Disrupting upstream translation in mRNAs leads to loss-of-function associated with human disease

Authors: David S.M. Lee[1,2], Joseph Park[1,2,3], Andrew Kromer[1], Regeneron Genetics Center[4], Daniel J. Rader[1,3,5], Marylyn D. Ritchie[1,2], Louis R. Ghanem[6,7]*†, Yoseph Barash[1,5,8]*.

Affiliations:

[1]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

[2]Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

[3]Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

[4]Regeneron Genetics Center, Regeneron Pharmaceuticals, Tarrytown, NY, USA.

[5]Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

[6]Division of Gastroenterology, Hepatology and Nutrition, The Children's Hospital of Philadelphia, Philadelphia, PA, USA.

[7]Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

[8]Department of Computer and Information Science, School of Engineering, University of Pennsylvania, Philadelphia, PA, USA.

*Correspondence to: lghanem@its.jnj.com, yosephb@upenn.edu

†Present address: Janssen Research and Development, Spring House, PA, USA.

ABSTRACT

Ribosome-profiling has uncovered pervasive translation in 5'UTRs, 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 demonstrate their impact on gene expression in human cells. Our results establish new mechanisms relating uORF variation to loss-of-function of downstream genes,

and demonstrate that translated uORFs are genetically constrained regulatory elements in 40% of human genes.

INTRODUCTION

The classic view of information processing in the cell by gene expression occurs through transcription and 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 gene expression by tuning translation rates of downstream protein-coding sequences, and potential uORFs have been identified in ~50% of all human protein-coding genes[1,2]. 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[2,3]. In contrast, less is known about the impact of genetic variation affecting translated uORFs. 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.

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 that introduce new stop codons, or create 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 these predicted uORF-disrupting variants, and replicate these associations in the UK Biobank (UKB), and by gene burden tests that aggregate rare protein-coding loss-of-function variants. Finally we validate the impact of these uORF variants on gene expression for our top phenome-wide significant associations with diabetes and anxiety

disorders in 5'UTRs from *PMVK* and *VPS53* respectively. These data demonstrate that SNVs in translated uORFs that create new stop codons, or strengthen existing stop codons can contribute to disease pathology by changing gene expression.

RESULTS

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

Translation initiation is the rate-limiting step controlling post-transcriptional gene expression[4], and rates of translation initiation can significantly impact mRNA stability[5–9]. 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, 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 expression of the downstream gene. 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. Since 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-introducing SNVs, 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 determining the relative enrichment for rare singleton (one sequenced allele) variants within the gnomAD database, adjusted for local mutation rates (see **Methods**). 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[3,10–14].

Using translated uORFs from 4392 genes identified by deep ribosome profiling of two human cell lines (**Suppl. Fig. 1**)[15], we mapped genetic variation from 71,702 whole-genome sequences in gnomAD (version 3)[11]. We identified the subset of SNVs creating UTCs by selecting those which mutated uORF codons to either TGA, TAG, or TAA in the mapped uORF reading frame (**Figure 1a**). We calculated MAPS scores for these UTC-creating SNVs, finding that they are under strong negative selection within 5'UTRs, comparable to that of missense mutations in protein-coding regions of the genome (**Figure 1b-i,ii**). Indeed, MAPS scores for these variants are significantly higher than all uORF variants (**Figure 1b-ii**, P<0.001), sets of uORF variants matched by their underlying trinucleotide mutation context (**Figure 1b-ii**, P<0.001 - see **methods**), and stop-codon introducing variants in ORFs in 3'UTRs, translated pseudogenes, and lncRNAs also mapped by ribosome-profiling from the same study (**Figure 1b-i,ii**, P=0.0041, **Suppl. Fig. 1**). Intriguingly, MAPS scores were highest for variants predicted to introduce strong (TAA) stop codons that are less susceptible to translational read-through [16–18]. The strong selective pressure to remove UTC-creating SNVs implies that these variants are also more likely to have functional biological consequences.

Crucially, we find that the enrichment in rare singleton variants appears to depend on the relationship between the new UTC-creating mutation and the CDS start codon. A subset of translated uORFs are known to overlap out-of-frame with their respective downstream coding sequences. These uORFs can repress translation of the downstream CDS by obscuring the protein-coding start codon [19,20]. Since uORF-mediated repression is contingent on the uORF overlapping out-of-frame with the CDS start codon, we hypothesized that UTC-introducing variants maintaining the uORF-CDS overlap would not exhibit similar MAPS scores to those that create new stop codons upstream of the CDS start. Indeed MAPS scores for UTCs that abolish the uORF-CDS overlap are significantly higher compared to all uORF-variants (P = 0.038), while MAPS scores for UTCs that preserve the overlap between the uORF reading frame and downstream CDS are indistinguishable from all uORF SNVs (P=0.4984, **Figure 1b-iii**). These findings are consistent with the hypothesis that selective pressures against UTC-introducing variants are restricted to those that have the capacity to disrupt downstream translation initiation

since the enrichment in singletons is observed only when a UTC is introduced proximal to the CDS start codon.

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 TAA > TAG > TGA[16,21,22]. Given the observed selection against UTCs in translated uORFs, and in particular against TAA-introducing variants, we next asked whether stop codon usage by translated uORFs is distinct from the background distribution of TGA, TAG, and TAA trinucleotides in 5'UTRs. To perform this comparison, we determined the relative frequency that TGA, TAG, or TAA 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 TGA, TAG, or TAA trinucleotides from uORF-containing UTRs with translated uORF sequences excluded. Strikingly, we find that translated uORF stop codons are significantly depleted of TAAs compared to background UTR distributions (**Figure 1c**), suggesting that weaker stop-codons (TGA, TAG) 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 TAA stop codons compared to the relative frequency of TAA trinucleotides in adjacent untranslated UTR sequences (uORF-TAA=19%, matched UTR-TAA=35% - **Suppl. Table 1**). In contrast, TGA 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).

Given the depletion of TAA-stop codons in translated uORFs, we next asked whether variants changing weaker stop codons (TGA, TAG) to TAA 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 1b-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 1b-ii**). Given that TAA codons can facilitate greater termination efficiency and more rapid ribosomal dissociation from mRNAs compared to TAG and TGA codons[16,23,24], 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 UTCs, stronger stop codons in uORFs may be disfavored because they decrease the probability that ribosomes reinitiate translation at downstream coding sequences.

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[25]. 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 (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[3].

