfections were repeated using the HeLa cell line. Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% (v/v) fetal bovine serum and antibiotics was used for all cell culture.

Luminometry assays

Luminescence was measured using the Promega Dual-Luciferase Reporter Assay System (E1910) following the manufacturer's protocol. Cells were lysed by adding 100 uL of lysis buffer. 10 uL of each lysate was transferred to a black opaque 96-well plate. The ratio of Firefly to Renilla luminescence with a microplate reader by automatic injection of the Luciferase Assay Reagent II and Stop & Glo reagents. Biological replicates were obtained by transfecting cells from separate passages on separate days using newly prepared reagents. Luminescence measurements were compared within each set of transfections and statistical significance was determined using a one-sided T-test comparing the firefly to Renilla expression ratio of each test construct normalized to the wild-type construct.

Code Availability

All scripts used in this analysis except for those generating PheWAS results and plots can be accessed from <https://www.bitbucket.org/biociiphers/uorf-paper-2020/src>

Data Availability

Data	Description	URL
gnomAD variants (version 3)	The set of variants obtained from 71,702 whole genome sequences used for MAPS analysis	https://gnomad.broadinstitute.org/downloads
Mapped Non-canonical ORFs	5'UTR (uORF), 3'UTR (dORF), long-noncoding RNA, and pseudogene ORFs mapped by the RibORF algorithm from ribosome-profiling data	https://doi.org/10.7554/eLife.08890.023
CSC scores	Codon-stability coefficient scores as determined by several techniques	https://doi.org/10.7554/eLife.45396.006

Software Availability

Software	Version	URL
Python	3.7.3	https://www.python.org/downloads/release/python-373/
R	3.6.1	https://cran.r-project.org/bin/windows/base/old/3.6.1/
bedtools	2.27.1	https://github.com/arq5x/bedtools2/releases
bcftools	1.9	http://samtools.github.io/bcftools/bcftools.html
Variant Effect Predictor (Ensembl)	98.2	https://useast.ensembl.org/info/docs/tools/vep/index.html

BIBLIOGRAPHY

1. Mirsky AE, Ris H. The desoxyribonucleic acid content of animal cells and its evolutionary significance. *J Gen Physiol.* 1951;34: 451–462.
2. Consortium IHGS, International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature.* 2001. pp. 860–921. doi:10.1038/35057062
3. Imanishi T, Itoh T, Suzuki Y, O'Donovan C, Fukuchi S, Koyanagi KO, et al. Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.* 2004;2: e162.
4. Clark MD, Hennig S, Herwig R, Clifton SW, Marra MA, Lehrach H, et al. An oligonucleotide fingerprint normalized and expressed sequence tag characterized zebrafish cDNA library. *Genome Res.* 2001;11: 1594–1602.
5. Britten RJ, Davidson EH. *Gene Regulation for Higher Cells: A Theory.* Science. 1969. pp. 349–357. doi:10.1126/science.165.3891.349
6. International HapMap Consortium. The International HapMap Project. *Nature.* 2003;426: 789–796.
7. Ikegawa S. A short history of the genome-wide association study: where we were and where we are going. *Genomics Inform.* 2012;10: 220–225.
8. Edwards SL, Beesley J, French JD, Dunning AM. Beyond GWASs: illuminating the dark road from association to function. *Am J Hum Genet.* 2013;93: 779–797.
9. Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Järvelä I. Identification of a variant associated with adult-type hypolactasia. *Nat Genet.* 2002;30: 233–237.
10. Lewinsky RH, Jensen TGK, Møller J, Stensballe A, Olsen J, Troelsen JT. T-13910 DNA variant associated with lactase persistence interacts with Oct-1 and stimulates lactase promoter activity in vitro. *Hum Mol Genet.* 2005;14: 3945–3953.
11. Ward LD, Kellis M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol.* 2012;30: 1095–1106.
12. Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. *Cell.* 2013;152: 1237–1251.
13. Kleinjan DA, van Heyningen V. Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am J Hum Genet.* 2005;76: 8–32.

