A platform for curated products from novel Open Reading Frames (nORFs) prompts reinterpretation of disease variants.

Matthew DC Neville[1†], Robin Kohze[1†], Chaitanya Erady[1], Narendra Meena[2], Matthew Hayden[3], David N. Cooper[3], Matthew Mort[3], Sudhakaran Prabakaran[1,2,4*]

[1]Department of Genetics, University of Cambridge, Downing Site, CB2 3EH, UK

[2]Department of Biology, Indian Institute of Science Education and Research, Pune, Maharashtra, 411008, India

[3]Institute of Medical Genetics, Cardiff University, Heath Park, Cardiff CF14 4XN, UK [4]St Edmund's College, University of Cambridge, CB3 0BN, UK

*Corresponding author, email: sp339@cam.ac.uk.

[†]These authors contributed equally to the work.

Abstract

Recent evidence from proteomics and deep massively parallel sequencing studies have revealed that eukaryotic genomes contain substantial numbers of as yet uncharacterised open reading frames (ORFs). We define these uncharacterised open reading frames as novel open reading frames (nORFs). nORFs in humans are mostly under 100 codons and found in diverse regions of the genome, including in long noncoding RNAs, pseudogenes, 3'UTRs, 5'UTRs, and alternative reading frames of canonical protein coding exons. There is therefore a pressing need to evaluate the potential functional importance of these unannotated transcripts and proteins in biological pathways and human disease on a larger scale, rather than one at a time. In this study, we outline the creation of a valuable nORFs dataset with experimental evidence of translation for the community, use measures of heritability and selection that reveal signals for functional importance in nORFs and demonstrate the potential implications for functional interpretation of genetic variants in nORFs. Our results indicate that some variants that have been previously classified as being benign or of uncertain significance may have to be reinterpreted.

Introduction

Recent evidence from proteomics, proteogenomics, ribosome profiling, and massively parallel sequencing studies have revealed that prokaryotic and eukaryotic genomes contain a substantial number of as yet uncharacterised and unannotated open reading frames (ORFs) (Firth and Brierley 2012; Miravet-Verde et al. 2019; Hellens et al. 2016; Albuquerque et al. 2015; Saghatelian and Couso 2015; Andrews and Rothnagel 2014; Prabakaran et al. 2014; Brunet et al. 2018). These ORFs have largely evaded detection due to the original conservative definition of a gene with annotation criteria of one coding sequence (CDS) per transcript, a minimum of 100 codons for each CDS, 'ATG' as the only start codon, and conservative definitions of Kozak sequences (Brunet et al. 2018; Plaza et al. 2017). There have also been recent advances in the sensitivity of proteomics methods such as ribosome profiling and mass spectrometry (MS), the ability to

sequence genomes and transcripts at a deeper depths, and in the ability to integrate these two data types, which our lab specializes in and calls *Systems Proteogenomics* (Prabakaran et al. 2014). These advances have revealed that we have underestimated the genome's coding potential, with many unannotated ORFs showing evidence of translation in humans alone (Ma et al. 2014; Jagannathan et al. 2019; Erady et al. 2019; Chen, J. et al., 2020). These ORFs are mostly under 100 codons and found in diverse regions of the genome, including in long noncoding RNAs (lncRNAs), pseudogenes, 3'UTRs, 5'UTRs, and alternative reading frames of canonical protein coding exons (Prabakaran et al. 2014; Brunet et al. 2018; Plaza et al. 2017).

Based on their genomic location and their size, these unannotated ORFs have been defined in numerous ways including as short ORFs (sORFs), small ORFs (smORFs), alternative ORFs (altORFs), and upstream/downstream ORFs (u/dORFs). The definitions for these labels, like those for original gene annotations, again set arbitrary bounds and even tend to vary between reports and species (Olexiouk et al. 2018). In this study, we have attempted to collate and reclassify all of these observed ORFs and we refer here to any unannotated ORF as a 'novel ORF' or 'nORF', which encompasses all of the above definitions. Specifically, a nORF is any ORF that can encode a not yet classified transcript or protein product, nor an isoform of one, with no bounds on number of codons, location, number of ORFs per transcript, nor start codon. Although nORFs may appear by chance in the genome (Olexiouk et al. 2018), in this study we have exclusively focused on nORFs with experimental evidence of translation from MS or ribosome profiling studies and have attempted to interpret their potential functional consequences.

For humans, two published databases of note: OpenProt (Brunet et al. 2019) and sORFs.org (Olexiouk et al. 2018) have compiled and analyzed translation data from ribosome profiling and MS studies to identify and share novel proteins. Evidence from these two databases has helped

challenge conventional gene annotations to provide critical data for the field of nORFs, but both still have important limitations because of ambiguous definitions of these nORFs. sORFs.org for instance only considers ORFs under 100 codons, presents many duplicate or highly similar entries, and shares data in formats difficult to use in downstream analyses. OpenProt, though more accessible, has far fewer entries with experimental evidence, partly due to only considering ORFs above 30 codons and limiting ORFs to ATG start codons and canonical transcripts. Additionally, annotation pipelines differ between the databases and are somewhat outdated, making comparisons difficult. Overall, the field lacks a consensus definition of what a nORF is and an accessible central resource with nORF data consolidated in a consistent manner. In this study we address these needs with the curation and redefinition of nORFs.

While mechanisms by which uORFs influence the translation of nearby canonical genes have been investigated (McGillivray et al. 2018; Whiffin et al. 2020), the functional consequence of nORFs more generally remains largely unexplored. The functional evidence which does exist has implicated nORFs in roles both related to and independent of nearby canonical genes (Brunet et al. 2018), including mRNA decapping (Pueyo et al. 2016), muscle regeneration (Matsumoto et al. 2017), and insulin secretion (Hu et al. 2016). Additionally, we have shown previously that nORF encoded protein-like products can form structures with potential biological functions (Erady et al. 2019), can be regulated by post-translational modifications (Jagannathan et al. 2019), are biologically regulated in mouse neurons (Prabakaran et al. 2014), and harbour deleterious mutations from cancer and other inherited diseases (Jagannathan et al. 2019; Erady et al. 2019). Despite these examples, the vast majority of nORFs have no known function and their exclusion from canonical genome annotations means that many of the studies that could uncover roles for nORFs do not even consider them.