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 matched UTR positions. This difference remained significant even when compared to potential stop-creating positions in 5'UTR sequences adjacent to translated uORFs (**Figure 1d**). 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 UTC introduced and the proportion of uORF genomic positions that are conserved. This trend was not observed for non-translated 5'UTR contexts. In all cases, the proportion of conserved bases for each class of potential stop-creating variant was significantly higher than those positions within all 5'UTRs in general, and particularly within untranslated regions of translated-uORF containing UTRs ($P<0.001$, **Figure 1d**). This complementary analysis supports our initial findings that uORF UTC variants are under strong negative selection within the human genome, and further strengthens the evidence that uORF UTC variants may functionally disrupt gene expression.

Upstream open reading frames are not under 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[15,26,27]. This has fostered significant interest in the possibility that translated, non-canonical ORFs represent an overlooked class of potentially functional micropeptides with biological functions independent of the downstream protein-coding sequences[27,28]. While previous genome-wide approaches to assess uORF coding potential use metrics measuring evolutionary conservation across species, using human variation data directly allows us to capture potential lineage-specific constraints that may have been absent from cross-species analyses. Moreover, the pattern of constraint against UTC-introducing variants might also reflect selection to preserve putative micropeptide functionality. To address this possibility, we asked whether uORFs broadly exhibit similar constraints against missense variation as known protein-coding regions of the genome that would imply peptide functionality. We compared MAPS scores for predicted missense versus synonymous variants in translated uORFs to those in canonical protein-coding regions of the genome (**Figure 2a**). The frequency spectrum for missense variants in uORFs were significantly lower than that of missense variants 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 2a-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 SNVs 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 2a-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 2a-v**).

Surprisingly, we observed that MAPS scores for both synonymous and missense variants in translated uORFs deviated significantly from all 5'UTR variation, implying that uORF variants are generally under a heightened degree of negative selection compared to synonymous variants in protein-coding sequences (**Figure 2a-iv**). 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 transitional efficiency by preserving codon optimality. Messenger RNAs that are enriched with more optimal codons are both more stable, and more efficiently translated by ribosomes [29]. Like UTCs, 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 [30–32], and 5'UTRs are generally known to be under selective pressures to maintain their capacity for facilitating translation initiation at the CDS [33,34].

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 2b**). Using experimentally-determined codon-stability coefficients (CSCs) [35], 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 2c**). 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-introducing variants (**Figure 1b**). 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 2c, Suppl. Figure 2**). 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 (**Suppl. Figure 3**) [35]. 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.

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[19]. 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 3a-i**), we distinguish between variants maintaining the start context by affecting the first position of the NTG trinucleotide from those that disrupt translation initiation by mutating the last two nucleotides in the uORF start codon (**Figure 3a-ii**). As expected, start-maintaining variants are no more enriched for singletons in gnomAD compared to synonymous protein coding variants. In contrast, start-disrupting SNVs are enriched for singletons at a level comparable to that of protein-coding missense SNVs, and SNVs introducing UTCs (**Figure 3b**). 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 3c**). 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, UTC-introducing, and stop-strengthening variants. 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.

uORF-disrupting variants associate genes with new disease phenotypes

The heightened MAPS score for UTC-introducing variants suggests that they are also likely to be functional. To explore the possibility that uORF UTC and stop-strengthening variants 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[36].

Using exome sequencing from the PMBB, we identified heterozygous and homozygous individuals carrying mutations in uORFs which introduce upstream termination codons and stop-strengthening mutations. For the latter class, we focused on variants that introduced TAA 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 TAA-UTCs). For each of these variants 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[37,38], including 5/6 of the strop-strengthening variants and 1/4 of the TAA-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 4, Suppl. Fig. 4**, **Table 1**).

Replication of novel associations in UK Biobank

Out of six novel associations reaching FDR < 0.1, two showed P < 0.05 in UK Biobank (UKB) with consistent direction of effect, validating and further strengthening the significance of our previous results. Of these replicated associations, a stop-strengthening variant in *BCL2L13* reached study-wide significance. 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. For non-replicated associations, variants in *NALCN* and *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. Overall, single-variant analysis in the UK Biobank confirmed novel associations in *PMVK* and *BCL2L13* respectively.

Disease-associated uORF variants change expression

To elucidate the possible biological consequences of UTC and stop-strengthening mutations, we selected our top two PheWAS association signals for functional assessment. To determine if these variants could affect gene expression, we measured the expression of a set of dual-luciferase reporters in HEK293T cells for *PMVK*, and *VPS53* uORF variants. We compared the expression of the wild-type 5'UTR sequence for *PMVK* and *VPS53* cloned upstream of a Firefly Luciferase ORF to two variant sequences - one with the 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 TGG codon to a TAG upstream termination codon (**Figure 5b**). Across all constructs, we observed a significant reduction in expression of the downstream ORF when the PheWAS-significant stop-strengthening mutation was introduced (**Figure 5**). Introducing a new UTC in the 5'UTR uORF of *VPS53* also significantly reduced reporter gene expression relative to the wild-type sequence. Similar results were obtained from assays performed in HeLa cells (**Suppl. Fig. 9**).

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

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

Results from reporter-gene experiments showed that UTCs and stop-strengthening variants could decrease expression of the downstream protein for *PMVK*, and *VPS53*. Our findings implied that uORF UTC and stop-strengthening variants 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 variants. 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 studies that utilized these two datasets [36,39].

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**). 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 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.

DISCUSSION

By combining large databases of human genetic variation with ribosome profiling, we have 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 2 of these associations in independent single-variant association tests in the UKB. Two associations involving stop-strengthening variants in *PMVK* and *SHMT2* were also

replicated using protein-coding mutations in loss-of-function gene burden tests. This result provides independent validation that *PMVK* and *SHMT2* loss-of-function is associated with diabetes and diseases of the salivary gland respectively, and further demonstrates that uORF stop-strengthening mutations produce the same phenotype as loss-of-function of downstream coding genes. In support of these conclusions,we have shown that introducing UTCs and stop-strengthening variants in translated uORFs decreases expression of downstream genes in reporter assays. These findings establish that uORF UTC and stop-strengthening variants can have functional consequences on gene expression and cause disease in humans. Based on the degree of enrichment in rare singleton variants in gnomAD, 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, compared to ~5-15% estimated to be under constraint for amino acid function (see **Suppl. Note 1**), although we have not assessed this further in the present study. 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 [27]. 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 variants in these genes may have additional utility for the diagnosis of rare disease.

