Characterising the loss-of-function impact of 5' untranslated region variants in 15,708 individuals

Nicola Whiffin[1,2,3†], Konrad J Karczewski[3,4], Xiaolei Zhang[1,2], Sonia Chothani[5], Miriam J Smith[6], D Gareth Evans[6], Angharad M Roberts[1,2], Nicholas M Quaife[1,2], Sebastian Schafer[5,7], Owen Rackham[5], Jessica Alföldi[3,4], Anne H O'Donnell-Luria[3,4], Laurent C Francioli[3,4], Genome Aggregation Database (gnomAD) Production Team, Genome Aggregation Database (gnomAD) Consortium, Stuart A Cook[1,5,7], Paul J R Barton[1,2], Daniel G MacArthur[3,4]* and James S Ware[1,2,3]*

1. National Heart and Lung Institute and MRC London Institute of Medical Sciences, Imperial College London, London, UK
2. NIHR Royal Brompton Cardiovascular Research Centre, Royal Brompton and Harefield National Health Service Foundation Trust, London, UK
3. Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
4. Analytical and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA
5. Program in Cardiovascular and Metabolic Disorders, Duke-NUS Medical School, 8 College Road, Singapore
6. NW Genomic Laboratory Hub, Centre for Genomic Medicine, Division of Evolution and Genomic Science, University of Manchester, St Mary's Hospital, Manchester, UK
7. National Heart Centre Singapore, Singapore

†To whom correspondence should be addressed: n.whiffin@imperial.ac.uk

*These authors contributed equally

Abstract

Upstream open reading frames (uORFs) are important tissue-specific *cis*-regulators of protein translation. Although isolated case reports have shown that variants that create or disrupt uORFs can cause disease, genetic sequencing approaches typically focus on protein-coding regions and ignore these variants. Here, we describe a systematic genome-wide study of variants that create and disrupt human uORFs, and explore their role in human disease using 15,708 whole genome sequences collected by the Genome Aggregation Database (gnomAD) project. We show that 14,897 variants that create new start codons upstream of the canonical coding sequence (CDS), and 2,406 variants disrupting the stop site of existing uORFs, are under strong negative selection. Furthermore, variants creating uORFs that overlap the CDS show signals of selection equivalent to coding missense variants, and uORF-perturbing variants are under strong selection when arising upstream of known disease genes and genes intolerant to loss-of-function variants. Finally, we identify specific genes where perturbation of uORFs is likely to represent an important disease mechanism, and report a novel uORF frameshift variant upstream of *NF2* in families with neurofibromatosis. Our results highlight uORF-perturbing variants as an important and under-recognised functional class that can contribute to penetrant human disease, and demonstrate the power of large-scale population sequencing data to study the deleteriousness of specific classes of non-coding variants.

Introduction

Upstream open reading frames (uORFs) are ORFs encoded within the 5' untranslated regions (5'UTRs) of protein coding genes. uORFs are found upstream of around half of all known genes[1], and are important tissue-specific *cis*-regulators of translation. Active translation of a uORF typically reduces downstream protein levels by up to 80%[1]. There are strong signatures of negative selection acting on these elements, with fewer upstream start codons (uAUGs) present in the human genome than would be expected by chance[1–3]. In addition, the start codons of uORFs have been shown to be the most conserved sites in 5'UTRs[1], supporting the importance of uORFs in the regulation of protein levels.

In humans, translation is initiated when the small ribosomal subunit, which scans from the 5' end of the mRNA, recognises an AUG start codon[4]. The likelihood of an AUG initiating translation is dependent on local sequence context, and in particular the degree of similarity to the Kozak consensus sequence[5,6]. uORFs can inhibit translation through multiple mechanisms. For some genes, uORFs may be translated into a small peptide which can directly inhibit translation by interacting with and stalling the elongating ribosome at or near the uORF stop codon, creating a 'roadblock' for other scanning ribosomes[7,8]. It is also possible for this small peptide to have a distinct biological function[9]; however, in general uORFs do not show strong evidence for conservation of their amino acid sequence[2,10]. For other genes, translation from a uAUG appears to be sufficient to inhibit translation of the downstream protein, with the small uORF peptide only produced as a by-product.

Mechanisms of leaky scanning (whereby a scanning ribosome may bypass an uAUG), re-initiation (where the small ribosomal subunit remains bound to the mRNA and translation is re-initiated at the canonical AUG), and the existence of internal ribosome entry sites (from which the ribosome can start scanning part-way along the RNA), can all act to attenuate inhibition by uORFs, adding to the complexity of translational regulation[10–12]. Termination at a uORF stop codon can also trigger the nonsense-mediated decay pathway, further magnifying the inhibitory effects of uORFs[11,13]. To date, studies of translational regulation by individual uORFs have mainly been restricted to model organisms.

Recently, large scale studies have assessed the global translational repression ability of uORFs: in vertebrates, uORF-containing transcripts are globally less efficiently translated than mRNAs lacking uORFs, with this effect mediated by features of both sequence and structure[2]. Similarly, polysome profiling of 300,000 synthetic 5'UTRs identified uORFs and uAUGs as strongly repressive of translation, with the strength of repression dependent on the surrounding Kozak consensus sequence[14].

Although 5'UTRs are typically not assessed for variation in either clinical or research settings, having been excluded from most exome capture target regions, there are several documented examples of variants that create or disrupt a uORF playing a role in human disease[1,15–21]. These studies have focused on single gene disorders or candidate gene lists, often when no causal variant was identified in the coding sequence. No study to date has characterised the baseline population incidence of these variants.

Here we describe a systematic genome-wide study of variants that create and disrupt human uORFs, and characterise the contribution of this class of variation to human genetic disease. We use the allele frequency spectrum of variants in 15,708 whole-genome sequenced individuals from the Genome Aggregation Database (gnomAD)[22] to explore selection against variants that either create uAUGs or remove the stop codon of existing uORFs. Finally, we demonstrate that these variants make an under-recognised contribution to genetic disease.

Results

uAUG-creating variants are under strong negative selection

To estimate the deleteriousness of variants that create a novel AUG start codon upstream of the canonical coding sequence (CDS), we assessed the frequency spectrum of uAUG-creating variants observed in gnomAD (Figure 1a). We identified all possible single nucleotide variants (SNVs) in the UTRs of 18,593 canonical gene transcripts (see Methods) that would create a new uAUG, yielding 562,196 possible SNVs, an average of 30.2 per gene (Figure 1b). Of these, 15,239 (2.7%) were observed at least once in whole genome sequence data from 15,708 individuals in gnomAD (Supplementary Figure 1a), upstream of 7,697 distinct genes.

We compared the mutability adjusted proportion of singletons (MAPS) score, a measure of the strength of selection acting against a variant class[23], for uAUG-creating SNVs to other classes of coding and non-coding SNVs (see methods). As negative selection acts to prevent deleterious variants from increasing in frequency, damaging classes of variants have skewed frequency spectra, with a higher proportion appearing as singletons (i.e. observed only once in the gnomAD data set),[23] reflected in a higher MAPS score. Whilst all observed UTR SNVs have an overall MAPS score almost identical to synonymous variants, uAUG-creating SNVs have a significantly higher MAPS score (permuted $P<1\text{x}10^{-4}$; Figure 1c), indicating a considerable selective pressure acting to remove these from the population.

We next evaluated subsets of uAUG-creating variants predicted to have distinct functional consequences. In addition to creating distinct uORFs, uAUGs may result in overlapping ORFs (oORFs) where the absence of an in-frame stop codon within the UTR results in an ORF that reads into the coding sequence, either in-frame (elongating the CDS), or out-of-frame (Figure 1a). uAUG-creating variants that form oORFs have a significantly higher MAPS score than uORF-creating variants (permuted $P<1\text{x}10^{-4}$), and equivalent to missense variants in coding regions (Figure 1d; Supplementary Figure 1a).

We also investigated the context of uAUG-creating variants and find that uAUGs created within 50 bp of the CDS have higher MAPS than those created further away (permuted $P=0.0042$), although this may be driven by the higher propensity of these variants to form oORFs. We did not observe a significantly greater MAPS score for uAUG-creating variants arising on a background of a strong Kozak consensus, though we observe a trend in this direction (Figure 1e).

Given that uAUGs are expected to dramatically decrease downstream protein levels, we hypothesised that uAUG-creating variants would behave similarly to pLoF variants and thus be more deleterious when arising upstream of genes intolerant to LoF variation. Indeed, we show a significantly higher MAPS score for uAUG-creating SNVs upstream of genes which are most intolerant to LoF variants (top sextile of LOEUF score[22]; 3,193 genes) when compared to those that are most tolerant (bottom sextile; permuted $P<1\text{x}10^{-4}$; Figure 1c).

Next, we calculated MAPS for uAUG-creating variants arising upstream of 1,659 genes known to cause developmental disorders (DD; confirmed or probable genes from the Developmental Disease Gene to Phenotype (DDG2P) database). While uAUG-creating variants upstream of all DD genes do not show a signal of selection above all observed uAUG-creating variants, the MAPS score is significantly inflated when limiting to 279 DD genes with a known dominant LoF mechanism (permuted $P=0.0012$; Figure 1f).