In this study, we have investigated the potential functional importance of a curated set of nORFs genome-wide. We begin at a broad scale, investigating the heritability associated with nORF regions for several human traits and diseases. We then narrow our focus from nORF regions to specific classes of nORF variants (e.g. nORF stop gained variants), to evaluate potential signals of negative selection, which would be indicative of functional importance. Finally, we move to specific genetic variants, such as those known to cause disease, to investigate whether their pathogenicity can be explained by their effect on nORFs. In particular, we highlight disease mutations that appear benign to canonical proteins but highly deleterious to nORFs, the clearest potential examples of nORF functional importance and hence warranting reinterpretation of some variants of benign or unknown significance.

Results

Data Overview

The nORFs dataset contains 194,407 ORFs curated from OpenProt (Brunet et al. 2019) and sORFs.org (Olexiouk et al. 2018) **(Fig. 1A),** which we have made publicly available on the nORFs.org platform **(Fig. 2)**. The curation steps **(Fig. 1A)** involved selecting unique ORFs with translation evidence from MS or ribosome profiling experiments that are distinct from each other **(Fig. 1B)** and from canonical proteins **(Fig. 1C)**. These nORFs were annotated with respect to canonical transcripts and CDS, and they are found in diverse locations in the genomes such as overlapping canonical CDS in alternate frames (altCDS), in UTRs, in noncoding RNAs (ncRNAs), and in intronic/intergenic regions **(Fig. 3A)**. From the 194,407 nORFs, we found that 98,577 (50.7%) fully overlap canonical CDS, 31,361 (16.1%) overlap CDS and intron regions, 28,067 (14.4%) overlap 5'UTRs, 5509 (2.8%) overlap 3'UTRs, 19,909 (10.2%) overlap ncRNAs, and 4,836 (2.5%) fully map to intronic or intergenic regions **(Fig. 3B)**. The length distribution of nORFs for each major annotation category falls mostly below 100 amino acids, with mean lengths of 39.8aa, 27.6aa, 29.9aa, and 54.4aa for UTRs, altCDS, intergenic, and ncRNA nORFs

respectively, much smaller, as previously reported (Jagannathan et al. 2019), than the mean canonical protein length of 557.3aa **(Fig. 3C)**. They are found spread throughout all 22 autosomes, both sex chromosomes and on mitochondrial DNA, similar to canoncial CDS **(Fig 3D)**.

We compared this nORF dataset with previously published uORF dataset (McGillivray et al. 2018 NAR 2018). We note that the sources of uORF entries from McGillivray et al. (Lee et al., Fritsch et al., and Gao et al.) are three of the ribosome profiling experiments also used as input for the sORFs.org dataset. Comparing the 188,802 “likely active” uORFs from McGillivray et al. 2018, with the 194,407 nORFs from this work, we find that there are 15,082 entries that are identical or highly similar (share stop codon but differ in start codon) between datasets. The majority of these shared entries fall, as expected, under the nORFs classified as 5’UTR (7,333) or 5’UTR-altCDS (3,681). The entries in the nORFs dataset not found in the uORF dataset can be attributed to the broader set of experiments used as input from sORFs.org and OpenProt, and the broader focus of any all unannotated ORFs, compared to the specific uORF focus of McGillivray et al. 2018. As the 188,802 “likely active” uORFs from McGillivray et al. 2018 would have been found in sORFs.org dataset, those not found in the nORFs dataset would have been filtered out at one of data curation steps performed (e.g. good/extreme ORFscore, longest ORF if similar, removing inframe entries (**Figure 1a.**)).

Heritability

We investigated the heritability of nORF regions in the genome to assess their importance to human traits and disease. To achieve this we applied Stratified LD score regression (S-LDSC) (Finucane et al. 2015; Gazal et al. 2017) with the baseline-LF model (Gazal et al. 2018) developed to assess both common (minor allele frequency (MAF) ≥ 5%) and low frequency

(0.5%≤ MAF <5%) heritability in complex traits. As applied by Gazal et al. (Gazal et al. 2018), we

used a UK10K (The UK10K Consortium 2015) LD reference panel, analyzing 40 heritable, complex UK Biobank (Bycroft et al. 2018) traits restricted to 409 thousand individuals with UK ancestry.

With this we analyzed all baseline-LF model annotations and custom annotations (see Methods). For 67 baseline-LF annotations and for our 7 custom annotations, we calculated heritability enrichments in each of the 40 UK Biobank traits. For each annotation, common variant enrichment (CVE) and low frequency variant enrichment (LFVE) were meta-analyzed across 27 independent traits (**Supplemental Table S1; Supplemental Table S2**). To interpret heritability enrichments of nORFs we focus on 4 custom annotations from canonical genes: transcribed regions, CDS, 5'UTR, and 3'UTR, and 3 custom annotations from nORFs: all nORFs, nORF regions overlapping canonical CDS (norfs_altCDS), and nORF regions not overlapping canonical CDS (nORFs_noCDS). Results from common variant heritability show that nORFs have a similar CVE (6.0 ± 1.2, $P = 8 \times 10^{-4}$) to canonical CDS (5.5 ± 0.7, $P = 7 \times 10^{-7}$), and that this CVE in nORFs is concentrated in the subset that overlaps canonical CDS (9.2 ± 1.6, $P = 2 \times 10^{-4}$) rather than those which do not (3.2 ± 1.2, *NS*) **(Fig. 4A).** In low frequency variant heritability we found that again nORFs have a comparable enrichment (17.6 ± 3.0, $P = 1 \times 10^{-8}$) to canonical CDS (23.6 ± 2.0, $P = 1 \times 10^{-31}$), but that the difference between LFVE in nORFs overlapping CDS (30.7 ± 4.5, $P = 7 \times 10^{-11}$) and nORFs not overlapping CDS (2.3 ± 3.0, *NS*) was more pronounced **(Fig. 4B)**.