14. Zoghbi HY, Beaudet AL. Epigenetics and Human Disease. *Cold Spring Harb Perspect Biol.* 2016;8: a019497.
15. Orkin SH, Kazazian HH Jr. The mutation and polymorphism of the human beta-globin gene and its surrounding DNA. *Annu Rev Genet.* 1984;18: 131–171.
16. Van der Ploeg LH, Konings A, Oort M, Roos D, Bernini L, Flavell RA. gamma-beta-Thalassaemia studies showing that deletion of the gamma- and delta-genes influences beta-globin gene expression in man. *Nature.* 1980;283: 637–642.
17. Kioussis D, Vanin E, deLange T, Flavell RA, Grosveld FG. Beta-globin gene inactivation by DNA translocation in gamma beta-thalassaemia. *Nature.* 1983;306: 662–666.
18. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet.* 2013;45: 580–585.
19. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science.* 2020;369: 1318–1330.
20. Ye Y, Zhang Z, Liu Y, Diao L, Han L. A Multi-Omics Perspective of Quantitative Trait Loci in Precision Medicine. *Trends Genet.* 2020;36: 318–336.
21. Cheng J, Maier KC, Avsec Ž, Rus P, Gagneur J. Cis-regulatory elements explain most of the mRNA stability variation across genes in yeast. *RNA.* 2017. pp. 1648–1659. doi:10.1261/rna.062224.117
22. 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;19: 158–174.
23. Hellen CU, Sarnow P. Internal ribosome entry sites in eukaryotic mRNA molecules. *Genes Dev.* 2001;15: 1593–1612.
24. Kearse MG, Wilusz JE. Non-AUG translation: a new start for protein synthesis in eukaryotes. *Genes Dev.* 2017;31: 1717–1731.
25. Green KM, Linsalata AE, Todd PK. RAN translation-What makes it run? *Brain Res.* 2016;1647: 30–42.
26. Nguyen L, Cleary JD, Ranum LPW. Repeat-Associated Non-ATG Translation: Molecular Mechanisms and Contribution to Neurological Disease. *Annu Rev Neurosci.* 2019;42: 227–247.
27. Barrett LW, Fletcher S, Wilton SD. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cell Mol Life Sci.* 2012;69: 3613–3634.
28. Gruber AJ, Zavolan M. Alternative cleavage and polyadenylation in health and disease. *Nat Rev Genet.* 2019. doi:10.1038/s41576-019-0145-z

29. Moszyńska A, Gebert M, Collawn JF, Bartoszewski R. SNPs in microRNA target sites and their potential role in human disease. *Open Biol.* 2017;7. doi:10.1098/rsob.170019
30. Steri M, Laura Idda M, Whalen MB, Orrù V. Genetic variants in mRNA untranslated regions. *Wiley Interdisciplinary Reviews: RNA.* 2018. p. e1474. doi:10.1002/wrna.1474
31. Sheets MD, Ogg SC, Wickens MP. Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro. *Nucleic Acids Res.* 1990;18: 5799–5805.
32. Mayr C. What Are 3' UTRs Doing? *Cold Spring Harb Perspect Biol.* 2019;11. doi:10.1101/cshperspect.a034728
33. Kimura M, Ohta T. On Some Principles Governing Molecular Evolution. *Proceedings of the National Academy of Sciences.* 1974. pp. 2848–2852. doi:10.1073/pnas.71.7.2848
34. Kimura M. Evolutionary Rate at the Molecular Level. *Nature.* 1968. pp. 624–626. doi:10.1038/217624a0
35. Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. *Genet Res.* 1974;23: 23–35.
36. Hermisson J, Pennings PS. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics.* 2005;169: 2335–2352.
37. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature.* 2016;536: 285–291.
38. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* 2010;20: 110–121.
39. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, et al. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature.* 2011;478: 476–482.
40. Chen K, Rajewsky N. Natural selection on human microRNA binding sites inferred from SNP data. *Nat Genet.* 2006;38: 1452–1456.
41. Savisaar R, Hurst LD. Both Maintenance and Avoidance of RNA-Binding Protein Interactions Constrain Coding Sequence Evolution. *Mol Biol Evol.* 2017;34: 1110–1126.
42. Chen G, Qiu C, Zhang Q, Liu B, Cui Q. Genome-Wide Analysis of Human SNPs at Long Intergenic Noncoding RNAs. *Human Mutation.* 2013. pp. 338–344. doi:10.1002/humu.22239
43. Zhang H, Shi X, Huang T, Zhao X, Chen W, Gu N, et al. Dynamic landscape and evolution of m6A methylation in human. *Nucleic Acids Res.* 2020;48: 6251–6264.