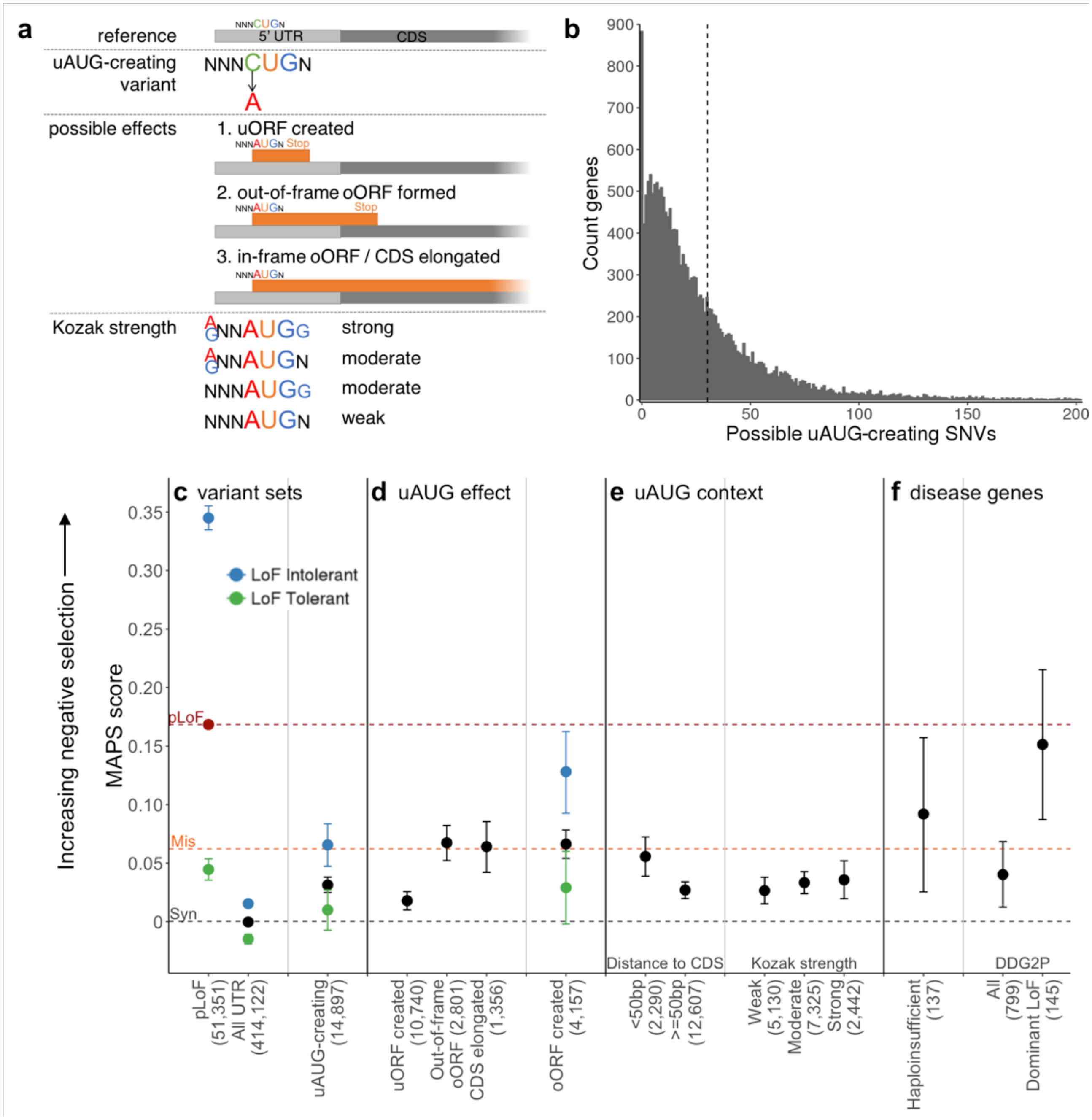


Figure 1: uAUG-creating variants have strong signals of negative selection, suggesting they are deleterious. (a) Schematic of uAUG-creating variants, their possible effects and how the strength of the surrounding Kozak consensus is determined. (b) The number of possible uAUG-creating SNVs in each of 18,593 genes, truncated at 200 (159 genes have >200). In total we identified 562,196 possible uAUG-creating SNVs, an average of 30.2 per gene (dotted line), with 883 genes having none. (c-f) MAPS scores (a measure of negative selection) for different variant sets. The number of observed variants for each set is shown in brackets. MAPS for classes of protein-coding SNVs are shown as dotted lines for comparison (synonymous - grey, missense - orange, and predicted loss-of-function (pLoF) - red point and red dotted line) (c) While overall UTR variants display a selection signature similar to synonymous variants, uAUG-creating variants have significantly higher MAPS (indicative of being more deleterious; permuted

$P<1\text{x}10^{-4}$). Variants are further subdivided into those upstream of, or within genes tolerant (green dot) and intolerant (blue dot) to LoF[22], with uAUG-creating variants upstream of LoF intolerant genes showing significantly stronger signals of selection than those upstream of LoF tolerant genes (permuted $P=5\text{x}10^{-4}$). pLoF variants are likewise stratified for comparison. (d) uAUG-creating variants that create an oORF or elongate the CDS show a significantly higher signal of selection that uORF-creating variants ($P<1\text{x}10^{-4}$; oORF created - out-of-frame oORF and CDS elongated combined). (e) The deleteriousness of uAUG-creating variants depends on the context into which they are created, with stronger selection against uAUG-creation close to the CDS, and with a stronger Kozak consensus sequence. (f) uAUG-creating variants are under strong negative selection upstream of genes manually curated as haploinsufficient[26] and developmental disorder genes reported to act via a dominant LoF mechanism. Abbreviations: CDS - coding sequence; uAUG - upstream AUG; uORF - upstream open reading frame; oORF - overlapping open reading frame; MAPS - mutability adjusted proportion of singletons; pLoF - predicted loss-of-function; DDG2P - Developmental Disease Gene to Phenotype.

Variants that disrupt stop codons of existing uORFs also show a signal of strong selection

As uAUG-creating variants that form oORFs have a significantly higher MAPS score than those with an in-frame UTR stop codon, we hypothesised that variants that disrupt the stop site of existing uORFs should also be under selection (Figure 2a). These stop-removing variants could either be SNVs that change the termination codon to one that codes for an amino acid, or frameshifting indels within the uORF sequence that cause the uORF to read through the normal stop codon. If there is no other in-frame stop codon before the CDS will result in an oORF.

We identified all possible SNVs that would remove the stop codon of a predicted uORF (n=169,206; see methods), and calculated the MAPS score for 2,406 such variants observed in gnomAD. Stop-removing SNVs have a nominally higher MAPS score than all UTR SNVs (permuted $P=0.030$). This difference is greater when specifically considering stop-removing SNVs which are upstream of LoF intolerant genes (permuted $P=0.0012$), result in an oORF (permuted $P=2\text{x}10^{-4}$), or where the uORF has either prior evidence of translation (documented in sorfs.org[24]; permuted $P=0.0049$), or a strong/moderate Kozak consensus (permuted $P=7\text{x}10^{-4}$; Figure 2b-e).

As the power of MAPS is limited by the small number of stop-removing variants in each category observed in gnomAD, we performed a complementary analysis investigating base level conservation at all uORF stop sites using PhyloP[25]. A significantly greater proportion of uORF stop site bases have PhyloP scores >2 (12.2%) compared to UTR bases matched by gene and distance from the CDS (10.8%; Fisher's $P=1.8\text{x}10^{-17}$; Figure 2f). This proportion is significantly higher where there is evidence supporting translation of the uORF (18.9%; Fisher's $P=3.6\text{x}10^{-83}$) or when removing the stop would result in an oORF (either in-frame or out-of-frame; 17.2% and 17.4% respectively; Fisher's $P=3.0\text{x}10^{-25}$ and $2.6\text{x}10^{-47}$ respectively). Furthermore, a greater proportion of stop site bases have PhyloP scores >2 when the uORF start codon has a strong or moderate Kozak when compared to a weak Kozak consensus (12.7% vs 10.9%; Fisher's $P=5.5\text{x}10^{-10}$; matched UTR bases $P=0.88$; Figure 2i).

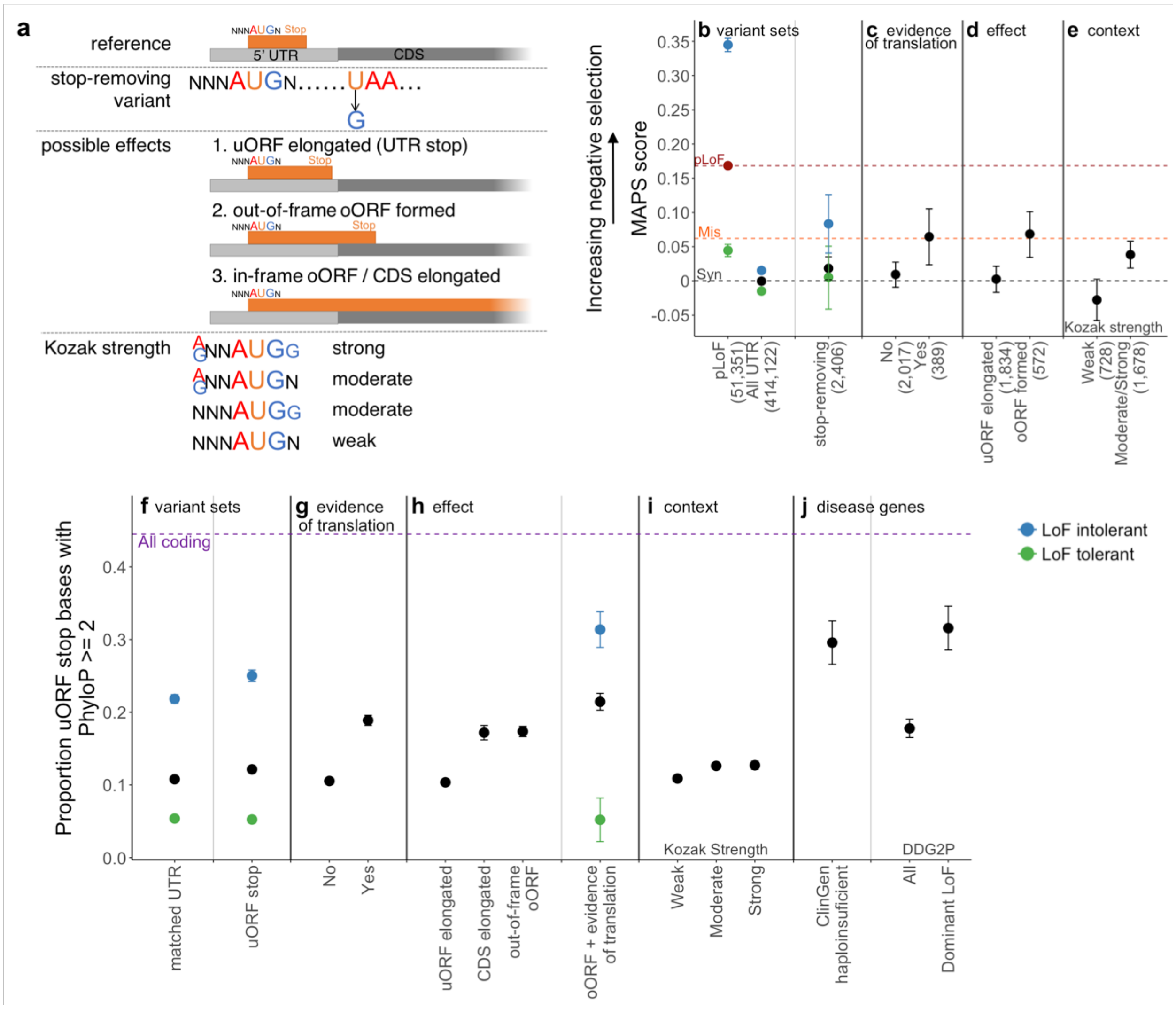


Figure 2: uORF stop codons are highly conserved and stop-removing variants show strong signals of negative selection. (a) Schematic of uORF stop-removing variants, their possible effects, and how the strength of the surrounding Kozak consensus is determined. (b-e) MAPS scores (a measure of negative selection, as in Fig. 1c) for different variant sets. The number of observed variants for each set is shown in brackets. MAPS for classes of protein-coding SNVs are shown as dotted lines for comparison (synonymous - black, missense - orange and predicted loss-of-function (pLoF) - red point and red dotted line). (b) Stop-removing SNVs have a nominally higher MAPS score than all UTR SNVs (permuted P=0.030). Variants are further subdivided into those upstream of, or within genes tolerant (green dot) and intolerant (blue dot) to LoF[22], with pLoF variants likewise stratified for comparison. Stop-removing SNVs with (c) evidence of translation (documented in sorfs.org) and (d) that create an oORF have signals of selection equivalent to missense variants. (e) A significantly higher MAPS is calculated for stop-removing variants where the uORF start site has a strong or moderate Kozak consensus when compared to those with a weak Kozak (permuted P=7x10^{-4}). (f-j) Since MAPS is only calculated on observed variants, we extended our analysis to look at the conservation of all possible uORF stop site bases, reporting the proportion of bases with phyloP scores > 2. All

coding bases are shown as a purple dotted line for comparison. (f) The stop sites of predicted uORFs are significantly more conserved than all UTR bases matched on gene and distance from the CDS (Fisher's P=1.8x10^{-17}). uORF stop bases are most highly conserved when (g) the uORF has evidence of translation, (h) the variant results in an oORF, (i) the uORF start site has a strong/moderate Kozak consensus, and (j) upstream of curated haploinsufficient genes and developmental genes with a known dominant LoF disease mechanism. Abbreviations: CDS - coding sequence; uORF - upstream open reading frame; oORF - overlapping open reading frame; MAPS - mutability adjusted proportion of singletons; pLoF - predicted loss-of-function; DDG2P - Developmental Disease Gene to Phenotype.