Our results suggest uORF translation has broad roles in regulating CDS expression. Translation initiation is rate-limiting for CDS expression 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 also inline with cis-regulatory relationships between uORFs and downstream coding sequences being frequently conserved across vertebrates, but features which confer strong repression on CDS expression - including strong uORF start codons, and longer uORF length - are less conserved[19,40]. For stop-strengthening variants, the increased translation termination efficiency could accelerate ribosomal release from the mRNA transcript thus decreasing downstream CDS translation. This suggested 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[41]. For UTC-introducing variants, the introduction of stop codons in the uORF may lead to either

ribosome stalling and subsequent collisions that further repress CDS expression[42,43]. This early translation termination within uORFs could 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[44–46], whether uORF-activated NMD broadly regulates gene 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[41].

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[15]. 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-variants within 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 uORFs 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[47]. As improvements in ribosome profiling facilitate deeper characterization of the translatome, observations of widespread translation in non-canonical ORFs should be interpreted cautiously in light of potential functionality.

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[48–50]. 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 [51,52], 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 genetic variants in both these genes confer an increased risk of diabetes through a similar mechanism, however further 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. Specifically the strength of the uORF start codon[53], intercistronic distance between the uORF stop codon and downstream coding gene[54], and additional contributions of secondary structure in the 5'UTR have all been shown to impact uORF regulatory functions previously[55]. The contributions of these factors to uORF-mediated translational regulation can be highly context specific, and dissecting these differences in regulation remains an interesting 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. 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 is a healthy, population-based biobank. 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[56–59], 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.

References

1. uAUG and uORFs in human and rodent 5′untranslated mRNAs. *Gene* **349**, 97–105 (2005).
2. Calvo, S. E., Pagliarini, D. J. & Mootha, V. K. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 7507–7512 (2009).
3. Whiffin, N. *et al.* Characterising the loss-of-function impact of 5' untranslated region variants in 15,708 individuals. *bioRxiv* 543504 (2019) doi:10.1101/543504.
4. Shah, P., Ding, Y., Niemczyk, M., Kudla, G. & Plotkin, J. B. Rate-Limiting Steps in Yeast Protein Translation. *Cell* vol. 153 1589–1601 (2013).
5. Chan, L. Y., Mugler, C. F., Heinrich, S., Vallotton, P. & Weis, K. Non-invasive measurement of mRNA decay reveals translation initiation as the major determinant of mRNA stability. *Elife* **7**, (2018).
6. 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* **5**, 420–433 (1999).
7. Schwartz, D. C. & Parker, R. mRNA Decapping in Yeast Requires Dissociation of the Cap Binding Protein, Eukaryotic Translation Initiation Factor 4E. *Mol. Cell. Biol.* **20**, 7933–7942 (2000).
8. Schwartz, D. C. & Parker, R. Mutations in Translation Initiation Factors Lead to Increased Rates of Deadenylation and Decapping of mRNAs inSaccharomyces cerevisiae. *Mol. Cell. Biol.* **19**, 5247–5256 (1999).
9. Beelman, C. A. & Parker, R. Differential effects of translational inhibition in cis and in trans on the decay of the unstable yeast MFA2 mRNA. *J. Biol. Chem.* **269**, 9687–9692 (1994).
10. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**,

285–291 (2016).

11. Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* 531210 (2019) doi:10.1101/531210.
12. Short, P. J. *et al.* De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* **555**, 611–616 (2018).
13. Lee, D. S. M., Ghanem, L. R. & Barash, Y. Integrative analysis reveals RNA G-quadruplexes in UTRs are selectively constrained and enriched for functional associations. *Nat. Commun.* **11**, 1–12 (2020).
14. Zhang, S. *et al.* Base-specific mutational intolerance near splice sites clarifies the role of nonessential splice nucleotides. *Genome Res.* **28**, 968–974 (2018).
15. Ji, Z., Song, R., Regev, A. & Struhl, K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. (2015) doi:10.7554/eLife.08890.
16. Cridge, A. G., Crowe-McAuliffe, C., Mathew, S. F. & Tate, W. P. Eukaryotic translational termination efficiency is influenced by the 3' nucleotides within the ribosomal mRNA channel. *Nucleic Acids Res.* **46**, 1927–1944 (2018).
17. Loughran, G. *et al.* Evidence of efficient stop codon readthrough in four mammalian genes. *Nucleic Acids Res.* **42**, 8928–8938 (2014).
18. 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.* **8**, e1002608 (2012).
19. Johnstone, T. G., Bazzini, A. A. & Giraldez, A. J. Upstream ORFs are prevalent translational repressors in vertebrates. *EMBO J.* **35**, 706–723 (2016).

20. Barbosa, C., Peixeiro, I. & Romão, L. Gene expression regulation by upstream open reading frames and human disease. *PLoS Genet.* **9**, e1003529 (2013).
21. Manuvakhova, M., Keeling, K. & Bedwell, D. M. Aminoglycoside antibiotics mediate context-dependent suppression of termination codons in a mammalian translation system. *RNA* **6**, 1044–1055 (2000).
22. Fearon, K., McClendon, V., Bonetti, B. & Bedwell, D. M. 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.* **269**, 17802–17808 (1994).
23. A direct estimation of the context effect on the efficiency of termination. *J. Mol. Biol.* **284**, 579–590 (1998).
24. Poole, E. S., Brown, C. M. & Tate, W. P. The identity of the base following the stop codon determines the efficiency of in vivo translational termination in Escherichia coli. *EMBO J.* **14**, 151–158 (1995).
25. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
26. Martinez, T. F. *et al.* Accurate annotation of human protein-coding small open reading frames. *Nat. Chem. Biol.* **16**, 458–468 (2020).
27. Chen, J. *et al.* Pervasive functional translation of noncanonical human open reading frames. *Science* **367**, 1140–1146 (2020).
28. Bazzini, A. A. *et al.* Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.* **33**, 981–993 (2014).
29. Hanson, G. & Coller, J. Codon optimality, bias and usage in translation and mRNA decay. *Nat. Rev. Mol. Cell Biol.* **19**, 20–30 (2018).