44. 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;581: 434–443.
45. Zhang S, Samocha KE, Rivas MA, Karczewski KJ, Daly E, Schmandt B, et al. Base-specific mutational intolerance near splice sites clarifies the role of nonessential splice nucleotides. *Genome Res*. 2018;28: 968–974.
46. 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. *Nat Commun*. 2020;11: 2523.
47. Tuller T, Zur H. Multiple roles of the coding sequence 5' end in gene expression regulation. *Nucleic Acids Res*. 2015;43: 13–28.
48. Kim HH, Kuwano Y, Srikantan S, Lee EK, Martindale JL, Gorospe M. HuR recruits let-7/RISC to repress c-Myc expression. *Genes Dev*. 2009;23: 1743–1748.
49. Mayr C. Regulation by 3'-Untranslated Regions. *Annu Rev Genet*. 2017;51: 171–194.
50. Kedde M, van Kouwenhove M, Zwart W, Oude Vrielink JAF, Elkon R, Agami R. A Pumilio-induced RNA structure switch in p27-3' UTR controls miR-221 and miR-222 accessibility. *Nat Cell Biol*. 2010;12: 1014–1020.
51. Meijlink F, Curran T, Miller AD, Verma IM. Removal of a 67-base-pair sequence in the noncoding region of protooncogene fos converts it to a transforming gene. *Proc Natl Acad Sci U S A*. 1985;82: 4987–4991.
52. Tushev G, Glock C, Heumüller M, Biever A, Jovanovic M, Schuman EM. Alternative 3' UTRs Modify the Localization, Regulatory Potential, Stability, and Plasticity of mRNAs in Neuronal Compartments. *Neuron*. 2018;98: 495–511.e6.
53. Ribeiro DM, Prod'homme A, Teixeira A, Zanzoni A, Brun C. The role of 3'UTR-protein complexes in the regulation of protein multifunctionality and subcellular localization. *Nucleic Acids Res*. 2020;48: 6491–6502.
54. Xu L, Peng L, Gu T, Yu D, Yao Y-G. The 3'UTR of human MAVS mRNA contains multiple regulatory elements for the control of protein expression and subcellular localization. *Biochim Biophys Acta Gene Regul Mech*. 2019;1862: 47–57.
55. Spasic A, Assmann SM, Bevilacqua PC, Mathews DH. Modeling RNA secondary structure folding ensembles using SHAPE mapping data. *Nucleic Acids Res*. 2018;46: 314–323.
56. Tian S, Das R. RNA structure through multidimensional chemical mapping. *Q Rev Biophys*. 2016;49: e7.
57. Watters KE, Lucks JB. Mapping RNA Structure In Vitro with SHAPE Chemistry and Next-Generation Sequencing (SHAPE-Seq). *Methods Mol Biol*. 2016;1490: 135–162.

58. Itzkovitz S, Hodis E, Segal E. Overlapping codes within protein-coding sequences. *Genome Res.* 2010;20: 1582–1589.
59. Goering R, Hudish LI, Guzman BB, Raj N, Bassell GJ, Russ HA, et al. FMRP promotes RNA localization to neuronal projections through interactions between its RGG domain and G-quadruplex RNA sequences. *Elife.* 2020;9. doi:10.7554/eLife.52621
60. Shafer RH, Smirnov I. Biological aspects of DNA/RNA quadruplexes. *Biopolymers.* 2000;56: 209–227.
61. Arora A, Maiti S. Differential biophysical behavior of human telomeric RNA and DNA quadruplex. *J Phys Chem B.* 2009;113: 10515–10520.
62. Zaccaria F, Fonseca Guerra C. RNA versus DNA G-Quadruplex: The Origin of Increased Stability. *Chemistry.* 2018;24: 16315–16322.
63. Kwok CK, Marsico G, Sahakyan AB, Chambers VS, Balasubramanian S. rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome. *Nat Methods.* 2016;13: 841–844.
64. Yang SY, Lejault P, Chevrier S, Boidot R, Robertson AG, Wong JMY, et al. Transcriptome-wide identification of transient RNA G-quadruplexes in human cells. *Nat Commun.* 2018;9: 4730.
65. Guo JU, Bartel DP. RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. *Science.* 2016;353. doi:10.1126/science.aaf5371
66. Agarwala P, Pandey S, Mapa K, Maiti S. The G-Quadruplex Augments Translation in the 5' Untranslated Region of Transforming Growth Factor β 2. *Biochemistry.* 2013;52: 1528–1538.
67. Kumari S, Bugaut A, Huppert JL, Balasubramanian S. An RNA G-quadruplex in the 5' UTR of the NRAS proto-oncogene modulates translation. *Nat Chem Biol.* 2007;3: 218–221.
68. Huang H, Zhang J, Harvey SE, Hu X, Cheng C. RNA G-quadruplex secondary structure promotes alternative splicing via the RNA-binding protein hnRNPF. *Genes Dev.* 2017;31: 2296–2309.
69. Subramanian M, Rage F, Tabet R, Flatter E, Mandel J-L, Moine H. G-quadruplex RNA structure as a signal for neurite mRNA targeting. *EMBO Rep.* 2011;12: 697–704.
70. Rouleau S, Glouzon J-PS, Brumwell A, Bisailon M, Perreault J-P. 3' UTR G-quadruplexes regulate miRNA binding. *RNA.* 2017;23: 1172–1179.
71. Beaudoin J-D, Perreault J-P. Exploring mRNA 3'-UTR G-quadruplexes: evidence of roles in both alternative polyadenylation and mRNA shortening. *Nucleic Acids Res.* 2013;41: 5898–5911.