The increased power of this analysis enables us to convincingly demonstrate that uORF stop sites upstream of (1) LoF intolerant genes, (2) genes manually curated as haploinsufficient[26], and (3) developmental disorder genes with a dominant LoF mechanism, are all highly conserved. Stop sites upstream of genes in these groups have 21.9%, 29.6% and 31.6% of bases with PhyloP >2, respectively (Fisher's P=8.2x10^{-250}, 4.7x10^{-43} and 1.4x10^{-52} compared to all stop site bases, respectively; Figure 2j), suggesting that removing these stop sites is likely to be deleterious.

Disease-causing variants represent the most deleterious uORF variant types and highlight specific genes that are sensitive to uAUG-creating and uORF stop-removing variants

We searched the Human Gene Mutation Database (HGMD)[27] and ClinVar[28] for uORF-creating or -disrupting variants, identifying 39 uAUG-creating and four stop-removing (likely) pathogenic/disease mutations in 37 different genes. All four stop-removing variants disrupt uORFs with uAUGs in a strong or moderate Kozak consensus and result in an oORF overlapping the CDS (Supplementary Table 2). Compared to all possible uAUG-creating variants in these 37 genes, the 39 reported disease-causing uAUG-creating variants (Supplementary Table 1) are significantly more likely to be created into a moderate or strong Kozak consensus (binomial P=3.5x10^{-4}), create an out-of-frame oORF (binomial P=1.1x10^{-5}), and be within 50bp of the CDS (binomial P=3.9x10^{-7}; Figure 3a). These results further illustrate the power of MAPS to identify variant classes most likely to be disease-causing.

This analysis highlights disease genes where aberrant translational regulation through uORFs is an important disease mechanism. Previous analysis of the *NF1* gene in 361 patients with neurofibromatosis identified four 5'UTR variants as putatively disease-causing[29]. While uAUG creation was proposed as the mechanism behind two of these variants, we now show that the other two variants both disrupt the stop codon of an existing uORF, resulting in an oORF which is out-of-frame with the CDS. This existing uORF has two start sites, both with strong Kozak consensus, and has prior evidence of active translation[24]. In figure 3b, we show these four variants along with an additional six stop-removing and ten uAUG-creating variants that would be predicted to also cause neurofibromatosis through the same mechanism if observed. In addition to these sixteen SNVs, indels that create high-impact uAUGs (oORF creating with strong/moderate Kozak consensus) or that cause a frameshift within the sequence of the existing uORF, resulting in an oORF, would also be predicted to cause disease.

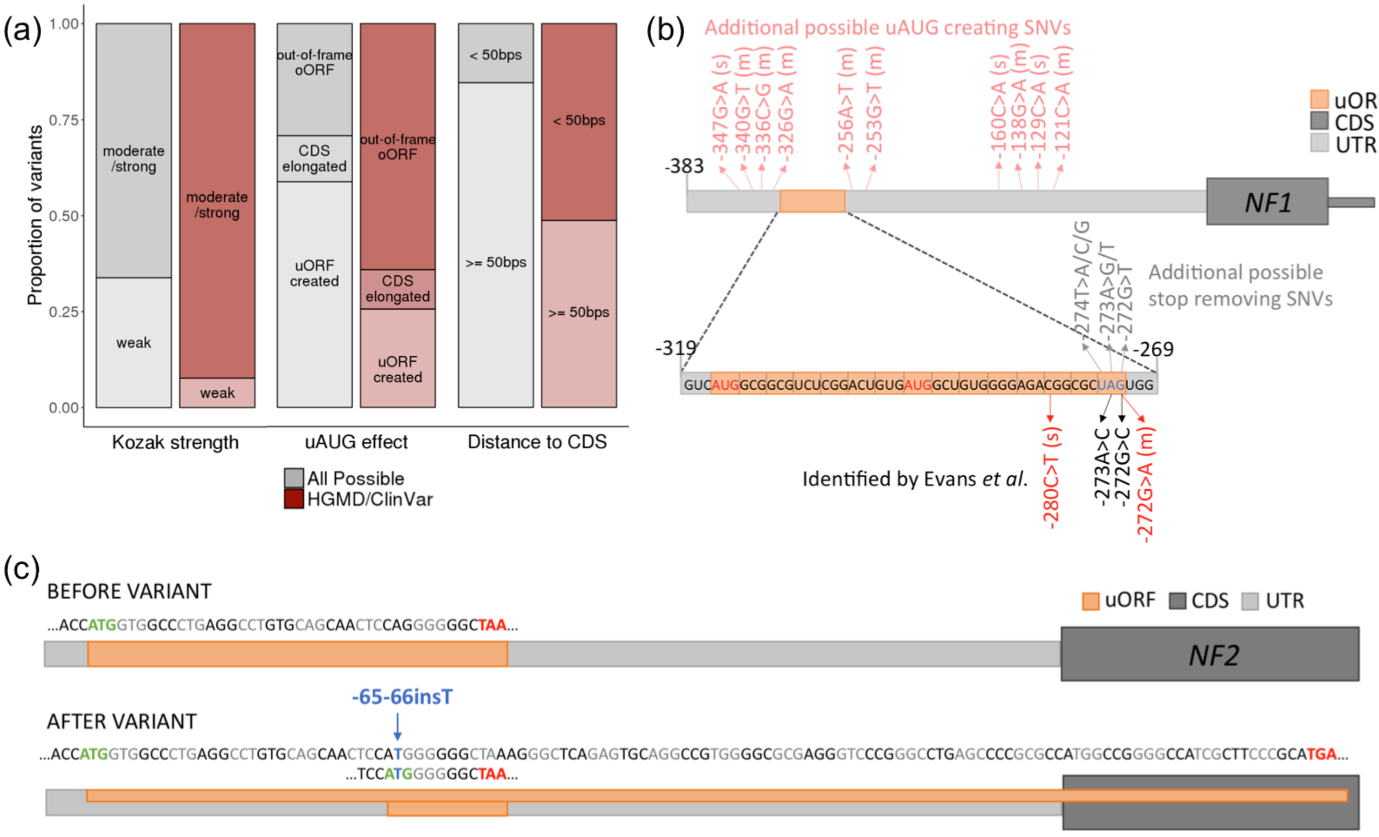


Figure 3: The role of uAUG-creating and uORF stop-removing variants in disease. (a) The proportion of 39 uAUG variants observed in HGMD and ClinVar (red bars) that fit into different sub-categories compared to all possible uAUG-creating SNVs (grey bars) in the same genes (n=1,022). Compared to all possible uAUG-creating variants, uAUG-creating variants observed in HGMD/ClinVar were significantly more likely to be created into a moderate or strong Kozak consensus (binomial P=3.5x10^{-4}), create an out-of-frame oORF (binomial P=1.1x10^{-5}), and be within 50 bp of the CDS (binomial P=3.9x10^{-7}). (b) Schematic of the *NF1* 5'UTR (light grey) showing the location of an existing uORF (orange) and the location of variants previously identified in patients with neurofibromatosis[29] in dark red (uAUG-creating) and black (stop-removing). uAUG-creating variants are annotated with the strength of the surrounding Kozak consensus in brackets ("s" for strong and "m" for moderate). All four published variants result in formation of an oORF out-of-frame with the CDS. Also annotated are the positions of all other possible uAUG-creating variants (light red; strong and moderate Kozak only), and stop-removing variants (grey) that would also create an out-of-frame oORF. (c) Schematic of the *NF2* 5'UTR (grey) showing the effects of the -65-66insT variant. The reference 5'UTR contains a uORF with a strong Kozak start site. Although the single-base insertion creates a novel uAUG which could be a new uORF start site, it also changes the frame of the existing uORF, so that it overlaps the CDS out-of-frame (forms an oORF). We predict this is the most likely mechanism of pathogenicity.

A second example is *IRF6*, where three uAUG-creating variants have been identified in seven patients with Van de Woude syndrome[30,31]. These variants all arise in the context of a strong or moderate Kozak consensus and result in an out-of-frame oORF. There are nine additional possible uAUG-creating variants that would be predicted to yield the same effect in *IRF6* (Supplementary Figure 2), suggesting it would be prudent to screen Van de Woude patients across all twelve sites.

Identifying genes where perturbing uORFs is likely to be important in disease

To guide the research and clinical identification of uAUG-creating and stop-removing variants (referred to collectively as uORF-perturbing variants), we set about identifying genes where these variants are likely to be of high importance. Investigating 17,715 genes with annotated 5'UTRs and at least one possible uORF-perturbing variant, we first identified 4,986 genes where uORF-perturbing variants are unlikely to be deleterious: genes with existing oORFs (strong/moderate Kozak or evidence of translation), with predicted high-impact uORF-perturbing SNVs of appreciable frequency in gnomAD (>0.1%), with no possible high-impact uORF-perturbing SNVs, or that are tolerant to LoF (see methods; Supplementary Figure 3a). Interestingly, these genes include 453 LoF intolerant (14.2% of most constrained LOEUF sextile) and 163 curated haploinsufficient or LoF disease genes (14.6%). Of the remaining 12,729 genes considered, 3,191 (25.1%) are LoF-intolerant, known haploinsufficient or LoF disease genes and hence are genes where uORF-perturbing variants have a high likelihood of being deleterious (Figure 4a). Despite only 18.0% of all classified genes falling into this high likelihood category (19.0% of all UTR bases when accounting for UTR length), 79% of uORF-perturbing variants in HGMD and ClinVar are found upstream of these genes (Fisher's P=1.6x10^{-9}; Figure 4b).

There are 296 genes that have at least 10 possible high-impact uORF-perturbing SNVs, and for which LoF and/or haploinsufficiency is a known mechanism of human disease (either curated as haploinsufficient, curated as acting via a LoF mechanism in DDG2P or with ≥10 high-confidence pathogenic LoF variants documented in ClinVar), including both *IRF6* and *NF1.* We predict these to be a fruitful set to search for additional disease-causing uORF-perturbing variants (Supplementary Table 3; Supplementary Figure 3b). To aid in the identification of uORF-perturbing variants we have created plugin for the Ensembl Variant Effect Predictor (VEP)[32] which annotates variants for predicted effects on translational regulation (available at https://github.com/ImperialCardioGenetics/uORFs).