30. Lin, Y. *et al.* Impacts of uORF codon identity and position on translation regulation. *Nucleic Acids Res.* **47**, 9358–9367 (2019).
31. Translational regulation of human methionine synthase by upstream open reading frames. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression* **1769**, 532–540 (2007).
32. Fervers, P., Fervers, F., Makałowski, W. & Jąkalski, M. Life cycle adapted upstream open reading frames (uORFs) in Trypanosoma congolense: A post-transcriptional approach to accurate gene regulation. *PLoS One* **13**, e0201461 (2018).
33. Bettany, A. J. *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* **5**, 187–198 (1989).
34. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-sequence determinants of gene expression in Escherichia coli. *Science* **324**, 255–258 (2009).
35. Wu, Q. *et al.* Translation affects mRNA stability in a codon-dependent manner in human cells. *Elife* **8**, (2019).
36. Park, J. *et al.* A genome-first approach to aggregating rare genetic variants in LMNA for association with electronic health record phenotypes. *Genet. Med.* **22**, 102–111 (2020).
37. Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1110 (2013).
38. Diogo, D. *et al.* Phenome-wide association studies across large population cohorts support drug target validation. *Nat. Commun.* **9**, 4285 (2018).
39. Park, J. *et al.* Exome-by-phenome-wide rare variant gene burden association with electronic health record phenotypes. *bioRxiv* 798330 (2019) doi:10.1101/798330.

40. Chew, G.-L., Pauli, A. & Schier, A. F. Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish. *Nat. Commun.* **7**, 11663 (2016).

41. Aliouat, A. *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.* **17**, 227–239 (2020).

42. Meijer, H. A. & Thomas, A. A. M. Ribosomes stalling on uORF1 in the Xenopus Cx41 5' UTR inhibit downstream translation initiation. *Nucleic Acids Res.* **31**, 3174–3184 (2003).

43. Fang, P., Wang, Z. & Sachs, M. S. Evolutionarily conserved features of the arginine attenuator peptide provide the necessary requirements for its function in translational regulation. *J. Biol. Chem.* **275**, 26710–26719 (2000).

44. Hurt, J. A., Robertson, A. D. & Burge, C. B. Global analyses of UPF1 binding and function reveal expanded scope of nonsense-mediated mRNA decay. *Genome Res.* **23**, 1636–1650 (2013).

45. Lee, M.-H. Translation repression by GLD-1 protects its mRNA targets from nonsense-mediated mRNA decay in C. elegans. *Genes Dev.* **18**, 1047–1059 (2004).

46. Gaba, A., Jacobson, A. & Sachs, M. S. Ribosome occupancy of the yeast CPA1 upstream open reading frame termination codon modulates nonsense-mediated mRNA decay. *Mol. Cell* **20**, 449–460 (2005).

47. Blobel, G. & Potter, V. R. Studies on free and membrane-bound ribosomes in rat liver. I. Distribution as related to total cellular RNA. *J. Mol. Biol.* **26**, 279–292 (1967).

48. Ward, N. C., Watts, G. F. & Eckel, R. H. Statin Toxicity. *Circ. Res.* **124**, 328–350 (2019).

49. Waters, D. D. *et al.* Predictors of new-onset diabetes in patients treated with atorvastatin: results from 3 large randomized clinical trials. *J. Am. Coll. Cardiol.* **57**, 1535–1545 (2011).

50. Preiss, D. *et al.* Risk of incident diabetes with intensive-dose compared with moderate-dose

statin therapy: a meta-analysis. *JAMA* **305**, 2556–2564 (2011).

51. Ference, B. A. *et al.* Variation in PCSK9 and HMGCR and Risk of Cardiovascular Disease and Diabetes. *N. Engl. J. Med.* **375**, 2144–2153 (2016).
52. Lotta, L. A. *et al.* Association Between Low-Density Lipoprotein Cholesterol-Lowering Genetic Variants and Risk of Type 2 Diabetes: A Meta-analysis. *JAMA* **316**, 1383–1391 (2016).
53. Ivanov, I. P., Loughran, G. & Atkins, J. F. uORFs with unusual translational start codons autoregulate expression of eukaryotic ornithine decarboxylase homologs. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 10079–10084 (2008).
54. Luukkonen, B. G., 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.* **69**, 4086–4094 (1995).
55. Dvir, S. *et al.* Deciphering the rules by which 5'-UTR sequences affect protein expression in yeast. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E2792–801 (2013).
56. Sivagnanasundaram, S. *et al.* 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.* **279**, 13–16 (2000).
57. Beffagna, G. *et al.* Regulatory mutations in transforming growth factor-beta3 gene cause arrhythmogenic right ventricular cardiomyopathy type 1. *Cardiovasc. Res.* **65**, 366–373 (2005).
58. Niesler, B. *et al.* Association between the 5' UTR variant C178T of the serotonin receptor gene HTR3A and bipolar affective disorder. *Pharmacogenetics* **11**, 471–475 (2001).
59. Pasaje, C. F. A. *et al.* WDR46 is a Genetic Risk Factor for Aspirin-Exacerbated Respiratory Disease in a Korean Population. *Allergy Asthma Immunol. Res.* **4**, 199–205 (2012).

60. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).

61. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–45 (2016).

62. Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).

63. Carroll, R. J., Bastarache, L. & Denny, J. C. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* vol. 30 2375–2376 (2014).

64. Price, A. L. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* **86**, 832–838 (2010).

Acknowledgements

We thank Dr. Benjamin Voight, Dr. Robert Heuckeroth, and members of the Barash lab for discussions and thoughtful feedback. D.L. is supported by the NIH grant 5T32HG000046-20. Y.B. and D.L. work was supported by R01 GM128096.

Contributions

D.L., L.G., and Y.B. conceived and designed the project. D.L. performed the analyses under the guidance of L.G. and Y.B. J.P. performed the PheWAS association studies under the guidance of D.R. and M.R. J.P. wrote the methods for the PheWAS analysis. D.L. wrote the paper and all authors contributed to editing the paper.