Higher ratios of LFVE/CVE have been associated with coding sequences, theorized to be due to natural selection keeping trait relevant variation at lower frequencies (Gazal et al. 2018). Here we found that canonical CDS showed the highest LFVE/CVE ratio (4.3x), with all nORFs (2.9x) and nORFs overlapping CDS (3.3x) showing high ratios but nORFs not overlapping CDS showing a ratio below one (0.7x) more comparable to that of 5'UTRs (0.5x) **(Fig. 4C)**. These results suggest that nORFs outside of coding regions may have less functional importance, but nORFs in the

altCDS and canonical CDS show a possible additive effect on heritability. From the results, however, we cannot distinguish between heritability coming from canonical CDS vs. nORFs. We attempt to disentangle the potential functional importance of nORFs from canonical CDS in the following analyses by examining specific nORF variant classes.

Mutability adjusted proportion of singletons (MAPS)

To examine the potential functional importance of nORFs separately from canonical CDS, we drew on variant frequencies from the Genome Aggregation Database (gnomAD) datasets, made up of 125,748 exome sequences and 15,708 genome sequences (Karczewski et al. 2020). Specifically, we used the mutability adjusted proportion of singletons (MAPS) score, which measures selection against classes of variants in a population (Lek et al. 2016; Karczewski et al. 2020). This measure is based on the principle that damaging classes of variants are kept at lower frequencies by natural selection. It compares the number of observed singletons for a particular variant class against the number of expected singletons under neutral selection, with a higher MAPS score being indicative of stronger selection against that variant class.

Variant bins for MAPS analysis were created using Ensembl's Variant Effect Predictor (VEP) (McLaren et al. 2016) to annotate the gnomAD exomes and genomes variants in the context of both nORFs and canonical genes. Selection patterns for nORFs are unclear when considering nORF annotations in isolation, likely due to consequences in canonical frames confounding the results (**Supplemental Fig. S1**). We therefore stratified our analysis by canonical consequence to examine the selection on nORF variants independently of their effect on canonical genes. We focus on seven annotations in canonical frames and five in nORFs, for a total of 35 variant bins which vary substantially in bin size **(Supplemental Fig. S2)**. For each bin, the MAPS score was calculated for the exomes **(Supplemental Table S3)** and genomes **(Supplemental Table S4)** dataset.

We observe that across most canonical consequences, variants annotated as stop lost or stop gained in nORFs show higher MAPS scores than the remainder of the canonical consequences, suggesting additional selective pressure on these variants **(Fig. 5)**. Several of the larger and therefore better powered bins showed significant differences in MAPS scores. For instance, for all exome dataset variants that are annotated as synonymous in canonical proteins, those which have stop lost or stop gained effects in nORFs show significantly higher MAPS scores than variants which fall outside of nORFs (both permuted $P < 1 \times 10^{-4}$) or are which are synonymous

in nORFs (both permuted $P < 1 \times 10^{-4}$; **Fig. 5A**). Similarly, when considering all canonical missense variants, those which have missense, stop lost, or stop gained effects in nORFs, all show significantly higher MAPS scores than variants which fall outside of nORFs (all permuted $P < 1 \times 10^{-4}$) or are which are synonymous in nORFs (all permuted $P < 1 \times 10^{-4}$; **Fig. 5A**). From the genomes, four of the five significant bins from the exomes analysis were also significantly different from variants falling outside of nORFs but not from synonymous nORF variants (**Fig. 5B; Supplementary Table S4**). We also observed in the genomes dataset that 5'UTR variants from canonical proteins showed significantly higher MAPS scores if they caused a stop gained effect in a nORF rather than a falling outside of nORFs (permuted $P = 1 \times 10^{-4}$) or synonymous in nORFs

(permuted $P = 7 \times 10^{-4}$; **Fig. 5B**). Overall, these results indicate selective pressure against deleterious nORF variants, suggesting that many of these variants may not be benign like current annotations would suggest.

Disease mutations in nORF contexts

Considering that stop lost and stop gained variants in nORFs show signals of negative selection, we investigated potential disease causing variants that could be due to these mutation types. We

first examined somatic cancer mutations from the Catalogue Of Somatic Mutations In Cancer (COSMIC) database (Tate et al. 2019). We annotated the 6.2 million coding and 19.7 million non coding somatic variants using VEP in the context of nORFs and then canonical annotations. Although COSMIC variant sets are expected to be dominated by passenger mutations, their functional interpretation is key to identifying the cancer causing genes and variants. We highlight 109K potential frameshift, stop gained, or stop lost variants in nORFs that have a less severe consequence in canonical genes **(Fig. 6A; Supplemental Table S5)**.

We then performed a similar analysis to annotate known human disease variants present in the Human Gene Mutation Database (HGMD) (Stenson et al. 2017) and ClinVar (Landrum et al. 2018) databases. We identified 1,852 variants from HGMD and 5,269 variants from ClinVar that are frameshift, stop gained, or stop lost variants in nORFs, but have less severe consequences in canonical genes **(Fig. 6B,C; Supplemental Table S6)**.

To create a short list of disease mutations most likely to have a nORF related cause, we further prioritized the COSMIC, HGMD and ClinVar disease mutations. Specifically, we identified top 20 cancer-associated genes with mutations with benign consequences in CDS but with deleterious consequences in the nORFs (**Supplemental Table S7**), 34 HGMD variants classified as disease causing **(Supplemental Table S8)** and 14 ClinVar variants classified as pathogenic or likely pathogenic **(Supplemental Table S9)** that have benign consequences in canonical annotations but stop loss or stop gain consequences in nORFs. We show an example where a theoretical synonymous disease variant has a stop gained effect on a nORF overlapping canonical CDS (**Fig. 6D,E**) which would normally be missed as a potential mechanism of pathogenicity.