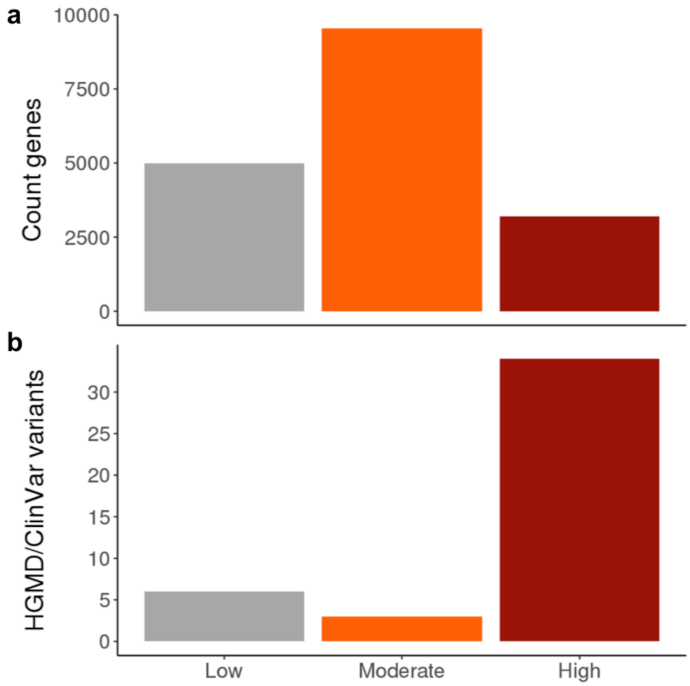


Figure 4: Identifying genes where uORF creating or disrupting variants are are likely to have a role in disease. Genes were split into three distinct categories representing a 'low', 'moderate' and 'high' likelihood that uORF-perturbing variants are important. Low likelihood genes include those with existing oORFs, common (>0.1%) oORF creating variants in gnomAD or that are tolerant to LoF. Those in the high likelihood category are remaining genes that are LoF-intolerant or where haploinsufficient or LoF is a known disease mechanism (see methods). (a) The number of genes in each of the three categories. (b) The number of uAUG-creating and uORF stop-removing variants in HGMD upstream of genes in each category. Although only 18.0% of all classified genes fall into the high likelihood category (19.0% of all UTR bases when adjusting for UTR length), 79% of uORF-perturbing variants identified in HGMD and ClinVar are found upstream of these genes (Fisher's P=1.6x10^{-9}).

A novel uORF frameshift variant as a cause of neurofibromatosis type 2

We analysed targeted sequencing data from a cohort of 1,134 unrelated individuals diagnosed with neurofibromatosis type 2, which is caused by LoF variants in one of these prioritised genes, *NF2*. We identified a single 5'UTR variant in two unrelated probands in this cohort (ENST00000338641:-66-65insT; GRCh37:chr22:29999922 A>AT) that segregates with disease in three additional affected relatives across the two families. This variant could act through two distinct uORF-disrupting mechanisms. While the insertion does create a new uAUG (in the context of a moderate Kozak consensus) an in-frame stop codon after only three codons would suggest only a weak effect on CDS translation. However, the *NF2* UTR contains an existing uORF with prior evidence of translation[24] and a strong Kozak consensus. The observed insertion changes the frame of this existing uORF, causing it to bypass the downstream stop codon and create an out-of-frame oORF (Figure 3c). This oORF is predicted to lower translation of *NF2*, consistent with the known LoF disease mechanism, however, functional follow-up is required to confirm this hypothesis.

Discussion

We used data from 15,708 whole human genomes to explore the global impact of variants that create or perturb uORFs in 5'UTRs, which can lead to altered translation of the downstream protein. We show that creating a new uORF and hence initiating translation from an uAUG is an important regulatory mechanism. Our data suggest that the major underlying mechanism of translational repression by uORFs is likely to be through competitive translation, since it is unlikely that novel peptides produced by uAUG-creating variants will be functional, and the most deleterious types of uAUG-creating and stop-removing variants are those that form oORFs.

Selective pressure on strongly translated uORFs has maintained features that promote re-initiation and prevent constitutive translational repression. Specifically, existing uORFs are selected to be short, further from the CDS, and to lack strong Kozak sequences[2]. This is in agreement with our results, which show a strongly skewed frequency spectrum for observed variants predicted to strongly inhibit translation, and an over-representation of these deleterious variants in disease cases.

We have defined a new category of variants, high-impact uORF-perturbing variants, a subset of which are likely to act as LoF by severely impacting translation. This class contains 145,398 possible SNVs (110,357 uAUG-creating and 35,041 stop-removing) across the genome, which are predicted to form oORFs from an uAUG with a strong or moderate Kozak consensus, or with prior evidence of translation. Of these, 3,213 (2.2%) are observed in the whole genome sequence data from gnomAD. In addition, uAUG-creating insertions and deletions or frameshifts that transform existing uORFs into oORFs would also be predicted to have a high impact.

Whilst uORF-perturbing variants resulting in constitutive translational repression are likely to have LoF effects, the complex mechanisms of translational regulation including leaky scanning, re-initiation, and the existence of internal ribosome entry sites makes it difficult to confidently predict the functional consequences of individual variants. Even variants predicted to be of high-impact may only result in partial LoF, reducing power to identify significant signals of selection. Confident interpretation of variants for a role in disease will require functional studies to assess the downstream impact of these variants on protein levels and/or additional genetic evidence, such as *de novo* occurrence or segregation with disease. It will also be interesting to study the impact of uORF-perturbing variants causing partial LoF on coding variant penetrance and their role in common disease phenotypes.

Even at a sample size of 15,708 individuals, we had limited power to observe uORF-perturbing variants, given their very small genomic footprint. Despite this, we identified specific genes such as *NF1*, *NF2*, and *IRF6*, where uORF perturbation appears to be an important disease mechanism. In anticipation of future studies with much larger cohorts of WGS cases, we have identified a set of genes where there is a high likelihood that this mechanism will contribute to disease. This will also be useful for rare disease diagnosis, where even if WGS is undertaken this class of pathogenic variation is likely not evaluated and under-diagnosed.

In this work, we used variant frequencies in a large population dataset to study the global impact of a specific class of non-coding variants with a predicted functional effect. Previous studies using non-coding constraint have focused on entire regulatory regions[33] or concentrated exclusively on splicing[34,35]. These and other studies[36] have concluded that signals of constraint and selection are likely confined to individual bases[33] and diluted out when studying larger regions. Our results support this assertion; as the signal of negative selection associated with all UTR variants is not discernible from synonymous variants. We show the power of grouping individual non-coding bases by functional effect to identify subsets of variants with strong signals of selection.

Acknowledgements

N.W. is supported by a Rosetrees and Stoneygate Imperial College Research Fellowship. This work was supported by the Wellcome Trust (107469/Z/15/Z), the Medical Research Council (UK), the NIHR Biomedical Research Unit in Cardiovascular Disease at Royal Brompton & Harefield NHS Foundation Trust and Imperial College London, the NIHR Imperial College Biomedical Research Centre, the Fondation Leducq (11 CVD-01), a Health Innovation Challenge Fund award from the Wellcome Trust and Department of Health, UK (HICF-R6–373), and by NIDDK U54DK105566 and NIGMS R01GM104371. DGE and MJS are funded by the NIHR Biomedical research centre Manchester (IS-BRC-1215-20007). LCF is supported by the Swiss National Science Foundation (Advanced Postdoc.Mobility 177853). NQ is supported by the Imperial College Academic Health Science Centre. The results published here are in part based upon data: 1) generated by The Cancer Genome Atlas managed by the NCI and NHGRI (accession: phs000178.v10.p8). Information about TCGA can be found at http://cancergenome.nih.gov, 2) generated by the Genotype-Tissue Expression Project (GTEx) managed by the NIH Common Fund and NHGRI (accession: phs000424.v7.p2), 3) generated by the Exome Sequencing Project, managed by NHLBI, 4) generated by the Alzheimer's Disease Sequencing Project (ADSP), managed by the NIA and NHGRI (accession: phs000572.v7.p4). The views expressed in this work are those of the authors and not necessarily those of any of the funders. We would like to thank Dr Christopher M Yates for his statistical advice.

References

1. 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).
2. Johnstone, T. G., Bazzini, A. A. & Giraldez, A. J. Upstream ORFs are prevalent translational repressors in vertebrates. *EMBO J.* **35**, 706–723 (2016).
3. Iacono, M., Mignone, F. & Pesole, G. uAUG and uORFs in human and rodent 5'untranslated mRNAs. *Gene* **349**, 97–105 (2005).
4. Kozak, M. Pushing the limits of the scanning mechanism for initiation of translation. *Gene* **299**, 1–34 (2002).
5. Kozak, M. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.* **15**, 8125–8148 (1987).
6. Noderer, W. L. *et al.* Quantitative analysis of mammalian translation initiation sites by FACS□seq. *Mol. Syst. Biol.* **10**, 748 (2014).
7. Hinnebusch, A. G., Ivanov, I. P. & Sonenberg, N. Translational control by 5'-untranslated regions of eukaryotic mRNAs. *Science* **352**, 1413–1416 (2016).
8. Jousse, C. *et al.* Inhibition of CHOP translation by a peptide encoded by an open reading frame localized in the chop 5′ UTR. *Nucleic Acids Res.* **29**, 4341–4351 (2001).
9. 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* **4**, e08890 (2015).
10. Couso, J.-P. & Patraquim, P. Classification and function of small open reading frames. *Nat. Rev. Mol. Cell Biol.* **18**, 575–589 (2017).
11. Wethmar, K. The regulatory potential of upstream open reading frames in eukaryotic gene expression. *Wiley Interdiscip. Rev. RNA* **5**, 765–778 (2014).
12. Wang, X. & Rothnagel, J. A. 5′□Untranslated regions with multiple upstream AUG codons

can support low□level translation via leaky scanning and reinitiation. *Nucleic Acids Res.* **32**, 1382–1391 (2004).