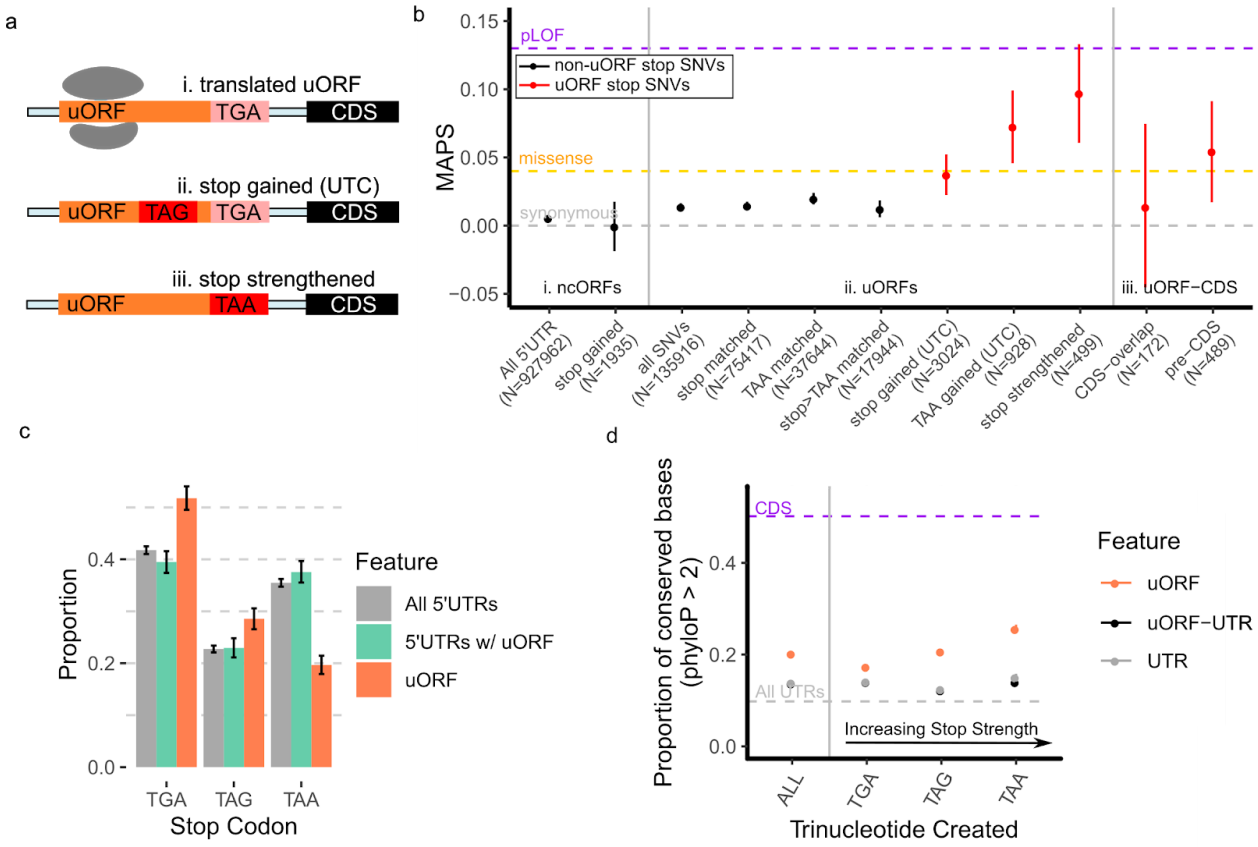

Figure 1: Stop-introducing, and stop-strengthening mutations in translated uORFs are under strong negative selection. (a) Examples of possible stop-gained (UTC) or stop-strengthened mutations in translated uORFs. (b) Mutability-Adjusted Proportion of Singletons (MAPS) scores for different classes of stop-introducing mutations within translated uORFs. (i) All 5'UTR SNVs, and nonsense mutations in ncORFs do not significantly deviate in MAPS scores from the CDS synonymous estimate. (ii) MAPS scores for uORF UTC-creating variants are significantly higher than matched non-UTC-creating uORF variants. This is also observed for TAA-creating, and stop-strengthening SNVs in translated uORFs. (iii) MAPS scores for UTCs that abolish the uORF-CDS overlap are more constrained compared to stop-creating variants that maintain the uORF-CDS overlap. Error bars represent bootstrapped 90% confidence intervals. (c) 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 (TGA, TAG) stop codons and depleted of the TAA stop codons compared to control sequences. Error bars represent 95% bootstrapped confidence intervals. (d) 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-UTRs, and non-uORF stop-creating in all UTR genomic positions. Error bars represent 90% bootstrapped confidence intervals.