72. Huppert JL, Bugaut A, Kumari S, Balasubramanian S. G-quadruplexes: the beginning and end of UTRs. *Nucleic Acids Res.* 2008;36: 6260–6268.
73. Fay JC, Wyckoff GJ, Wu CI. Positive and negative selection on the human genome. *Genetics.* 2001;158: 1227–1234.
74. Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, Reymond A, et al. Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nature Genetics.* 2006. pp. 223–227. doi:10.1038/ng1710
75. Aggarwala V, Voight BF. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nat Genet.* 2016;48: 349–355.
76. Agarwala P, Kumar S, Pandey S, Maiti S. Human Telomeric RNA G-Quadruplex Response to Point Mutation in the G-Quartets. *J Phys Chem B.* 2015;119: 4617–4627.
77. Resch AM, Ogurtsov AY, Rogozin IB, Shabalina SA, Koonin EV. Evolution of alternative and constitutive regions of mammalian 5'UTRs. *BMC Genomics.* 2009;10: 162.
78. Mayr C. Evolution and Biological Roles of Alternative 3'UTRs. *Trends Cell Biol.* 2016;26: 227–237.
79. Mockenhaupt S, Makeyev EV. Non-coding functions of alternative pre-mRNA splicing in development. *Semin Cell Dev Biol.* 2015;47-48: 32–39.
80. Brown AA, Viñuela A, Delaneau O, Spector TD, Small KS, Dermitzakis ET. Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues. *Nat Genet.* 2017;49: 1747–1751.
81. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012;489: 57–74.
82. Li Q, Brown JB, Huang H, Bickel PJ. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics.* 2011. pp. 1752–1779. doi:10.1214/11-aas466
83. Jourdain AA, Koppen M, Wydro M, Rodley CD, Lightowlers RN, Chrzanowska-Lightowlers ZM, et al. GRSF1 regulates RNA processing in mitochondrial RNA granules. *Cell Metab.* 2013;17: 399–410.
84. Pietras Z, Wojcik MA, Borowski LS, Szewczyk M, Kulinski TM, Cysewski D, et al. Dedicated surveillance mechanism controls G-quadruplex forming non-coding RNAs in human mitochondria. *Nat Commun.* 2018;9: 2558.
85. Havrilla JM, Pedersen BS, Layer RM, Quinlan AR. A map of constrained coding regions in the human genome. *Nat Genet.* 2018. doi:10.1038/s41588-018-0294-6

86. Lord J, Gallone G, Short PJ, McRae JF, Ironfield H, Wynn EH, et al. Pathogenicity and selective constraint on variation near splice sites. *Genome Res.* 2019;29: 159–170.
87. di Iulio J, Bartha I, Wong EHM, Yu H-C, Lavrenko V, Yang D, et al. The human noncoding genome defined by genetic diversity. *Nat Genet.* 2018;50: 333–337.
88. 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 Res.* 2014;42: D980–5.
89. Wain LV, Vaez A, Jansen R, Joehanes R, van der Most PJ, Erzurumluoglu AM, et al. Novel Blood Pressure Locus and Gene Discovery Using Genome-Wide Association Study and Expression Data Sets From Blood and the Kidney. *Hypertension.* 2017. doi:10.1161/HYPERTENSIONAHA.117.09438
90. Siitonen A, Nalls MA, Hernández D, Gibbs JR, Ding J, Ylikotila P, et al. Genetics of early-onset Parkinson’s disease in Finland: exome sequencing and genome-wide association study. *Neurobiol Aging.* 2017;53: 195.e7–195.e10.
91. van de Geijn B, McVicker G, Gilad Y, Pritchard JK. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat Methods.* 2015;12: 1061–1063.
92. Gros J, Rosu F, Amrane S, De Cian A, Gabelica V, Lacroix L, et al. Guanines are a quartet’s best friend: impact of base substitutions on the kinetics and stability of tetramolecular quadruplexes. *Nucleic Acids Res.* 2007;35: 3064–3075.
93. Lee JY, Kim DS. Dramatic effect of single-base mutation on the conformational dynamics of human telomeric G-quadruplex. *Nucleic Acids Res.* 2009;37: 3625–3634.
94. Garcia-Moreno M, Noerenberg M, Ni S, Järvelin AI, González-Almela E, Lenz CE, et al. System-wide Profiling of RNA-Binding Proteins Uncovers Key Regulators of Virus Infection. *Mol Cell.* 2019;74: 196–211.e11.
95. Li Z, Nagy PD. Diverse roles of host RNA binding proteins in RNA virus replication. *RNA Biol.* 2011;8: 305–315.
96. Lavezzo E, Berselli M, Frasson I, Perrone R, Palù G, Brazzale AR, et al. G-quadruplex forming sequences in the genome of all known human viruses: A comprehensive guide. *PLoS Comput Biol.* 2018;14: e1006675.
97. Kikin O, D’Antonio L, Bagga PS. QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res.* 2006;34: W676–82.
98. Garant J-M, Perreault J-P, Scott MS. Motif independent identification of potential RNA G-quadruplexes by G4RNA screener. *Bioinformatics.* 2017;33: 3532–3537.
99. Bedrat A, Lacroix L, Mergny J-L. Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res.* 2016;44: 1746–1759.