13. Chugunova, A., Navalayeu, T., Dontsova, O. & Sergiev, P. Mining for Small Translated ORFs. *J. Proteome Res.* **17**, 1–11 (2018).
14. Sample, P. J. *et al.* Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. *bioRxiv* 310375 (2018). doi:10.1101/310375
15. Schulz, J. *et al.* Loss-of-function uORF mutations in human malignancies. *Sci. Rep.* **8**, 2395 (2018).
16. Liu, L. *et al.* Mutation of the CDKN2A 5' UTR creates an aberrant initiation codon and predisposes to melanoma. *Nat. Genet.* **21**, 128–132 (1999).
17. Wen, Y. *et al.* Loss-of-function mutations of an inhibitory upstream ORF in the human hairless transcript cause Marie Unna hereditary hypotrichosis. *Nat. Genet.* **41**, 228–233 (2009).
18. Occhi, G. *et al.* A novel mutation in the upstream open reading frame of the CDKN1B gene causes a MEN4 phenotype. *PLoS Genet.* **9**, e1003350 (2013).
19. Barbosa, C., Peixeiro, I. & Romão, L. Gene expression regulation by upstream open reading frames and human disease. *PLoS Genet.* **9**, e1003529 (2013).
20. Matthes, T. *et al.* Severe hemochromatosis in a Portuguese family associated with a new mutation in the 5'-UTR of the HAMP gene. *Blood* **104**, 2181–2183 (2004).
21. von Bohlen, A. E. *et al.* A mutation creating an upstream initiation codon in the SOX9 5' UTR causes acampomelic campomelic dysplasia. *Mol Genet Genomic Med* **5**, 261–268 (2017).
22. 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
23. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**,

285–291 (2016).

24. Olexiouk, V., Crappé, J. & Verbruggen, S. sORFs. org: a repository of small ORFs identified by ribosome profiling. *Nucleic acids* (2015).

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. Riggs, E. R. *et al.* Copy number variant discrepancy resolution using the ClinGen dosage sensitivity map results in updated clinical interpretations in ClinVar. *Hum. Mutat.* **39**, 1650–1659 (2018).

27. Stenson, P. D. *et al.* The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalised genomics. *Hum. Genomics* **4**, 69 (2009).

28. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–5 (2014).

29. Evans, D. G. *et al.* Comprehensive RNA Analysis of the NF1 Gene in Classically Affected NF1 Affected Individuals Meeting NIH Criteria has High Sensitivity and Mutation Negative Testing is Reassuring in Isolated Cases With Pigmentary Features Only. *EBioMedicine* **7**, 212–220 (2016).

30. de Lima, R. L. L. F. *et al.* Prevalence and nonrandom distribution of exonic mutations in interferon regulatory factor 6 in 307 families with Van der Woude syndrome and 37 families with popliteal pterygium syndrome. *Genet. Med.* **11**, 241–247 (2009).

31. Kondo, S. *et al.* Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. *Nat. Genet.* **32**, 285–289 (2002).

32. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

33. Short, P. J. *et al.* De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* **555**, 611–616 (2018).

34. Zhang, S. *et al.* Base-specific mutational intolerance near splice sites clarifies the role of

nonessential splice nucleotides. *Genome Res.* **28**, 968–974 (2018).

35. doi:10.1101/256636

36. An, J.-Y. *et al.* Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* **362**, (2018).

Supplementary Table 1: uAUG-creating variants catalogued as disease mutations (DM) in HGMD[1] or (Likely) Pathogenic in ClinVar[2]. All positions are in GRCh37.

variant	Kozak strength	effect	gene	gene class	distance to start	clinvar	PubMed IDs
1-43424429-C-T	Strong	out-of-frame oORF	SLC2A1	8 (High)	107		28378819
1-209975332-G-T	Moderate	out-of-frame oORF	IRF6	8 (High)	19		19282774
1-209975361-T-A	Moderate	out-of-frame oORF	IRF6	8 (High)	49		12219090
1-209979367-C-T	Strong	out-of-frame oORF	IRF6	8 (High)	151		19282774
5-14871567-G-A	Moderate	CDS elongated	ANKH	8 (High)	12		12297987
5-36877266-C-T	Weak	uORF created	NIPBL	8 (High)	95		17661813
6-137143759-C-T	Strong	out-of-frame oORF	PEX7	8 (High)	46	Phytanic_acid_storage_disease;Pathogenic;38871	12325024
7-107301244-A-G	Moderate	uORF created	SLC26A4	8 (High)	62		19204907
7-117120115-C-T	Moderate	out-of-frame oORF	CFTR	8 (High)	35		21837768
9-21974860-C-A	Strong	out-of-frame oORF	CDKN2A	8 (High)	35	Hereditary_cancer-predisposing_syndrome|Hereditary_cutaneous_melanoma|Melanoma-pancreatic_cancer_syndrome;Pathogenic;182414	9916806
9-130616761-G-A	Moderate	out-of-frame oORF	ENG	8 (High)	128	Osler_hemorrhagic_telangiectasia_syndrome;Pathogenic;407113	21967607
9-133327612-C-T	Moderate	out-of-frame oORF	ASS1	3 (Low)	5	Citrullinemia_type_I;Likely_pathogenic;203632	19006241
11-5248280-C-T	Moderate	out-of-frame oORF	HBB	8 (High)	29	beta_Thalassemia;Pathogenic;393702	1717406

11-17409692-G-A	Moderate	out-of-frame oORF	KCNJ11	3 (Low)	55		12364426
14-55369403-G-A	Moderate	out-of-frame oORF	GCH1	8 (High)	23		10825351
17-29422056-G-A	Moderate	out-of-frame oORF	NF1	8 (High)	272		27322474
17-70117348-G-A	Moderate	out-of-frame oORF	SOX9	8 (High)	185		28546996
19-11200076-C-A	Weak	uORF created	LDLR	8 (High)	149	Familial_hypercholesterolemia;Likely_pathogenic;250946	22698793
19-11200127-C-T	Moderate	uORF created	LDLR	8 (High)	99	Familial_hypercholesterolemia;Pathogenic;440535	NA
19-11200128-G-A	Strong	out-of-frame oORF	LDLR	8 (High)	97	Familial_hypercholesterolemia;Likely_pathogenic;430743	NA
19-11200202-AC-A	Moderate	uORF created	LDLR	8 (High)	22		25248394
22-50523373-G-A	Moderate	out-of-frame oORF	MLC1	8 (High)	43		25497041
X-38211811-A-G	Weak	uORF created	OTC	8 (High)	141	Ornithine_carbamoyltransferase_deficiency;Likely_pathogenic;487341	NA
X-49114969-C-A	Moderate	out-of-frame oORF	FOXP3	8 (High)	8		16371377
X-148579835-ATG-A	Moderate	uORF created	IDS	3 (Low)	124	Mucopolysaccharidosis,_MPS-II;Pathogenic;10496	1303211
X-154250832-T-C	Moderate	out-of-frame oORF	F8	8 (High)	7		22958177
2-96931137-G-A	Strong	out-of-frame oORF	TMEM127	7 (High)	19	Pheochromocytoma;Likely_pathogenic;126961	21156949
2-157189174-G-A	Moderate	uORF created	NR4A2	7 (High)	310		19429166
4-6271704-G-T	Moderate	out-of-frame oORF	WFS1	7 (High)	44		27395765

5-147211193-G-A	Moderate	uORF created	SPINK1	7 (High)	54		10835640, 27171515, 26228362, 21610753
7-143013247-C-A	Strong	out-of-frame oORF	CLCN1	7 (High)	59		23771340
12-121416385-C-T	Moderate	uORF created	HNF1A	7 (High)	188		10649494
17-66508599-G-A	Strong	out-of-frame oORF	PRKAR1A	7 (High)	97		12424709
2-25387652-G-T	Moderate	out-of-frame oORF	POMC	6 (Mod)	11	Proopiomelanocortin_deficiency;Pathogenic;13355	9620771, 27906547, 23649472
6-26087649-G-A	Moderate	out-of-frame oORF	HFE	6 (Mod)	20		21175851
22-19710933-C-G	Moderate	CDS elongated	GP1BB	6 (Mod)	162	Bernard-Soulier_syndrome,_type_B;Pathogenic;16041	8703016
11-299504-G-A	Strong	CDS elongated	IFITM5	5 (Low)	15	Osteogenesis_imperfecta_type_5;Pathogenic;37143	22863190
19-35773456-G-A	Strong	out-of-frame oORF	HAMP	3 (Low)	25		15198949
1-151372055-G-A	Moderate	CDS elongated	PSMB4	4 (Low)	9	PROTEASOME-ASSOCIATED_AUTOINFLAMMATORY_SYNDROME_3;Pathogenic;548956	28848544

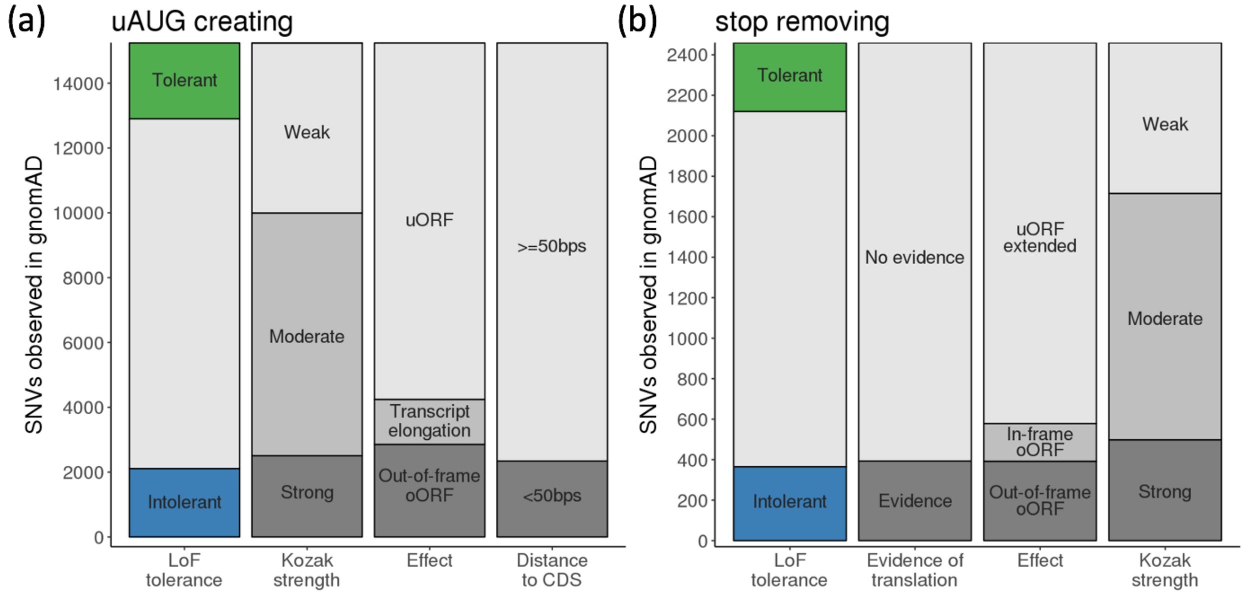
Supplementary Table 2: Stop-removing variants catalogued as disease mutations (DM) in HGMD[1] or (Likely) Pathogenic in ClinVar[2].

variant	Kozak strength	effect	gene	gene class	evidence of translation	clinvar	PubMed IDs
4-159593534-A-G	Strong	out-of-frame oORF	ETFDH	8	N		23628458
8-21988118-T-C	Moderate	CDS elongated	HR	8	Y	Hypotrichosis_4;Pathogenic;7344	NA
17-29422055-A-C	Strong	out-of-frame oORF	NF1	8	Y		27322474
X-68049525-T-C	Strong	out-of-frame oORF	EFNB1	8	N		23335590