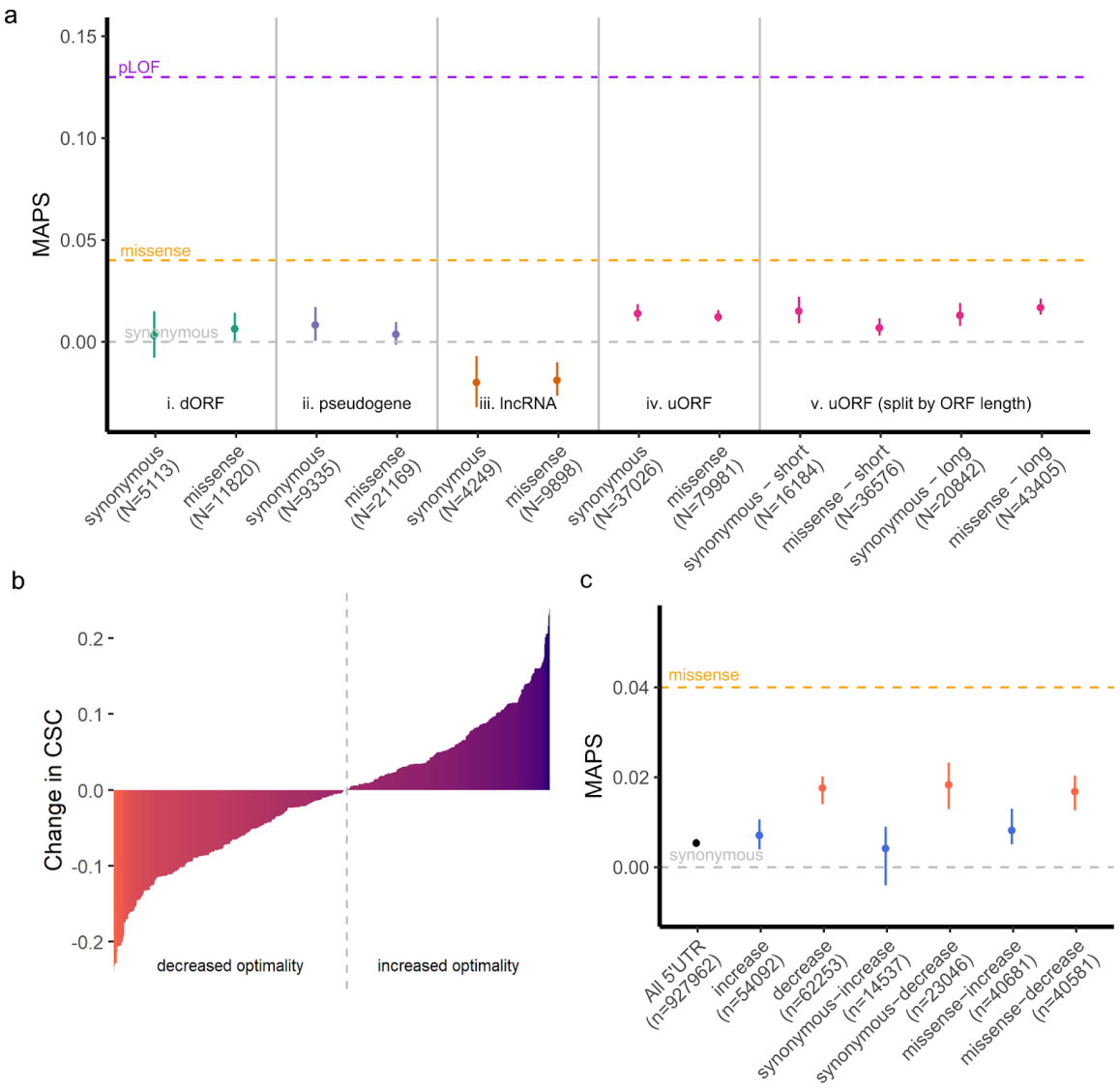

Figure 2: **uORFs do not exhibit selective pressure to maintain amino acid identity.** (a) MAPS scores for single nucleotide variants within each ncORF category separated by predicted consequence (synonymous or missense) in each ORF. (i-iv) Allele frequencies for predicted missense SNVs are not significantly lower 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. (b) Translated uORF variants ranked by predicted change in codon optimality using codon stability coefficient (CSC) scores from SLAM-seq [35]. Grey dotted line denotes boundary separating optimality increasing versus decreasing SNVs. (c) 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.

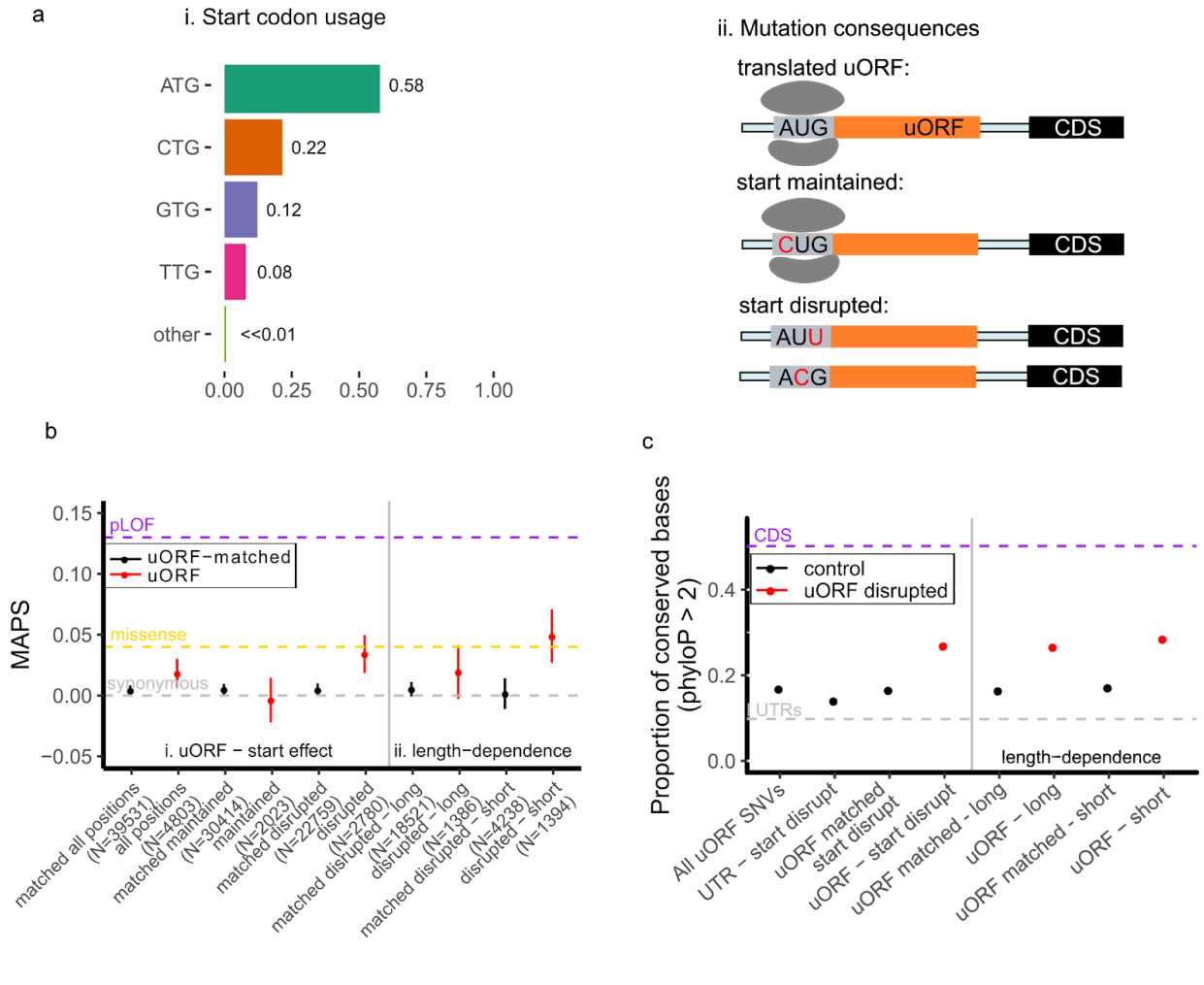

Figure 3: Selective pressure to preserve uORF start codons. (a) (i) Distribution of start codon usage for experimentally mapped translated uORFs, and (ii) possible consequences of mutations affecting uORF start codons. (b) (i) MAPS scores for start-disrupting SNVs are compared to uORF variants matched 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. (c) phyloP estimates for possible start codon disrupting positions in uORFs 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.