100. Fickett JW. Finding genes by computer: the state of the art. *Trends in Genetics*. 1996. pp. 316–320. doi:10.1016/0168-9525(96)10038-x
101. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*. 2009;324: 218–223.
102. 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;13: 165–170.
103. Ingolia NT, Lareau LF, Weissman JS. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*. 2011;147: 789–802.
104. Ji Z, Song R, Regev A, Struhl K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife*. 2015. doi:10.7554/elife.08890
105. Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJS, Jackson SE, et al. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep*. 2014;8: 1365–1379.
106. Martinez TF, Chu Q, Donaldson C, Tan D, Shokhirev MN, Saghatelian A. Accurate annotation of human protein-coding small open reading frames. *Nat Chem Biol*. 2020;16: 458–468.
107. Chen J, Brunner A-D, Cogan JZ, Nuñez JK, Fields AP, Adamson B, et al. Pervasive functional translation of noncanonical human open reading frames. *Science*. 2020;367: 1140–1146.
108. Kastenmayer JP, Ni L, Chu A, Kitchen LE, Au W-C, Yang H, et al. Functional genomics of genes with small open reading frames (sORFs) in *S. cerevisiae*. *Genome Res*. 2006;16: 365–373.
109. Galindo MI, Pueyo JI, Fouix S, Bishop SA, Couso JP. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol*. 2007;5: e106.
110. Magny EG, Pueyo JI, Pearl FMG, Cespedes MA, Niven JE, Bishop SA, et al. Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science*. 2013;341: 1116–1120.
111. Anderson DM, Anderson KM, Chang C-L, Makarewich CA, Nelson BR, McAnally JR, et al. A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell*. 2015;160: 595–606.
112. Anderson DM, Makarewich CA, Anderson KM, Shelton JM, Bezprozvannaya S, Bassel-Duby R, et al. Widespread control of calcium signaling by a family of SERCA-inhibiting micropeptides. *Sci Signal*. 2016;9: ra119.

113. Nelson BR, Makarewich CA, Anderson DM, Winders BR, Troupes CD, Wu F, et al. A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science*. 2016;351: 271–275.
114. Niu L, Lou F, Sun Y, Sun L, Cai X, Liu Z, et al. A micropeptide encoded by lncRNA MIR155HG suppresses autoimmune inflammation via modulating antigen presentation. *Sci Adv*. 2020;6: eaaz2059.
115. Bhatta A, Atianand M, Jiang Z, Crabtree J, Blin J, Fitzgerald KA. A Mitochondrial Micropeptide Is Required for Activation of the Nlrp3 Inflammasome. *J Immunol*. 2020;204: 428–437.
116. Banerjee S, Ghoshal S, Stevens JR, McCommis KS, Gao S, Castro-Sepulveda M, et al. Hepatocyte expression of the micropeptide adropin regulates the liver fasting response and is enhanced by caloric restriction. *J Biol Chem*. 2020;295: 13753–13768.
117. Spencer HL, Sanders R, Boulberdaa M, Meloni M, Cochrane A, Spiroski A-M, et al. The LINC00961 transcript and its encoded micropeptide, small regulatory polypeptide of amino acid response, regulate endothelial cell function. *Cardiovasc Res*. 2020;116: 1981–1994.
118. Zhang S, Reljić B, Liang C, Kerouanton B, Francisco JC, Peh JH, et al. Mitochondrial peptide BRAWNIN is essential for vertebrate respiratory complex III assembly. *Nat Commun*. 2020;11: 1312.
119. Makarewich CA, Baskin KK, Munir AZ, Bezprozvannaya S, Sharma G, Khemtong C, et al. MOXI Is a Mitochondrial Micropeptide That Enhances Fatty Acid β -Oxidation. *Cell Rep*. 2018;23: 3701–3709.
120. Matsumoto A, Pasut A, Matsumoto M, Yamashita R, Fung J, Monteleone E, et al. mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature*. 2017;541: 228–232.
121. Ji Z, Song R, Regev A, Struhl K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *Elife*. 2015;4: e08890.
122. Johnstone TG, Bazzini AA, Giraldez AJ. Upstream ORFs are prevalent translational repressors in vertebrates. *EMBO J*. 2016;35: 706–723.
123. 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;106: 7507–7512.
124. Chew G-L, Pauli A, Schier AF. Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish. *Nat Commun*. 2016;7: 11663.
125. Neafsey DE, Galagan JE. Dual modes of natural selection on upstream open reading frames. *Mol Biol Evol*. 2007;24: 1744–1751.