Discussion

Following the advent of proteogenomics, ribosome profiling, and massively parallel sequencing studies, a key observation was that the entire genome has the potential to encode transcriptional and translational products. It was observed that noncanonical transcription and translation is not

bound by classical motifs for transcriptional start or stop sites, polyadenylation, AUG start codons, single CDS per transcript, or numerous other signatures associated with the conventional gene definitions. Beyond the lack of conventional signatures to identify them, there is no consensus on how nORFs should be classified, with research groups often focusing on specific types or sizes of nORFs. We have undertaken a systematic analysis to collate and reclassify these nORFs into an accessible dataset available to the wider community. This dataset was created with the goal of facilitating investigations into nORF signatures for transcription, translation, regulation, and function. In this study, we curated and annotated 194,407 nORFs with translation evidence from MS or ribosome profiling and assessed their functional significance using global genomic properties. We found signals of functional importance for nORFs from heritability of common and low-frequency variants, negative selection against classes of nORF variants, and disease mutations potentially explained by nORFs consequences.

Our observations demonstrate that nORFs show large heritability enrichments characteristic of CDS in both common and low frequency variation. We also show that these enrichments are vastly different when dividing nORFs into those that overlap canonical CDS and those that do not (**Fig. 4**). The nORF regions that do not overlap CDS show modest heritability enrichments, largely similar to other noncoding regions such as UTRs, which are not indicative of functional importance on the level of canonical CDS. By contrast, nORF regions overlapping canonical CDS show large heritability enrichments, higher than only canonical CDS. This suggests possible functional importance in nORFs and CDS showing an additive effect on heritability. An alternative explanation for these enrichments is that nORFs regions investigated are not causally adding heritability, but instead have been identified in a subset of highly enriched canonical CDS due to confounding factors such as gene expression, gene identity, or sequence composition. Future investigations might attempt to control for these factors by using within-gene control sites (i.e canonical CDS not overlapped by nORFs from only those genes with an overlapping nORF).

When considering both the canonical and nORF consequences of variants, MAPS scores reveal selection acting to keep nORF stop lost and stop gained variants at lower frequencies than other variants with the same canonical consequence (**Fig. 5**). This signal was significant in exonic regions from the exomes dataset and in 5'UTRs from the genomes dataset, possibly due to each being better powered in these respective areas. We also note that where there is signal of selection, the magnitude of that selection for stop gained variants in nORFs appears notably smaller than that of stop gained variants in canonical frames. This gap can be attributed to several possible reasons. Firstly, of the 194,407 nORFs in our dataset there are surely both false positive detections of translation, and detected translation of nORFs that do not create functional products, which would dilute the selection signal. Secondly, nORF products that are functional may have more specialized, cell specific, or context specific functions than canonical proteins. This could mean that the selection pressure against a 'true' set of stop gained variants from functional nORFs could be weaker than the pressure acting on canonical genes, and lead to a lower expected MAPS score. Nevertheless, these results suggest that nORF encoded protein products may have functional importance, motivating further analysis of disease mutations where they may be relevant.

Investigation of this showed that numerous variants in disease mutation databases could potentially have nORF related mechanisms of pathogenicity such as stop lost, stop gained, or frameshift mutations. We identified candidate HGMD disease mutations and ClinVar pathogenic/likely-pathogenic mutations with benign effects in canonical genes for which we believe nORF consequences should be considered as possible mechanisms of pathogenicity, similar to uORF perturbing variants recently found to be disease causing (Whiffin et al. 2020). These examples highlight the potential impact of annotating disease mutations for their nORF consequence.

Although this study has added valuable insights into non-canonical translation products, it does have limitations. First, some entries gathered in our dataset may be false positive detections of translation, or translation events of proteins with no function. This may dilute signals of functional importance, and should be kept in mind when using the dataset. The difference in entry count is clearly weighted towards sORFs (12% vs 88%), however the difference in sequence context is not quite as pronounced (22% vs 78%) due to the length distribution of OpenProt being substantially higher. We believe that both databases add substantial value to the nORFs dataset with the advantages of the sORFs database being that it focuses on Ribo-seq (OpenProt is primarily MS) and it does not use several of the constraints of OpenProt, which only considers ORFs above 30 codons and limits ORFs to ATG start codons and canonical transcripts. While we acknowledge that there may be concerns as to the sORF scoring methods, we have filtered out over 90% of the 2.1 million sORFs.org entries to the 209K with the filtering methods described in the manuscript (ORFscore with unique genomic mappings, longest ORF at shared stop sites, removing in-frame entries) which we believe substantially reduces false-positives. Despite these measures, one could certainly hypothesize that the filtering strategies of OpenProt make it more likely to contain functional nORFs. Our investigations of this possibility have shown that perhaps the opposite is true, however it is difficult to make conclusive statements because of lack of statistical power when analyzing the smaller OpenProt dataset by itself. In an early analysis of averaged heritability partitioned across 11 UK Biobank traits, split by OpenProt and sORFs we found that sORFs show more heritability than OpenProt entries, however the confidence intervals here are mostly overlapping (**Supplemental Fig. S4**). We also re-ran our MAPS analysis using only OpenProt entries and found that they did not show particularly strong selection (**Supplemental Fig. S5**), however this analysis is limited by lack of power as shown by the large confidence intervals. Based on these results, we find it unlikely there are strong functional signals from OpenProt being contaminated by entries from sORFs.org. In addition, the calculation of

heritability and MAPS scores based on the source of the database suffers from statistical noise, making it difficult to draw meaningful conclusions, compared to the better powered joint analyses.