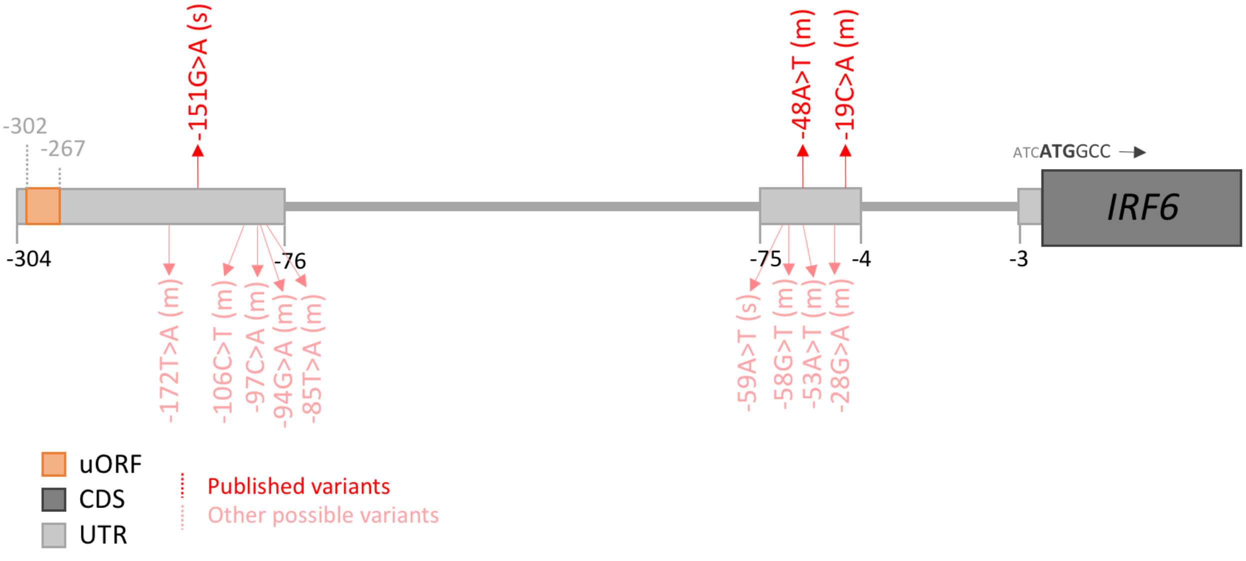
Supplementary Table 3: Genes with ≥10 possible predicted high-impact uAUG-creating or stop-removing SNVs, and for which LoF and/or haploinsufficiency is a known mechanism of human disease (either curated as haploinsufficient, curated as acting via a LoF mechanism in DDG2P or with ≥10 high-confidence pathogenic LoF variants documented in ClinVar).

https://github.com/ImperialCardioGenetics/uORFs/blob/master/data_files/SupplementaryTable3

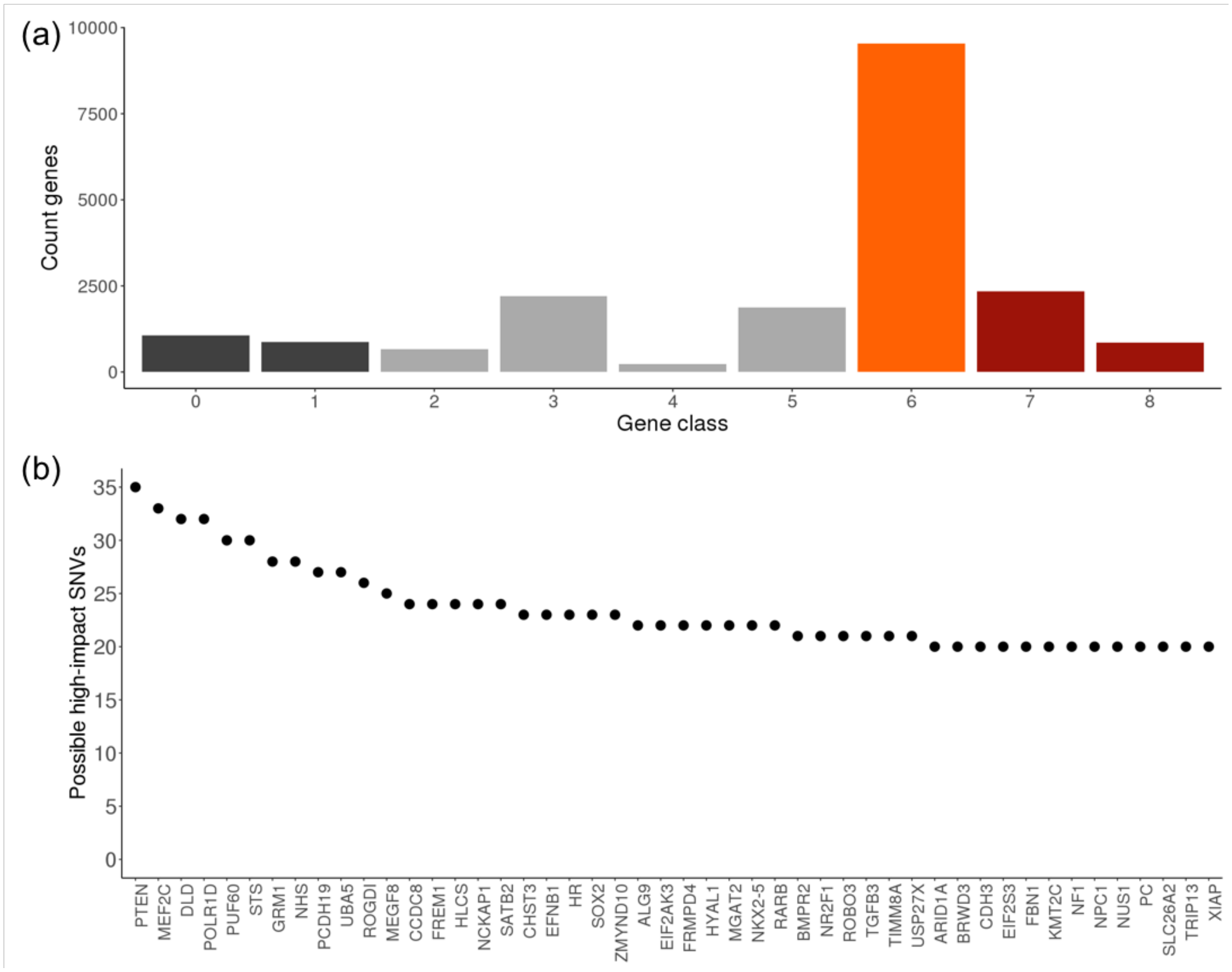
Supplementary Figure 1: Observed numbers of uAUG-creating and stop-removing SNVs in gnomAD.



Supplementary Figure 2: Schematic of the 5'UTR of IRF6 showing the location of uAUG-creating variants identified by de Lima et al. (bright red) and all other possible uAUG-creating SNVs (faded red) which would be created into a strong or moderate Kozak consensus and form an out-of-frame oORF. The strength of the Kozak consensus is shown in brackets ("s" for strong, "m" for moderate).



Supplementary Figure 3: Identifying genes where there is a high likelihood that uORF-perturbing variants will be deleterious. (a) Plot of all 18,593 by category (see methods) coloured by unknown (dark grey), low (light grey), moderate (orange) and high likelihood (dark red) that uORF-perturbing variants will be deleterious. (b) Genes in class 8 (not classified with a 'Low' likelihood and where LoF and/or haploinsufficiency is a known mechanism of human disease) with ≥20 possible high-impact uAUG-creating and stop-removing SNVs.



Methods

Ethics statement

We have complied with all relevant ethical regulations. This study was overseen by the Broad Institute's Office of Research Subject Protection and the Partners Human Research Committee, and was given a determination of Not Human Subjects Research. Informed consent was obtained from all participants.

Definition of 5'UTRs

The start and end positions and sequence of the 5'UTRs of all protein-coding genes were downloaded from Ensembl biomart (Human genes GRCh37.p13) and filtered to only include canonical transcripts. Genes with no annotated 5'UTR on the canonical transcript were removed.

Identification and classification of uAUG-creating variants

Reading through each UTR from start to end (5' to 3'), we identified all instances where a SNV would create an ATG. We recorded the positions of all possible stop codons (TAA, TGA and TAG) and annotated each uAUG-creating variant with whether or not there was an in-frame stop codon within the UTR. To annotate the strength of the Kozak consensus into which the uAUG was formed we assessed the positions at -3 and +3 relative to the A of the AUG, known to be the most important bases for dictating strength of translation. If both the -3 base was either A or G and the +3 was G, Kozak was annotated as 'Strong', if either of these conditions was true, Kozak was deemed to be 'Moderate' and if neither was the case 'Weak'. uAUG-creating variants were also annotated with the distance to, and the frame relative to the coding sequence (CDS).

Identification and classification of uORF stop-removing variants

Existing uORFs were defined as the combination of an ATG and in-frame stop codon (TAA, TGA or TAG) within a UTR. Each predicted uORF was annotated with the positions of all alternative downstream in-frame stop codons within the UTR and with the frame relative to the coding sequence. The Kozak strength of each uORF was defined as outlined above for uAUG-creating variants. Where multiple uAUGs converge on the same stop codon, the uORF is annotated with the strongest Kozak consensus. To identify uORFs with prior evidence of translation we downloaded all human small open reading frames (sORFs) from sorfs.org, a public repository of sORFs identified in humans, mice and fruit flies using ribosome profiling [Olexiouk *et al*. 2015]. Predicted uORFs were marked as having prior evidence if the annotated stop codon matched an entry from sorfs.org.

Stop-removing variants were identified as SNVs that would change the base of a stop codon to any sequence that would not retain the stop (i.e. did not create another of TAA, TGA or TAG).

Calculating MAPS

For each set of variants we computed the mutability adjusted proportion of singletons, or MAPS. The basis of this approach has previously been described[23]. Briefly, for each substitution, accounting for 1 base of surrounding context (e.g. ACG -> ATG), we calculated the proportion of all possible variants (-3.9885 < GERP < 2.6607, 15x < gnomAD coverage < 60x) that are observed in intergenic/intronic autosomal regions in a downsampled set of 1000 gnomAD whole-genomes. For C>T changes at CpG sites, variant proportions are calculated separately for three distinct bins of methylation. These proportions are then scaled so that the weighted genome-wide average is the human per-base, per-generation mutation rate (1.2e-8). The creation of these context-dependent mutation rates is more fully described in our companion paper[22].

To determine the transformation between these mutation rates and the expected proportion of singletons, for each substitution and context (and methylation bin for CpGs), we regress the mutation rates against the observed proportion of singletons for synonymous variants. We use synonymous as a relatively neutral class of variants which should not be subject to any biases being investigating in UTRs, but that are distinct from bases used to define the model.