126. Raney A, Law GL, Mize GJ, Morris DR. Regulated translation termination at the upstream open reading frame in s-adenosylmethionine decarboxylase mRNA. *J Biol Chem.* 2002;277: 5988–5994.
127. Karagyozov L, Godfrey R, Böhmer S-A, Petermann A, Hölters S, Östman A, et al. The structure of the 5'-end of the protein-tyrosine phosphatase PTPRJ mRNA reveals a novel mechanism for translation attenuation. *Nucleic Acids Research.* 2008. pp. 4443–4453. doi:10.1093/nar/gkn391
128. uAUG and uORFs in human and rodent 5'untranslated mRNAs. *Gene.* 2005;349: 97–105.
129. Shah P, Ding Y, Niemczyk M, Kudla G, Plotkin JB. Rate-Limiting Steps in Yeast Protein Translation. *Cell.* 2013. pp. 1589–1601. doi:10.1016/j.cell.2013.05.049
130. Chan LY, Mugler CF, Heinrich S, Vallotton P, Weis K. Non-invasive measurement of mRNA decay reveals translation initiation as the major determinant of mRNA stability. *Elife.* 2018;7. doi:10.7554/eLife.32536
131. LaGRANDEUR T, Parker R. The cis acting sequences responsible for the differential decay of the unstable MFA2 and stable PGK1 transcripts in yeast include the context of the translational start codon. *RNA.* 1999;5: 420–433.
132. Schwartz DC, Parker R. mRNA Decapping in Yeast Requires Dissociation of the Cap Binding Protein, Eukaryotic Translation Initiation Factor 4E. *Mol Cell Biol.* 2000;20: 7933–7942.
133. Schwartz DC, Parker R. Mutations in Translation Initiation Factors Lead to Increased Rates of Deadenylation and Decapping of mRNAs in *Saccharomyces cerevisiae*. *Mol Cell Biol.* 1999;19: 5247–5256.
134. Beelman CA, Parker R. Differential effects of translational inhibition in cis and in trans on the decay of the unstable yeast MFA2 mRNA. *J Biol Chem.* 1994;269: 9687–9692.
135. Whiffin N, Karczewski KJ, Zhang X, Chothani S, Smith MJ, Gareth Evans D, et al. Characterising the loss-of-function impact of 5' untranslated region variants in 15,708 individuals. *bioRxiv.* 2019. p. 543504. doi:10.1101/543504
136. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv.* 2019. p. 531210. doi:10.1101/531210
137. Short PJ, McRae JF, Gallone G, Sifrim A, Won H, Geschwind DH, et al. De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature.* 2018;555: 611–616.

138. Lee DSM, Ghanem LR, Barash Y. Integrative analysis reveals RNA G-quadruplexes in UTRs are selectively constrained and enriched for functional associations. *Nat Commun.* 2020;11: 1–12.
139. Cridge AG, Crowe-McAuliffe C, Mathew SF, Tate WP. Eukaryotic translational termination efficiency is influenced by the 3' nucleotides within the ribosomal mRNA channel. *Nucleic Acids Res.* 2018;46: 1927–1944.
140. Loughran G, Chou M-Y, Ivanov IP, Jungreis I, Kellis M, Kiran AM, et al. Evidence of efficient stop codon readthrough in four mammalian genes. *Nucleic Acids Res.* 2014;42: 8928–8938.
141. Floquet C, Hatin I, Rousset J-P, Bidou L. Statistical analysis of readthrough levels for nonsense mutations in mammalian cells reveals a major determinant of response to gentamicin. *PLoS Genet.* 2012;8: e1002608.
142. Manuvakhova M, Keeling K, Bedwell DM. Aminoglycoside antibiotics mediate context-dependent suppression of termination codons in a mammalian translation system. *RNA.* 2000;6: 1044–1055.
143. Fearon K, McClendon V, Bonetti B, Bedwell DM. Premature translation termination mutations are efficiently suppressed in a highly conserved region of yeast Ste6p, a member of the ATP-binding cassette (ABC) transporter family. *J Biol Chem.* 1994;269: 17802–17808.
144. A direct estimation of the context effect on the efficiency of termination. *J Mol Biol.* 1998;284: 579–590.
145. Poole ES, Brown CM, Tate WP. The identity of the base following the stop codon determines the efficiency of in vivo translational termination in *Escherichia coli*. *EMBO J.* 1995;14: 151–158.
146. Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming ES, et al. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.* 2014;33: 981–993.
147. Hanson G, Collier J. Codon optimality, bias and usage in translation and mRNA decay. *Nat Rev Mol Cell Biol.* 2018;19: 20–30.
148. Lin Y, May GE, Kready H, Nazzaro L, Mao M, Spealman P, et al. Impacts of uORF codon identity and position on translation regulation. *Nucleic Acids Res.* 2019;47: 9358–9367.
149. Translational regulation of human methionine synthase by upstream open reading frames. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression.* 2007;1769: 532–540.