Second, it is by no means a comprehensive catalogue of translation products in the human genomes; more nORFs are sure to be found as more investigations of translation products are carried out. Third, heritability enrichment estimates for nORF regions do not directly estimate the contribution of nORFs, but of any causal heritability signals in the region investigated. This is particularly relevant for the nORF regions that overlap canonical CDS where the relative contribution of these factors cannot reliably be distinguished. Despite this caveat the increase in heritability enrichment is an interesting finding that suggests a possible contribution of nORFs to the heritability of traits and disease. Last, for disease mutations potentially explained by nORFs, we caution that individual causal mechanisms cannot be confidently determined without weighing original translation evidence associated with the nORF, other possible mechanisms (e.g. regulatory), and potential follow-up functional analysis. Nevertheless, these disease mutation examples demonstrate the potential of annotating variants for their consequence in nORFs to explain their pathogenicity.

In this work, we developed a consistent, comprehensive, and accessible nORF resource that will aid future investigations into non-canonical translation products for the community. We have used this dataset to make insights into the potential functional impacts of nORFs in the human genome. We have shown heritability enrichments associated with nORFs, particularly those overlapping canonical CDS. We then demonstrated selective pressure acting on potentially deleterious nORF variants, suggesting their potential functional importance. Finally, we annotated disease mutations with nORF consequences, demonstrating a potential to uncover plausible mechanisms and to generate hypothesis of their pathogenicity. In future investigations this technique may be a valuable addition for discerning pathogenic mechanisms for rare disease diagnosis or in common

disease phenotypes. If these investigations are successful nORFs could be a set of new potential drug targets for disease treatment.

Methods

Selection of sources for evidence of nORFs

Three existing databases with entries that qualify as nORFs were considered for inclusion in the nORFs dataset: OpenProt (Brunet et al. 2019), sORFs.org (Olexiouk et al. 2018), and SmProt (Hao et al. 2018). SmProt was not used due to inconsistencies in data (e.g. incorrect genomic coordinate annotations) and lack of details in their methods to reanalyse the data, specifically in regards to their MS evidence (Olexiouk et al. 2018). By contrast, OpenProt and sORFs.org have shown commitment to providing consistent, verifiable, and maintained data, and were therefore used as the main sources for the nORFs dataset.

OpenProt (Release 1.3) predicts all possible ORFs with an ATG start codon and a minimum length of 30 codons that map to an Ensembl (Zerbino et al. 2018) or RefSeq (O'Leary et al. 2016) transcript. They identified 607,456 alternate ORFs (altORFs) that are neither canonical ORFs, nor an isoform of those ORFs, but in noncoding regions or an alternate frame to canonical CDS. Although OpenProt maps to both Ensembl and RefSeq transcripts, we focus exclusively on the Ensembl annotations for compatibility with the sORFs.org dataset and other downstream analyses. From the altORFs mapped to Ensembl transcripts, we consider the 26,480 altORFs with translation evidence from MS (21,708), ribosome profiling (5,059), or both (398).

The sORFs.org database (downloaded April 30, 2019) uses notably different inclusion criteria, annotating 'sORFs' with translation evidence from 43 human ribosome profiling experiments, then adding MS evidence found in publicly available datasets. The sORFs are defined as ORFs

between 10 and 100 codons using any of four start codons: 'ATG', 'CTG', 'TTG', or 'GTG', and are not restricted to known transcripts.

Curation of nORFs

The curation steps we performed to create a nORF dataset are detailed in **Fig. 1**. The final dataset that we created a) contains only nORFs with translation evidence from either MS or ribosome profiling b) contains no duplicate or highly similar entries and c) contains only ORFs clearly distinct from currently annotated canonical proteins.

We used 607,456 predicted altORFs from OpenProt and filtered to the 26,480 entries with MS or ribosome profiling evidence of translation. From over 2.1 million sORFs.org entries with 'good' or 'extreme' ORFscore (Bazzini et al. 2014), 502,056 entries with unique genomic mappings were extracted **(Fig. 1A)**. The next step involved processing similar entries in the sORFs.org dataset that shared the same stop site and amino acid sequences up to differing start sites. A characteristic example is shown in **Fig. 1B** where in an alternative frame of the final coding exon of the *MRPS21* gene, sORFs.org provides evidence for five small ORFs sharing the same end site and differing only by their start site. This is common in the sORFs.org dataset because of the ambiguity in ribosome profiling experiments to identify the correct translation start site, unless specifically using methods that search for them (e.g. ribosome profiling with antibiotics used to trap newly initiated ribosomes at start codons) (Olexiouk et al. 2018; Weaver et al. 2019). Although ideally the correct start site(s) would be identified through experiments, this data is not currently available. For consistency and simplicity, we have selected the longest ORF in these cases, which may not always represent the true translated ORF, but will always encompass all ORFs identified at these sites. We emphasize this ambiguity in the correct start site as an important limitation to be kept in mind when using the dataset. In all, the selection of the longest ORF at ambiguous start sites further reduced extracted sORFs.org entries to 209,543.

Next, the OpenProt and sORFs.org datasets were merged, 1,028 redundant entries between the datasets were removed, and 1,976 cases of ambiguous start sites between the two datasets were resolved by again taking the longest ORF, resulting in a merged total of 233,021 entries. The small number of overlapping or similar entries between the two datasets can be partly attributed to different inclusion criteria for ORFs between the databases (i.e. ORF length, start codon, transcript requirement) and the main source of entries (sORFs from ribosome profiling and OpenProt predominantly from MS).

Finally, we separated all entries that were in-frame with canonical CDS, as the translation evidence from these entries cannot be unambiguously resolved as to whether they are from a canonical protein product or an independent nORF embedded within a canonical protein. We identified 38,614 such entries and removed them, leaving a total of 194,407 entries in the final nORFs dataset. An example case is shown in **Fig. 1C** where two small ORFs overlap the CDS of the *RICA* gene. One of these ORFs is in the same frame as the *RICA* CDS and was therefore filtered out, whereas the second ORF is in a different frame and retained in the dataset. Following this final curation step all entries in the nORF dataset that overlap canonical CDS are in a different frame from and do not share amino acid sequence with that CDS.