For a given list of possible variants, annotated with gnomAD allele counts using Hail (https://hail.is), we take only those that are observed in gnomAD and annotate each with the transformed mutation rate given the variant context (which now corresponds to the expected chance this site will be a singleton), and sum these values across the entire variant list to give an expected number of singletons. Variants are excluded if they are outliers on coverage in gnomAD (15x < coverage < 60x), were found on the X or Y chromosome, or were filtered out of the gnomAD whole genomes.

Finally, this expected number of singletons is compared to the number of sites that are observed as singletons in gnomAD, to estimate MAPS.

$$MAPS = (observed\ singletons - expected\ singletons)\ /\ total\ observed\ variants$$

Confidence intervals were calculated using bootstrapping. For a list of *n* observed variants, *n* variant sites are sampled at random with replacement and used to calculate MAPS. This is repeated over 10,000 permutations before the 5th and 95th percentiles of the resulting MAPS distribution are taken as confidence intervals.

P-values we calculated using the same bootstrapping approach but for each permutation MAPS was calculated for each of the two variant sets of interest, A and B. The *P*-value was defined as the proportion of permutations where MAPS of B was less than MAPS of A.

$$P = \Sigma\,[(MAPS(B) - MAPS(A)) < 0]\ /\ permutations$$

For coding variants, MAPS was calculated using the predicted impact on the canonical transcript.

Using PhyloP to assess base-level conservation

Per-base vertebrate PhyloP scores were extracted from the Combined Annotation Dependent Depletion (CADD) version v1.4 GRCh37 release files and used to annotate lists of all possible coding, UTR and uORF stop bases. To remove biases due to gene context and distance from the coding sequence, we created a set of matched UTR bases which comprised the 3 bases immediately upstream and downstream of the stop. Conserved bases were defined as those with PhyloP >= 2. We also checked for a significant difference between the entire distribution of scores using a Wilcoxon rank sum test for all stop-removing compared to matched UTR bases (P=8.1x10^{-9}).

Identifying disease gene lists

Developmental disease genes were downloaded from The Developmental Disorders Genotype-Phenotype Database (DDG2P) on the 6th October 2018. We included only genes categorised as ‘confirmed’ or ‘probable’. Genes with a known dominant LoF mechanism were identified using the ‘allelic requirement’ and ‘mutation consequence’ annotations.

Genes intolerant and tolerant to LoF variants were identified using data from Karczewski *et al.* 2019[22]. Genes were ordered by their loss-of-function observed/expected upper bound fraction (LOEUF) scores and the top and bottom sextiles were categorised as tolerant and intolerant respectively.

We downloaded data from The Clinical Genome Resource (ClinGen) Dosage Sensitivity Map on 21st January 2019 (https://www.ncbi.nlm.nih.gov/projects/dbvar/clingen/). Genes manually curated as haploinsufficient were defined as those with a score of 3 (sufficient evidence). In addition, we added genes curated as severe or moderately haploinsufficient by the MacArthur lab (https://github.com/macarthur-lab/gene_lists/tree/master/lists).

Searching for uORF-perturbing variants in HGMD and ClinVar

Lists of all possible uAUG-creating and stop-removing SNVs were intersected with all DM variants from HGMD pro release 2018.1 and all ClinVar Pathogenic or Likely Pathogenic variants from the August 2018 release (clinvar_20180805.vcf). In addition, we created a list of all possible 1-5bp deletions that would create an uAUG, annotated as described for SNVs above, and also searched for these variants. We did not investigate small insertions or deletion >5bps due to the inhibitory number of possible variants.

Sub-classifying genes

uAUG-creating variants were classified as 'high-impact' if they are formed into a high or moderate Kozak consensus and if they either form an oORF or result in transcript elongation. Stop removing variants were similarly classified as 'high-impact' if the original uORF start site has a strong or moderate Kozak and/or the uORF is documented in sorfs.org and the variants results in a oORF or a transcript elongation.

Genes were divided into nine categories according to the following logic.
Class 0 - genes with no annotated 5'UTR on the canonical transcript
Class 1 - genes with no possible uAUG-creating or stop-removing SNVs identified
Class 2 - remaining genes with no possible SNVs of predicted high-impact
Class 3 - remaining genes where the UTR has a high-confidence oORF (strong/moderate Kozak or documented in sorfs.orf) indicating creating a second would be of low-impact
Class 4 - remaining genes where one or more identified high-impact SNVs have AF > 0.1% in gnomAD (genomes AC>15)
Class 5 - remaining genes that are intolerant to LoF variants
Class 8 - remaining genes curated as haploinsufficient by ClinGen or the MacArthur lab, curated as acting via a loss-of-function mechanism in DDG2P or with >=10 high-confidence Pathogenic LoF variants in ClinVar (known LoF disease genes)
Class 7 - remaining genes intolerant to LoF variants or with >=2 high-confidence Pathogenic LoF variants in ClinVar
Class 6 - all genes not classified into any other class

The nine gene classes were grouped into three categories corresponding to low (classes 2,3,4 and 5), moderate (class 6) and high (classes 7 and 8) likelihood that high-impact uORF-perturbing variants would have a deleterious effect.

Sequencing of individuals with neurofibromatosis type 2

A cohort of 1,134 unrelated individuals with neurofibromatosis type 2 were recruited to the Centre for Genomic Medicine at St Mary's Hospital, Manchester. All patients were sequenced across the *NF2* gene. Two individuals were identified to carry a single 5'UTR variant (ENST00000338641:-66-65insT; GRCh37:chr22:29999922 A>AT). Both carriers were confirmed to have no variants in *SMARCB1* or *LZTRA1* and no coding variants in *NF2*. The -66-65insT variant segregated with disease in 3 affected siblings in one family and in affected parent and child in another.

Software/data availability

To aid in the identification of uORF-perturbing variants we have created a VEP plugin which annotates variants for predicted effects on translational regulation. This script is freely available at https://github.com/ImperialCardioGenetics/uORFs. All possible uAUG-creating and stop-removing SNVs for canonical Gencode transcripts along with likelihood classifications for all genes are also available for download at https://github.com/ImperialCardioGenetics/uORFs.

Group authors

Genome Aggregation Database Production Team: Jessica Alföldi[1,2], Irina M. Armean[3,1,2], Eric Banks[4], Louis Bergelson[4], Kristian Cibulskis[4], Ryan L Collins[1,5,6], Kristen M. Connolly[7], Miguel Covarrubias[4], Beryl Cummings[1,2,8], Mark J. Daly[1,2,9], Stacey Donnelly[1], Yossi Farjoun[4], Steven Ferriera[10], Laurent Francioli[1,2], Stacey Gabriel[10], Laura D. Gauthier[4], Jeff Gentry[4], Namrata Gupta[10,1], Thibault Jeandet[4], Diane Kaplan[4], Konrad J. Karczewski[1,2], Kristen M. Laricchia[1,2], Christopher Llanwarne[4], Eric V. Minikel[1], Ruchi Munshi[4], Benjamin M Neale[1,2], Sam Novod[4], Anne H. O'Donnell-Luria[1,11,12], Nikelle Petrillo[4], Timothy Poterba[9,2,1], David Roazen[4], Valentin Ruano-Rubio[4], Andrea Saltzman[1], Kaitlin E. Samocha[13], Molly Schleicher[1], Cotton Seed[9,2], Matthew Solomonson[1,2], Jose Soto[4], Grace Tiao[1,2], Kathleen Tibbetts[4], Charlotte Tolonen[4], Christopher Vittal[9,2], Gordon Wade[4], Arcturus Wang[9,2,1], Qingbo Wang[1,2,6], James S Ware[14,15,1], Nicholas A Watts[1,2], Ben Weisburd[4], Nicola Whiffin[14,15,1]

1. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
2. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA
3. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom
4. Data Sciences Platform, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
5. Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA
6. Program in Bioinformatics and Integrative Genomics, Harvard Medical School, Boston, MA 02115, USA
7. Genomics Platform, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
8. Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, 02115, USA
9. Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
10. Broad Genomics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
11. Division of Genetics and Genomics, Boston Children's Hospital, Boston, Massachusetts 02115, USA
12. Department of Pediatrics, Harvard Medical School, Boston, Massachusetts 02115, USA
13. Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK
14. National Heart & Lung Institute and MRC London Institute of Medical Sciences, Imperial College London, London UK
15. Cardiovascular Research Centre, Royal Brompton & Harefield Hospitals NHS Trust, London UK

Genome Aggregation Database Consortium: Carlos A Aguilar Salinas[1], Tariq Ahmad[2], Christine M. Albert[3,4], Diego Ardissino[5], Gil Atzmon[6,7], John Barnard[8], Laurent Beaugerie[9], Emelia J. Benjamin[10,11,12], Michael Boehnke[13], Lori L. Bonnycastle[14], Erwin P. Bottinger[15], Donald W Bowden[16,17,18], Matthew J Bown[19,20], John C Chambers[21,22,23], Juliana C. Chan[24], Daniel Chasman[3,25], Judy Cho[15], Mina K. Chung[26], Bruce Cohen[27,25], Adolfo Correa[28], Dana Dabelea[29], Mark J. Daly[30,31,32], Dawood Darbar[33], Ravindranath Duggirala[34], Josée Dupuis[35,36], Patrick T. Ellinor[30,37], Roberto Elosua[38,39,40], Jeanette Erdmann[41,42,43], Tõnu Esko[30,44], Martti Färkkilä[45], Jose Florez[46], Andre Franke[47], Gad Getz[48,49,25], Benjamin Glaser[50], Stephen J. Glatt[51], David Goldstein[52,53], Clicerio Gonzalez[54], Leif Groop[55,56], Christopher Haiman[57], Craig Hanis[58], Matthew Harms[59,60], Mikko Hiltunen[61], Matti M. Holi[62], Christina M. Hultman[63,64], Mikko Kallela[65], Jaakko Kaprio[56,66], Sekar Kathiresan[67,68,25], Bong-Jo Kim[69], Young Jin Kim[69], George Kirov[70], Jaspal Kooner[23,22,71], Seppo Koskinen[72], Harlan M. Krumholz[73], Subra Kugathasan[74], Soo Heon Kwak[75], Markku Laakso[76,77], Terho Lehtimäki[78], Ruth J.F. Loos[15,79], Steven A.