150. Fervers P, Fervers F, Makalowski W, Jankalski M. Life cycle adapted upstream open reading frames (uORFs) in *Trypanosoma congolense*: A post-transcriptional approach to accurate gene regulation. *PLoS One*. 2018;13: e0201461.
151. Bettany AJ, Moore PA, Cafferkey R, Bell LD, Goodey AR, Carter BL, et al. 5'-secondary structure formation, in contrast to a short string of non-preferred codons, inhibits the translation of the pyruvate kinase mRNA in yeast. *Yeast*. 1989;5: 187–198.
152. Kudla G, Murray AW, Tollervey D, Plotkin JB. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science*. 2009;324: 255–258.
153. Wu Q, Medina SG, Kushawah G, DeVore ML, Castellano LA, Hand JM, et al. Translation affects mRNA stability in a codon-dependent manner in human cells. *Elife*. 2019;8. doi:10.7554/eLife.45396
154. 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;8: 2395.
155. Park J, Levin MG, Haggerty CM, Hartzel DN, Judy R, Kember RL, et al. A genome-first approach to aggregating rare genetic variants in LMNA for association with electronic health record phenotypes. *Genet Med*. 2020;22: 102–111.
156. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol*. 2013;31: 1102–1110.
157. Diogo D, Tian C, Franklin CS, Alanne-Kinnunen M, March M, Spencer CCA, et al. Phenome-wide association studies across large population cohorts support drug target validation. *Nat Commun*. 2018;9: 4285.
158. Park J, Katz N, Zhang X, Lucas AM, Verma A, Judy RL, et al. Exome-by-phenome-wide rare variant gene burden association with electronic health record phenotypes. *bioRxiv*. 2019. p. 798330. doi:10.1101/798330
159. Aliouat A, Hatin I, Bertin P, François P, Stierlé V, Namy O, et al. Divergent effects of translation termination factor eRF3A and nonsense-mediated mRNA decay factor UPF1 on the expression of uORF carrying mRNAs and ribosome protein genes. *RNA Biol*. 2020;17: 227–239.
160. Zhang Y, Pelechano V. High-throughput 5'P sequencing reveals environmental regulated ribosome stalls at termination level. doi:10.1101/2020.06.22.165134
161. Meijer HA, Thomas AAM. Ribosomes stalling on uORF1 in the *Xenopus* Cx41 5' UTR inhibit downstream translation initiation. *Nucleic Acids Res*. 2003;31: 3174–3184.
162. Fang P, Wang Z, Sachs MS. Evolutionarily conserved features of the arginine attenuator peptide provide the necessary requirements for its function in translational regulation. *J Biol Chem*. 2000;275: 26710–26719.

163. Hurt JA, Robertson AD, Burge CB. Global analyses of UPF1 binding and function reveal expanded scope of nonsense-mediated mRNA decay. *Genome Res.* 2013;23: 1636–1650.
164. Lee M-H. Translation repression by GLD-1 protects its mRNA targets from nonsense-mediated mRNA decay in *C. elegans*. *Genes Dev.* 2004;18: 1047–1059.
165. Gaba A, Jacobson A, Sachs MS. Ribosome occupancy of the yeast CPA1 upstream open reading frame termination codon modulates nonsense-mediated mRNA decay. *Mol Cell.* 2005;20: 449–460.
166. Blobel G, Potter VR. Studies on free and membrane-bound ribosomes in rat liver. I. Distribution as related to total cellular RNA. *J Mol Biol.* 1967;26: 279–292.
167. García-Cazorla À, Verdura E, Juliá-Palacios N, Anderson EN, Goicoechea L, Planas-Serra L, et al. Impairment of the mitochondrial one-carbon metabolism enzyme SHMT2 causes a novel brain and heart developmental syndrome. *Acta Neuropathol.* 2020;140: 971–975.
168. Ward NC, Watts GF, Eckel RH. Statin Toxicity. *Circ Res.* 2019;124: 328–350.
169. Waters DD, Ho JE, DeMicco DA, Breazna A, Arsenault BJ, Wun C-C, et al. Predictors of new-onset diabetes in patients treated with atorvastatin: results from 3 large randomized clinical trials. *J Am Coll Cardiol.* 2011;57: 1535–1545.
170. Preiss D, Seshasai SRK, Welsh P, Murphy SA, Ho JE, Waters DD, et al. Risk of incident diabetes with intensive-dose compared with moderate-dose statin therapy: a meta-analysis. *JAMA.* 2011;305: 2556–2564.
171. Ference BA, Robinson JG, Brook RD, Catapano AL, Chapman MJ, Neff DR, et al. Variation in PCSK9 and HMGCR and Risk of Cardiovascular Disease and Diabetes. *N Engl J Med.* 2016;375: 2144–2153.
172. Lotta LA, Sharp SJ, Burgess S, Perry JRB, Stewart ID, Willems SM, et al. Association Between Low-Density Lipoprotein Cholesterol-Lowering Genetic Variants and Risk of Type 2 Diabetes: A Meta-analysis. *JAMA.* 2016;316: 1383–1391.
173. Wen Y, Liu Y, Xu Y, Zhao Y, Hua R, Wang K, et al. Loss-of-function mutations of an inhibitory upstream ORF in the human hairless transcript cause Marie Unna hereditary hypotrichosis. *Nat Genet.* 2009;41: 228–233.
174. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol.* 2017;186: 1026–1034.
175. Sivagnanasundaram S, Morris AG, Gaitonde EJ, McKenna PJ, Mollon JD, Hunt DM. A cluster of single nucleotide polymorphisms in the 5'-leader of the human dopamine D3 receptor gene (DRD3) and its relationship to schizophrenia. *Neurosci Lett.* 2000;279: 13–16.