Annotation of nORFs

We annotated each nORF with reference to human GENCODE (v30) gene annotations (Frankish et al. 2019). The annotation categories included nORFs mapping to UTRs or CDS of protein coding transcripts, ncRNAs, or intergenic regions. When multiple annotations were possible, due to multiple transcripts in a region, annotations were prioritized by first selecting full overlaps with protein coding transcripts, particularly those that overlap canonical CDS in an alternative reading frame (altCDS), followed by full overlaps with ncRNA transcripts, then by partial transcript

overlaps, and finally intronic or intergenic regions. Our detailed prioritization summary is shown in **Supplemental Fig. 3**.

Using GENCODE 34 (latest version) our pipeline identifies 194,291 rather than 194,407 nORFs, meaning that between releases 30 and 34, 116 nORFs became part of canonical CDS as newly identified genes or as part of new coding transcripts of existing genes. We find it encouraging that some nORFs are becoming canonical CDS and plan to regularly update our GENCODE reference in future iterations of the nORFs database.

Database and web platform

To reduce the threshold of accessibility, databases need to be accessible with minimal requirements of tools or prior knowledge. We therefore built an online platform with Representational State Transfer (REST) application programming interface (API) functionality. This online platform acts as an entry and lookup point for individual entries, while the REST API is feature compatible with existing bioinformatics pipelines. We made the curated and annotated GRCh38 raw dataset available in BED and GTF format as well as a downloadable nORFs.org UCSC track. Considering reproducible research guidelines, we used git as a versioning tool and uploaded the repository to GitHub under an MIT license (https://github.com/PrabakaranGroup/nORFs.org).

Stratified LD score regression (S-LDSC) heritability analysis

As applied previously (Gazal et al. 2018), we obtained summary statistics for 40 heritable, complex UK Biobank (Bycroft et al. 2018) traits (downloaded from https://data.broadinstitute.org/alkesgroup/UKBB/UKBB_409K/) that were restricted to 409 K individuals with UK ancestry. We then generated an LD reference panel for UK ancestry to match the summary statistics with 3,567 UK10K (The UK10K Consortium 2015, 10) whole-genome sequencing (WGS) samples from the ALSPAC and TWINSUK cohorts.

With these inputs, we analyzed a total of 177 genomic annotations, each corresponding to a defined set of variants, for their heritability enrichment. Of the 177, 163 are together known as the previously described baseline-LF model (Gazal et al. 2018). We added to the analysis 14 custom annotations, from seven functional annotations doubled for common variants and low frequency variants. Of these seven, three custom annotations were nORF related: one for all nORFs, and 2 in which nORFs were split at the variant level to those regions which overlap canonical CDS (norfs_altCDS), and those which do not (nORFs_noCDS). The remaining 4 were canonical annotations from GENCODE: transcribed regions, CDS, 5'UTRs, and 3'UTRs. It should be noted that similar annotations appear to be already present in the baseline-LF model, but they were generated from a different reference set than our nORFs (UCSC 2013) (Gusev et al. 2014) and their 'Coding' annotation contains UTRs, which ours does not.

For the baseline-LD functional annotations and our custom annotations, we calculated common variant enrichment (CVE) and low frequency variant enrichment (LFVE) for each of the 40 UK Biobank traits. CVE is the proportion of common heritability (h^2_C) divided by the proportion of common single nucleotide polymorphisms (SNPs) in the annotation, while LFVE is proportion of low-frequency heritability (h^2_{LF}) divided by the proportion of low frequency SNPs in the annotation:

$$CVE = Prop(h^{2}_{C})/Prop(common\ SNPs)$$

$$LFVE = Prop(h^{2}_{LF})/Prop(low\ frequency\ SNPs)$$

Meta-analysis of results was conducted using random-effects meta-analyses in the rmeta package on 27 independent traits (Gazal et al. 2018), indicated in **Supplemental Table S1**. All standard errors were computed using a block jackknife procedure (Bulik-Sullivan et al. 2015).

Mutability adjusted proportion of singletons (MAPS) analysis

We calculated MAPS with gnomAD genomes and exomes by using publicly available code at https://github.com/macarthur-lab/gnomad_lof. We modified the code to include variant bins based on both nORF consequences and canonical consequences, rather than only canonical consequences. We selected five nORF consequences of interest: missense, synonymous, stop lost, stop gained, and noncoding (intergenic + upstream gene + downstream gene) and 7 canonical consequences of interest: missense, synonymous, ncRNA, 5'UTR, 3'UTR, intronic and intergenic. For each of these 35 (5x7) bins, MAPS calibrated expected variant frequencies to account for 1 surrounding base of context and CpG methylation, two factors known to influence the mutability of base pairs (Lek et al. 2016). The transformation between variant frequencies and the expected proportion of singletons was regressed against the observed proportion of synonymous variants in canonical proteins. As the MAPS score given to variant classes is a relative metric, this means that synonymous variants in canonical proteins were set as 0 and higher scores reflected more negative selection. We reported MAPS scores for bins with at least 100 variants in the gnomAD exomes or genomes dataset respectively.

P-values were calculated using a bootstrapping approach as applied previously (Whiffin et al. 2020). For a given bin with *n* variants, *n* variants were randomly sampled with replacement and used to calculate MAPS for two bins or interest: bin A and bin B. This was repeated over 10,000 permutations with the P-value being the proportion of permutations where MAPS of bin B was less than MAPS of bin A.

Variant annotation

Variant annotation was carried out using version 96 of VEP (McLaren et al. 2016) to investigate the consequences of variants in the context of canonical frames and nORFs. Variant sets were obtained for annotation as VCFs. These included gnomAD genomes and exomes (release 2.1.1)

(Karczewski et al. 2020), HGMD (pro release 2019.2) (Stenson et al. 2017), ClinVar (release 2019 0708) (Landrum et al. 2018), and COSMIC coding and noncoding mutations (v89) (Tate et al. 2019). Each set of variants was annotated for their most severe consequence as defined by VEP with respect to a) canonical gene annotations, corresponding to GENCODE 30 in GRCh38 or GENCODE 30 lifted over to GRCh37 and b) nORF annotations provided as a custom GTF in the appropriate genome assembly.