Lubitz[30,37], Ronald C.W. Ma[24,80,81], Daniel G. MacArthur[31,30], Jaume Marrugat[82,39], Kari M. Mattila[78], Steven McCarroll[32,83], Mark I McCarthy[84,85,86], Dermot McGovern[87], Ruth McPherson[88], James B. Meigs[89,25,90], Olle Melander[91], Andres Metspalu[44], Benjamin M Neale[30,31], Peter M Nilsson[92], Michael C O'Donovan[70], Dost Ongur[27,25], Lorena Orozco[93], Michael J Owen[70], Colin N.A. Palmer[94], Aarno Palotie[56,32,31], Kyong Soo Park[75,95], Carlos Pato[96], Ann E. Pulver[97], Nazneen Rahman[98], Anne M. Remes[99], John D. Rioux[100,101], Samuli Ripatti[56,66,102], Dan M. Roden[103,104], Danish Saleheen[105,106,107], Veikko Salomaa[108], Nilesh J. Samani[19,20], Jeremiah Scharf[30,32,67], Heribert Schunkert[109,110], Moore B. Shoemaker[111], Pamela Sklar*[112,113,114], Hilkka Soininen[115], Harry Sokol[9], Tim Spector[116], Patrick F. Sullivan[63,117], Jaana Suvisaari[108], E Shyong Tai[118,119,120], Yik Ying Teo[118,121,122], Tuomi Tiinamaija[56,123,124], Ming Tsuang[125,126], Dan Turner[127], Teresa Tusie-Luna[128,129], Erkki Vartiainen[66], James S Ware[130,131,30], Hugh Watkins[132], Rinse K Weersma[133], Maija Wessman[123,56], James G. Wilson[134], Ramnik J. Xavier[135,136]

1. Unidad de Investigacion de Enfermedades Metabolicas. Instituto Nacional de Ciencias Medicas y Nutricion. Mexico City
2. Peninsula College of Medicine and Dentistry, Exeter, UK
3. Division of Preventive Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA.
4. Division of Cardiovascular Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA.
5. Department of Cardiology, University Hospital, 43100 Parma, Italy
6. Department of Biology, Faculty of Natural Sciences, University of Haifa, Haifa, Israel
7. Departments of Medicine and Genetics, Albert Einstein College of Medicine, Bronx, NY, USA, 10461
8. Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44122, USA
9. Sorbonne Université, APHP, Gastroenterology Department, Saint Antoine Hospital, Paris, France
10. NHLBI and Boston University's Framingham Heart Study, Framingham, Massachusetts, USA.
11. Department of Medicine, Boston University School of Medicine, Boston, Massachusetts, USA.
12. Department of Epidemiology, Boston University School of Public Health, Boston, Massachusetts, USA.
13. Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan 48109
14. National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA
15. The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY
16. Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC, USA
17. Center for Genomics and Personalized Medicine Research, Wake Forest School of Medicine, Winston-Salem, NC, USA
18. Center for Diabetes Research, Wake Forest School of Medicine, Winston-Salem, NC, USA
19. Department of Cardiovascular Sciences, University of Leicester, Leicester, UK
20. NIHR Leicester Biomedical Research Centre, Glenfield Hospital, Leicester, UK
21. Department of Epidemiology and Biostatistics, Imperial College London, London, UK
22. Department of Cardiology, Ealing Hospital NHS Trust, Southall, UK
23. Imperial College Healthcare NHS Trust, Imperial College London, London, UK
24. Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong, China.
25. Department of Medicine, Harvard Medical School, Boston, MA
26. Departments of Cardiovascular Medicine, Cellular and Molecular Medicine, Molecular Cardiology, and Quantitative Health Sciences, Cleveland Clinic, Cleveland, Ohio, USA.
27. McLean Hospital, Belmont, MA
28. Department of Medicine, University of Mississippi Medical Center, Jackson, Mississippi, USA
29. Department of Epidemiology, Colorado School of Public Health, Aurora, Colorado, USA.
30. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
31. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA

32. Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA
33. Department of Medicine and Pharmacology, University of Illinois at Chicago
34. Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX, USA
35. Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA
36. National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, MA 01702, USA
37. Cardiac Arrhythmia Service and Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA
38. Cardiovascular Epidemiology and Genetics, Hospital del Mar Medical Research Institute (IMIM). Barcelona, Catalonia, Spain
39. CIBER CV, Barcelona, Catalonia, Spain
40. Departament of Medicine, Medical School, University of Vic-Central University of Catalonia. Vic, Catalonia, Spain
41. Institute for Cardiogenetics, University of Lübeck, Lübeck, Germany
42. 1. DZHK (German Research Centre for Cardiovascular Research), partner site Hamburg/Lübeck/Kiel, 23562 Lübeck, Germany
43. University Heart Center Lübeck, 23562 Lübeck, Germany
44. Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia
45. Helsinki University and Helsinki University Hospital, Clinic of Gastroenterology, Helsinki, Finland.
46. Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital; Programs in Metabolism and Medical & Population Genetics, Broad Institute; Department of Medicine, Harvard Medical School
47. Institute of Clinical Molecular Biology (IKMB), Christian-Albrechts-University of Kiel, Kiel, Germany
48. Bioinformatics Program, MGH Cancer Center and Department of Pathology
49. Cancer Genome Computational Analysis, Broad Institute.
50. Endocrinology and Metabolism Department, Hadassah-Hebrew University Medical Center, Jerusalem, Israel
51. Department of Psychiatry and Behavioral Sciences; SUNY Upstate Medical University
52. Institute for Genomic Medicine, Columbia University Medical Center, Hammer Health Sciences, 1408, 701 West 168th Street, New York, New York 10032, USA.
53. Department of Genetics & Development, Columbia University Medical Center, Hammer Health Sciences, 1602, 701 West 168th Street, New York, New York 10032, USA.
54. Centro de Investigacion en Salud Poblacional. Instituto Nacional de Salud Publica MEXICO
55. Lund University, Sweden
56. Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland
57. Lund University Diabetes Centre
58. Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX 77030
59. Department of Neurology, Columbia University
60. Institute of Genomic Medicine, Columbia University
61. Institute of Biomedicine, University of Eastern Finland, Kuopio, Finland
62. Department of Psychiatry, PL 320, Helsinki University Central Hospital, Lapinlahdentie, 00 180 Helsinki, Finland
63. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
64. Icahn School of Medicine at Mount Sinai, New York, NY, USA
65. Department of Neurology, Helsinki University Central Hospital, Helsinki, Finland.
66. Department of Public Health, Faculty of Medicine, University of Helsinki, Finland
67. Center for Genomic Medicine, Massachusetts General Hospital, Boston, Massachusetts 02114, USA
68. Cardiovascular Disease Initiative and Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
69. Center for Genome Science, Korea National Institute of Health, Chungcheongbuk-do, Republic of Korea.
70. MRC Centre for Neuropsychiatric Genetics & Genomics, Cardiff University School of Medicine, Hadyn Ellis Building, Maindy Road, Cardiff CF24 4HQ
71. National Heart and Lung Institute, Cardiovascular Sciences, Hammersmith Campus, Imperial College London, London, UK.
72. Department of Health, THL-National Institute for Health and Welfare, 00271 Helsinki, Finland.
73. Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, Connecticut3Center for Outcomes Research and Evaluation, Yale-New Haven Hospital, New Haven, Connecticut.
74. Division of Pediatric Gastroenterology, Emory University School of Medicine, Atlanta, Georgia, USA.
75. Department of Internal Medicine, Seoul National University Hospital, Seoul, Republic of Korea
76. The University of Eastern Finland, Institute of Clinical Medicine, Kuopio, Finland

77. Kuopio University Hospital, Kuopio, Finland
78. Department of Clinical Chemistry, Fimlab Laboratories and Finnish Cardiovascular Research Center-Tampere, Faculty of Medicine and Health Technology, Tampere University, Finland
79. The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY
80. Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong, China.
81. Hong Kong Institute of Diabetes and Obesity, The Chinese University of Hong Kong, Hong Kong, China.
82. Cardiovascular Research REGICOR Group, Hospital del Mar Medical Research Institute (IMIM). Barcelona, Catalonia.
83. Department of Genetics, Harvard Medical School, Boston, MA, USA
84. Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Old Road, Headington, Oxford, OX3 7LJ UK
85. Wellcome Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK
86. Oxford NIHR Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford OX3 9DU, UK
87. F Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA.
88. Atherogenomics Laboratory, University of Ottawa Heart Institute, Ottawa, Canada
89. Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, 02114
90. Program in Population and Medical Genetics, Broad Institute, Cambridge, MA
91. Department of Clinical Sciences, University Hospital Malmo Clinical Research Center, Lund University, Malmo, Sweden.
92. Lund University, Dept. Clinical Sciences, Skane University Hospital, Malmo, Sweden
93. Instituto Nacional de Medicina Genómica (INMEGEN), Mexico City, 14610, Mexico
94. Medical Research Institute, Ninewells Hospital and Medical School, University of Dundee, Dundee, UK.
95. Department of Molecular Medicine and Biopharmaceutical Sciences, Graduate School of Convergence Science and Technology, Seoul National University, Seoul, Republic of Korea
96. Department of Psychiatry, Keck School of Medicine at the University of Southern California, Los Angeles, California, USA.
97. Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA
98. Division of Genetics and Epidemiology, Institute of Cancer Research, London SM2 5NG
99. Medical Research Center, Oulu University Hospital, Oulu, Finland and Research Unit of Clinical Neuroscience, Neurology, University of Oulu, Oulu, Finland.
100. Research Center, Montreal Heart Institute, Montreal, Quebec, Canada, H1T 1C8
101. Department of Medicine, Faculty of Medicine, Université de Montréal, Québec, Canada
102. Broad Institute of MIT and Harvard, Cambridge MA, USA
103. Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA.
104. Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, USA.
105. Department of Biostatistics and Epidemiology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA
106. Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA
107. Center for Non-Communicable Diseases, Karachi, Pakistan
108. National Institute for Health and Welfare, Helsinki, Finland
109. Deutsches Herzzentrum München, Germany
110. Technische Universität München
111. Division of Cardiovascular Medicine, Nashville VA Medical Center and Vanderbilt University, School of Medicine, Nashville, TN 37232-8802, USA.
112. Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA
113. Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA
114. Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA
115. Institute of Clinical Medicine, neurology, University of Eastern Finad, Kuopio, Finland
116. Department of Twin Research and Genetic Epidemiology, King's College London, London UK
117. Departments of Genetics and Psychiatry, University of North Carolina, Chapel Hill, NC, USA

118. Saw Swee Hock School of Public Health, National University of Singapore, National University Health System, Singapore
119. Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
120. Duke-NUS Graduate Medical School, Singapore
121. Life Sciences Institute, National University of Singapore, Singapore.
122. Department of Statistics and Applied Probability, National University of Singapore, Singapore.
123. Folkhälsan Institute of Genetics, Folkhälsan Research Center, Helsinki, Finland
124. HUCH Abdominal Center, Helsinki University Hospital, Helsinki, Finland
125. Center for Behavioral Genomics, Department of Psychiatry, University of California, San Diego
126. Institute of Genomic Medicine, University of California, San Diego
127. Juliet Keidan Institute of Pediatric Gastroenterology, Shaare Zedek Medical Center, The Hebrew University of Jerusalem, Israel
128. Instituto de Investigaciones Biomédicas UNAM Mexico City
129. Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán Mexico City
130. National Heart & Lung Institute & MRC London Institute of Medical Sciences, Imperial College London, London UK
131. Cardiovascular Research Centre, Royal Brompton & Harefield Hospitals NHS Trust, London UK
132. Radcliffe Department of Medicine, University of Oxford, Oxford UK
133. Department of Gastroenterology and Hepatology, University of Groningen and University Medical Center Groningen, Groningen, the Netherlands
134. Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS 39216, USA
135. Program in Infectious Disease and Microbiome, Broad Institute of MIT and Harvard, Cambridge, MA, USA
136. Center for Computational and Integrative Biology, Massachusetts General Hospital