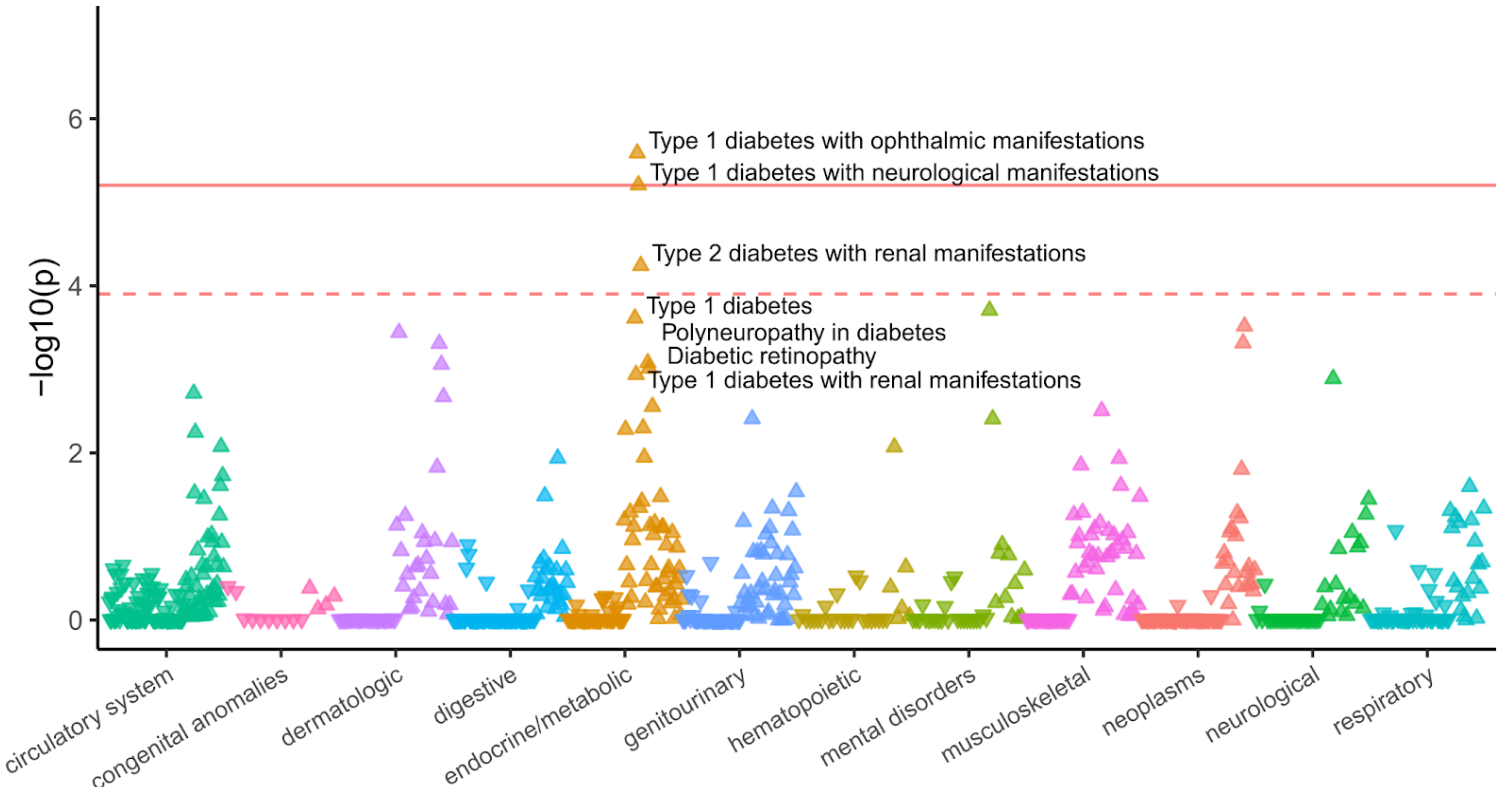

Figure 4: Phenome-wide association study (PheWAS) of predicted stop-strengthening variant in a translated uORF in PMVK. PheWAS plot of translated uORF stop-strengthening variant in the 5'UTR of PMVK (N = 65 carriers) in the Penn Medicine BioBank. ICD-9 and ICD-10 Phecodes are organized and plotted by 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).