When examining possible disease mutations that could be explained by nORF consequences, we first filtered variants from the disease mutations databases (COSMIC, HGMD, and ClinVar) to remove those with strongly deleterious annotations in canonical proteins (i.e. essential splice, frameshift, stop gained, stop lost, start lost). We then further filtered these variant sets to those with possible pathogenic consequences in nORFs (stop lost, stop gained, and frameshift).

Software availability

The code used to curate, annotate, and analyze the nORFs dataset is publicly available at https://github.com/PrabakaranGroup/nORF-data-prep and uploaded as Supplemental Code File 1 and 2. To share the nORFs dataset we have also created nORFs.org, an open source platform for the nORFs dataset with implementation available at https://github.com/PrabakaranGroup/nORFs.org and uploaded as Supplementay Code File 3. A UCSC track download is also provided on the nORFs.org API page.

Competing Interest Statement

SP is a cofounder of NonExomics.

Acknowledgments

We would like to thank RosettaHuB (https://rosettahub.com) for helping us to build applications using Amazon Web Services. SP is funded by the Cambridge-DBT lectureship; CE is funded by

Dr. Manmohan Singh scholarship; RK is funded by BBSRC Fellowship; NM is funded by Govt of India and Trinity Barlow fellowship. We thank the anonymous reviewers and the editor for their critical and thoughtful comments.

*Author contributions***:** MN did the nORF classification with help from CE and NM. RK built the database and web platform. MN did the MAPS, S-LDSC and variant analysis. MN and SP interpreted the data and wrote the manuscript with assistance from RK and CE. MH, DC, MM curated and provided the HGMD dataset. SP designed and supervised the project.

References

Albuquerque JP, Tobias-Santos V, Rodrigues AC, Mury FB, da Fonseca RN. 2015. small ORFs: A new class of essential genes for development. *Genet Mol Biol* **38**: 278–283.

Andrews SJ, Rothnagel JA. 2014. Emerging evidence for functional peptides encoded by short open reading frames. *Nat Rev Genet* **15**: 193–204.

Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming ES, Vejnar CE, Lee MT, Rajewsky N, Walther TC, et al. 2014. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J* **33**: 981–993.

Brunet MA, Brunelle M, Lucier J-F, Delcourt V, Levesque M, Grenier F, Samandi S, Leblanc S, Aguilar J-D, Dufour P, et al. 2019. OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucleic Acids Res* **47**: D403–D410.

Brunet MA, Levesque SA, Hunting DJ, Cohen AA, Roucou X. 2018. Recognition of the polycistronic nature of human genes is critical to understanding the genotype-phenotype relationship. *Genome Res* **28**: 609–624.

Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson N, Daly MJ, Price AL, Neale BM. 2015. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**: 291–295.

Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, et al. 2018. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**: 203.

Chen, J. et al., 2020. Pervasive functional translation of noncanonical human open reading frames. *Science*, 367(6482), pp.1140–1146.

Erady C, Chong D, Meena N, Puntambekar S, Chauhan R, Umrania Y, Andreani A, Nel J, Wayland MT, Pina C, et al. 2019. Translational products encoded by novel ORFs may form protein-like structures and have biological functions. *bioRxiv*. http://biorxiv.org/lookup/doi/10.1101/567800 (Accessed July 28, 2019).

Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, Anttila V, Xu H, Zang C, Farh K, et al. 2015. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**: 1228–1235.

Firth AE, Brierley I. 2012. Non-canonical translation in RNA viruses. *J Gen Virol* **93**: 1385–1409.

Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al. 2019. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**: D766–D773.

Gazal S, Finucane HK, Furlotte NA, Loh P-R, Palamara PF, Liu X, Schoech A, Bulik-Sullivan B, Neale BM, Gusev A, et al. 2017. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nat Genet* **49**: 1421–1427.

Gazal S, Loh P-R, Finucane HK, Ganna A, Schoech A, Sunyaev S, Price AL. 2018. Functional architecture of low-frequency variants highlights strength of negative selection across coding and noncoding annotations. *Nat Genet* **50**: 1600.

Gusev A, Lee SH, Trynka G, Finucane H, Vilhjálmsson BJ, Xu H, Zang C, Ripke S, BulikSullivan B, Stahl E, et al. 2014. Partitioning Heritability of Regulatory and Cell-TypeSpecific Variants across 11 Common Diseases. *Am J Hum Genet* **95**: 535–552.

Hao Y, Zhang L, Niu Y, Cai T, Luo J, He S, Zhang B, Zhang D, Qin Y, Yang F, et al. 2018. SmProt: a database of small proteins encoded by annotated coding and noncoding RNA loci. *Brief Bioinform* **19**: 636–643.

Hellens RP, Brown CM, Chisnall MAW, Waterhouse PM, Macknight RC. 2016. The Emerging World of Small ORFs. *Trends Plant Sci* **21**: 317–328.

Jagannathan NS, Meena N, Bhayankaram KP, Prabakaran S. 2019. Proteins encoded by Novel ORFs have increased disorder but can be biochemically regulated and harbour deleterious mutations. *bioRxiv*. http://biorxiv.org/lookup/doi/10.1101/562835 (Accessed July 28, 2019).

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, et al. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**: 434–443.

Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, et al. 2018. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* **46**: D1062–D1067.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**: 285–291.

Ma J, Ward CC, Jungreis I, Slavoff SA, Schwaid AG, Neveu J, Budnik BA, Kellis M, Saghatelian A. 2014. Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. *J Proteome Res* **13**: 1757–1765.

McGillivray P, Ault R, Pawashe M, Kitchen R, Balasubramanian S, Gerstein M. 2018. A comprehensive catalog of predicted functional upstream open reading frames in humans. *Nucleic Acids Res* **46**: 3326–3338.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. 2016. The Ensembl Variant Effect Predictor. *Genome Biol* **17**: 122.