176. Beffagna G, Occhi G, Nava A, Vitiello L, Ditadi A, Basso C, et al. Regulatory mutations in transforming growth factor-beta3 gene cause arrhythmogenic right ventricular cardiomyopathy type 1. *Cardiovasc Res.* 2005;65: 366–373.
177. Niesler B, Flohr T, Nöthen MM, Fischer C, Rietschel M, Franzek E, et al. Association between the 5' UTR variant C178T of the serotonin receptor gene HTR3A and bipolar affective disorder. *Pharmacogenetics.* 2001;11: 471–475.
178. Pasaje CFA, Bae JS, Park B-L, Cheong HS, Kim J-H, Uh S-T, et al. WDR46 is a Genetic Risk Factor for Aspirin-Exacerbated Respiratory Disease in a Korean Population. *Allergy Asthma Immunol Res.* 2012;4: 199–205.
179. Luo E-C, Nathanson JL, Tan FE, Schwartz JL, Schmok JC, Shankar A, et al. Large-scale tethered function assays identify factors that regulate mRNA stability and translation. *Nat Struct Mol Biol.* 2020;27: 989–1000.
180. Ivanov IP, Loughran G, Atkins JF. uORFs with unusual translational start codons autoregulate expression of eukaryotic ornithine decarboxylase homologs. *Proc Natl Acad Sci U S A.* 2008;105: 10079–10084.
181. Luukkonen BG, Tan W, Schwartz S. Efficiency of reinitiation of translation on human immunodeficiency virus type 1 mRNAs is determined by the length of the upstream open reading frame and by intercistronic distance. *J Virol.* 1995;69: 4086–4094.
182. Ferreira JP, Overton KW, Wang CL. Tuning gene expression with synthetic upstream open reading frames. *Proc Natl Acad Sci U S A.* 2013;110: 11284–11289.
183. Dvir S, Velten L, Sharon E, Zeevi D, Carey LB, Weinberger A, et al. Deciphering the rules by which 5'-UTR sequences affect protein expression in yeast. *Proc Natl Acad Sci U S A.* 2013;110: E2792–801.
184. 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;19: 229.
185. Wang F, Zuroske T, Watts JK. RNA therapeutics on the rise. *Nat Rev Drug Discov.* 2020;19. doi:10.1038/d41573-020-00078-0
186. Shen X, Corey DR. Chemistry, mechanism and clinical status of antisense oligonucleotides and duplex RNAs. *Nucleic Acids Res.* 2018;46: 1584–1600.
187. Geary RS, Henry SP, Grillone LR. Fomivirsen: clinical pharmacology and potential drug interactions. *Clin Pharmacokinet.* 2002;41: 255–260.
188. Roberts TC, Langer R, Wood MJA. Advances in oligonucleotide drug delivery. *Nat Rev Drug Discov.* 2020;19: 673–694.
189. Kharel P, Balaratnam S, Beals N, Basu S. The role of RNA G-quadruplexes in human diseases and therapeutic strategies. *Wiley Interdiscip Rev RNA.* 2020;11: e1568.

190. Liang X-H, 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;34: 875–880.
191. Liang X-H, Shen W, Crooke ST. Specific Increase of Protein Levels by Enhancing Translation Using Antisense Oligonucleotides Targeting Upstream Open Frames. *Adv Exp Med Biol.* 2017;983: 129–146.
192. Xu Y, Poggio M, Jin HY, Shi Z, Forester CM, Wang Y, et al. Translation control of the immune checkpoint in cancer and its therapeutic targeting. *Nat Med.* 2019;25: 301–311.
193. Liang X-H, 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;45: 9528–9546.
194. Sasaki S, Sun R, Bui H-H, Crosby JR, Monia BP, Guo S. Steric Inhibition of 5' UTR Regulatory Elements Results in Upregulation of Human CFTR. *Molecular Therapy.* 2019. pp. 1749–1757. doi:10.1016/j.ymthe.2019.06.016
195. Zamiri B, Reddy K, Macgregor RB Jr, Pearson CE. TMPyP4 porphyrin distorts RNA G-quadruplex structures of the disease-associated r(GGGGCC)_n repeat of the C9orf72 gene and blocks interaction of RNA-binding proteins. *J Biol Chem.* 2014;289: 4653–4659.
196. Simone R, Balendra R, Moens TG, Preza E, Wilson KM, Heslegrave A, et al. G-quadruplex-binding small molecules ameliorate FTD/ALS pathology and. *EMBO Mol Med.* 2018;10: 22–31.
197. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols.* 2009. pp. 1184–1191. doi:10.1038/nprot.2009.97
198. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics.* 2005;21: 3439–3440.
199. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal.* 2011;17: 10–12.
200. Yee TW. *Vector Generalized Linear and Additive Models: With an Implementation in R.* Springer; 2015.
201. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 2010;38: e164.
202. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 2016;44: D733–45.

203. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet.* 2016;99: 877–885.
204. Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics.* 2014. pp. 2375–2376. doi:10.1093/bioinformatics/btu197
205. Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, Wei L-J, et al. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet.* 2010;86: 832–838.