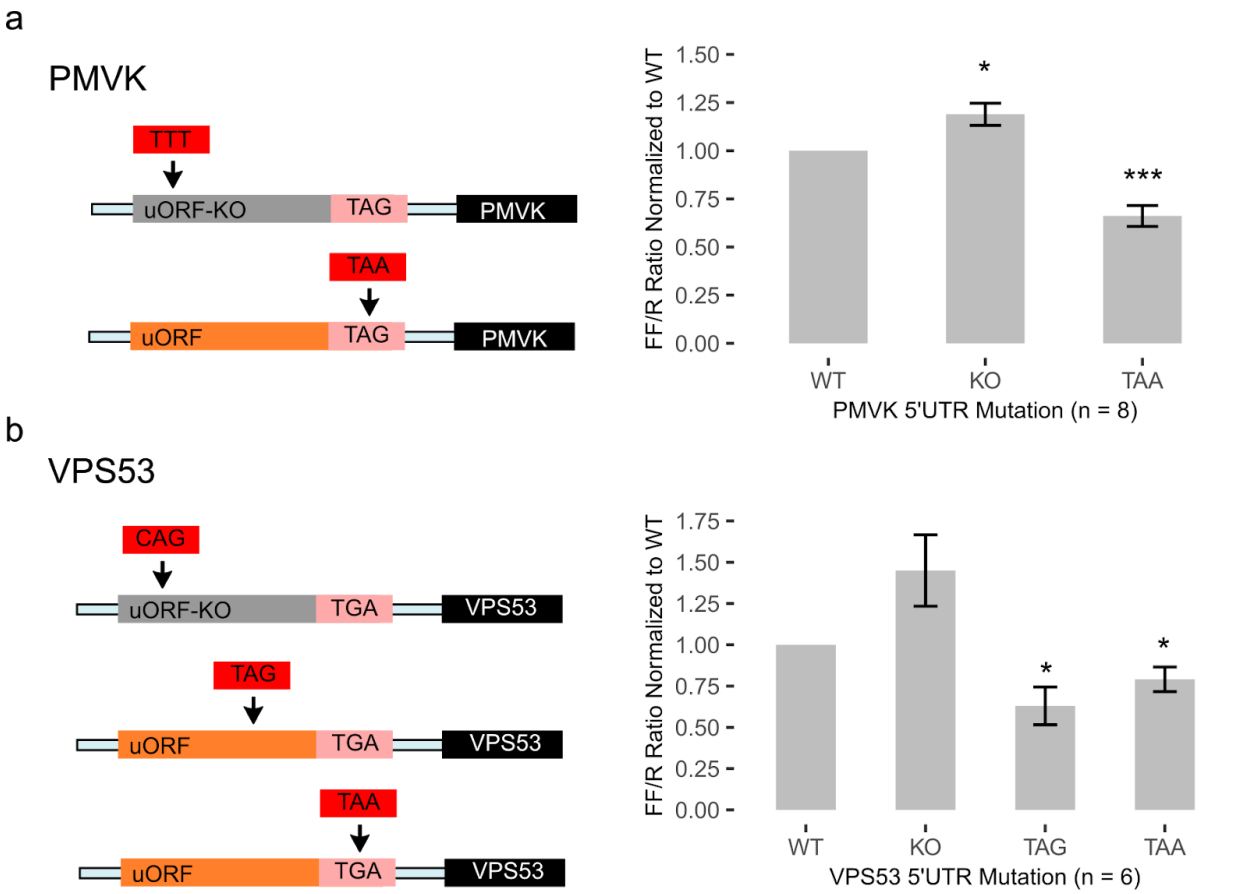

Figure 5: Reporter gene assays for translated uORF stop-introducing and stop-strengthening variants. Dual-luciferase reporter assay quantifies relative expression for uORFs with UTC and stop-strengthening variants that are associated with EHR phenotypes by PheWAS. Experimental 5'UTRs for (a) *PMVK* and (b) *VPS53* 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) and *** (>>0.001). Error bars represent mean + S.E.M. of at least 6 independent experiments.

Table 1: Significant Novel Associations in PheWAS of Penn Medicine BioBank

Variant properties			Novel associations**					Replication		
Gene	SNP	uORF effect	Phenotype (Phecode)	OR (SE)	P value	N cases	N controls	UKB	PMBB LOF	UKBB LOF
PMVK	rs181302437	Stop-strengthened (TAG>TAA)	250.13 (T1D with ophthalmic manifestations)	27.29 ± 0.70	2.58e-06	23	5189	Yes (diabetes mellitus P = 0.048)	Abnormal fasting blood glucose (P = 0.0234874)	Yes (250.13, P = 0.00727)
			250.14 (T1D with neurological manifestations)	22.71± 0.69	6.20e-06	25	5189			
			250.22 (T2D with renal manifestations)	7.79 ± 0.69	5.73e-05	136	5189			
VPS53	rs35915949	Stop-strengthened (TGA>TAA)	300.10 (Anxiety disorder)	0.64 ± 0.10	4.23e-06	1060	6939	No	No	No
			300.00 (Anxiety disorders)	0.69 ± 0.09	2.00 e-05	1249	6939	No	No	No
NALCN **	rs139848407	UTC (CAA>TAA)	270.33 (Amyloidosis)	38.92 ± 0.84	1.34e-05	30	7727	No (insufficient cases in UKB)	No	Yes (parent code 270, P=0.0264)
*BCL2L13***	rs140799351	Stop-strengthened (TGA>TAA)	610.00 (Benign mammary dysplasias)	270.57 ± 1.34	2.80e-05	55	7689	Yes (Other signs and symptoms in breast P = 1.79e-05, Malignant neoplasm of testis P = 2.09e-4)	No (not enough LOF variants)	No (not enough LOF variants)
			187.20 (Malignmant neoplasm of the testes)	331.41 ± 1.39	3.03e-05	26	7700			
			187.00 (Cancer of other male genital organs)	220.01 ± 1.35	6.31e-05	34	7700			
SHMT2 **	rs28365863	Stop-strengthened (TAG>TAA)	527.00 (Diseases of the salivary glands)	6.37 ± 0.46	5.27e-05	90	9774	No (insufficient cases in UKB)	Yes (527.00, P = 0.005515)	No (insufficient cases in UKB)
MOAP1 **	rs116450723	UTC (TAC>TAA)	350.00 (Abnormal movement)	4.99 ± 0.42	1.22e-04	362	9414	No (variant not present in UKB)	No	No (variant not present in UKB)

*T1D: type 1 diabetes, T2D: type 2 diabetes, UTC, upstream termination codon,

** Associations reaching FDR < 0.1.

Methods

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. [15]. 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.

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 [11].

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 [11]. 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 [10,11]. 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 [11]. 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 [3,10,11]. 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_ablatoin, 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 TAA-introducing variants we selected any variant that produced an in-frame TAA 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 [35]. 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 [3]. 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 [3].

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 (TGA, TAG, TAA) 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 [11]. For each iteration, one stop codon was randomly selected from each 5'UTR and the proportion of TGA, TAG, and TAA trinucleotides selected from all 5'UTR sequences were calculated. This procedure was repeated 10,000 times to form a distribution of TGA, TAG, and TAA 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 TAA stop codons used in translated uORFs were calculated by determining the number of bootstrap iterations where the frequency of TAA codons from uORFs was higher compared to non-uORF sequences. P-values for enrichment of TGA and TAG sequences were calculated by determining the fraction of sampled iterations where fewer TGA and TAG 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 TGA, TAG, or TAA stop codon. Some positions could be mutated to either a TAG or TAA codon and these were considered separately from potential TAG or TAA-creating positions. We have included all potential stop-introducing positions in **Suppl. File 1**.

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.

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 [36]. 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 34,629 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 [60] as 5' untranslated region (5' UTR), predicted loss-of-function (pLOF), or missense variants according to the NCBI Reference Sequence (RefSeq) database [60,61]. pLOF variants were defined as frameshift insertions/deletions, gain/loss of stop codon, or disruption of canonical splice site dinucleotides. Predicted deleterious missense variants were defined as those with Rare Exonic Variant Ensemble Learner (REVEL) [62] 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" [37,63]. 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 TAA UTCs, or strengthen existing uORF stop sites and carried by individuals in PMBB for the discovery experiment [37]. Each disease phenotype was tested for association with each uORF variant using a logistic regression model adjusted for age, age^2, 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 [64]. 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, age^2, 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, age^2, 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 2**. 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.

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	https://doi.org/10.7554/eLife.45396.006

	techniques	

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