Miravet-Verde S, Ferrar T, Espadas-García G, Mazzolini R, Gharrab A, Sabido E, Serrano L, Lluch-Senar M. 2019. Unraveling the hidden universe of small proteins in bacterial genomes. *Mol Syst Biol* **15**. https://www.embopress.org/doi/abs/10.15252/msb.20188290 (Accessed August 3, 2019).

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**: D733–D745.

Olexiouk V, Van Criekinge W, Menschaert G. 2018. An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res* **46**: D497–D502.

Plaza S, Menschaert G, Payre F. 2017. In Search of Lost Small Peptides. *Annu Rev Cell Dev Biol* **33**: 391–416.

Prabakaran S, Hemberg M, Chauhan R, Winter D, Tweedie-Cullen RY, Dittrich C, Hong E, Gunawardena J, Steen H, Kreiman G, et al. 2014. Quantitative profiling of peptides from RNAs classified as noncoding. *Nat Commun* **5**: 5429.

Saghatelian A, Couso JP. 2015. Discovery and characterization of smORF-encoded bioactive polypeptides. *Nat Chem Biol* **11**: 909–916.

Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, Hussain M, Phillips AD, Cooper DN. 2017. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet* **136**: 665–677.

Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, et al. 2019. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **47**: D941–D947.

The UK10K Consortium. 2015. The UK10K project identifies rare variants in health and disease. *Nature* **526**: 82–90.

The UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47**: D506–D515.

Weaver J, Mohammad F, Buskirk AR, Storz G. 2019. Identifying Small Proteins by Ribosome Profiling with Stalled Initiation Complexes ed. J. Vogel. *mBio* **10**. http://mbio.asm.org/lookup/doi/10.1128/mBio.02819-18 (Accessed July 20, 2019).

Whiffin N, Karczewski KJ, Zhang X, Chothani S, Smith MJ, Evans DG, Roberts AM, Quaife NM, Schafer S, Rackham O, et al. 2020. Characterising the loss-of-function impact of 5' untranslated region variants in 15,708 individuals. *Nat Commun* **11**: 2523.

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, et al. 2018. Ensembl 2018. *Nucleic Acids Res* **46**: D754–D761.

Legends

Figure 1. Flow chart for the curation of the nORFs dataset. (A) Steps illustrating the workflow to curate nORFs entries. From OpenProt, all predicted human altProts were filtered to entries with MS or ribosome profiling evidence (26,480). From sORFs.org all human sORFs with an ORFscore of good or extreme were filtered to unique entries (502,056), and then summarized to the longest ORF at sites with multiple ORFs. Entries were then merged, the longest ORF was selected at multiple ORF sites and in-frame entries were removed, leaving a total of 194,407 nORFs in the final dataset. (B) An example of selecting the longest ORF for 5 small ORFs in an alternative frame of the final coding exon of the *MRPS21* gene. In these cases where the ORFs share the same end site and differ only by their start site, we retain the longest ORF, indicated by the orange arrow and remove the shorter ORFs, indicated by the red cross. (C) An example of removing in-frame entries where two small ORFs overlap the CDS of the *RICA* gene. The ORF in the same frame as the *RICA* CDS is removed from the dataset as indicated by the red cross, whereas the second ORF in a different frame is retained in the dataset, indicated by the orange arrow.

Figure 2. Overview of Platform. The nORFs.org platform contains 6 individual pages (3 shown above) that introduce the platform, methods and nORF entries. The nORF detail page (right) is divided into 3 sections: the meta data section includes the unique identifier, genomic position and experimental sources, the (biodalliance) genome browser displays genes, repeats, conservation and epigenetic information such as DNase I binding sites, and Histone modifications (H3K4me1-me3), and the protein sequence section which can be utilized for biostatistical pipelines to display variants, topology and alternative splicing.

Figure 3. nORF genomic annotations. (A) Schematic of common nORF locations with respect to a typical protein coding gene and a ncRNA. (B) Number of nORFs per genomic annotation. (C) Distribution of amino acid length of major categories of nORF annotations and canonical UniprotKB/Swiss-prot proteins **(The UniProt Consortium 2019)**. (D) Distribution of nORFs and canonical CDS from GENCODE throughout human chromosomes. Canonical and nORF scales are not proportionate.

Figure 4. Meta-analysis of heritability partitioned across 27 UK Biobank traits for nORF regions.
Heritability enrichment was compared for canonical gene annotation from GENCODE vs nORF annotation for (A) common variation enrichment (CVE), defined as the proportion of common variant heritability explained by the annotation divided by the proportion of common variants in the annotation (B) lowfrequency variant enrichment (LFVE), defined similarly to CVE and (C) the LFVE/CVE ratio. Higher enrichments suggest more functional importance in the studied traits and diseases.

Figure 5. nORF stop lost and stop gained variants show signals of negative selection. The mutabilityadjusted proportion of singletons (MAPS) was calculated for 35 variant bins of SNVs from gnomAD (A) exomes and (B) genomes. Higher values indicate an enrichment of lower frequency variants, suggesting negative selection. The canonical annotation of the bin is indicated along the x-axis, while the nORF annotation is indicated by colour. Noncoding refers to variants falling outside of nORF regions. Dotted lines correspond to results from bins of only canonical annotations previously reported (Karczewski et al. 2020). (*) permuted P adj < 0.05 vs noncoding bin and < 0.05 vs synonymous bin with the same canonical consequence.

Figure 6. Reinterpreting COSMIC, HGMD and ClinVar mutations in the context of nORFs. The canonical consequence and nORF consequence of (A) 109K somatic cancer mutations from COSMIC, (B) 1.8K disease mutations from HGMD, and (C) 5.3K disease mutations from ClinVar. Bins with 10 or fewer variants not shown. These mutations would likely be interpreted as benign or missense in canonical genes but may have more severe consequences in nORFs. (D) A theoretical example of a disease variant that results in a synonymous mutation in canonical CDS but (E) a stop gain mutation in a nORF from an alternative reading frame.

Figure 1

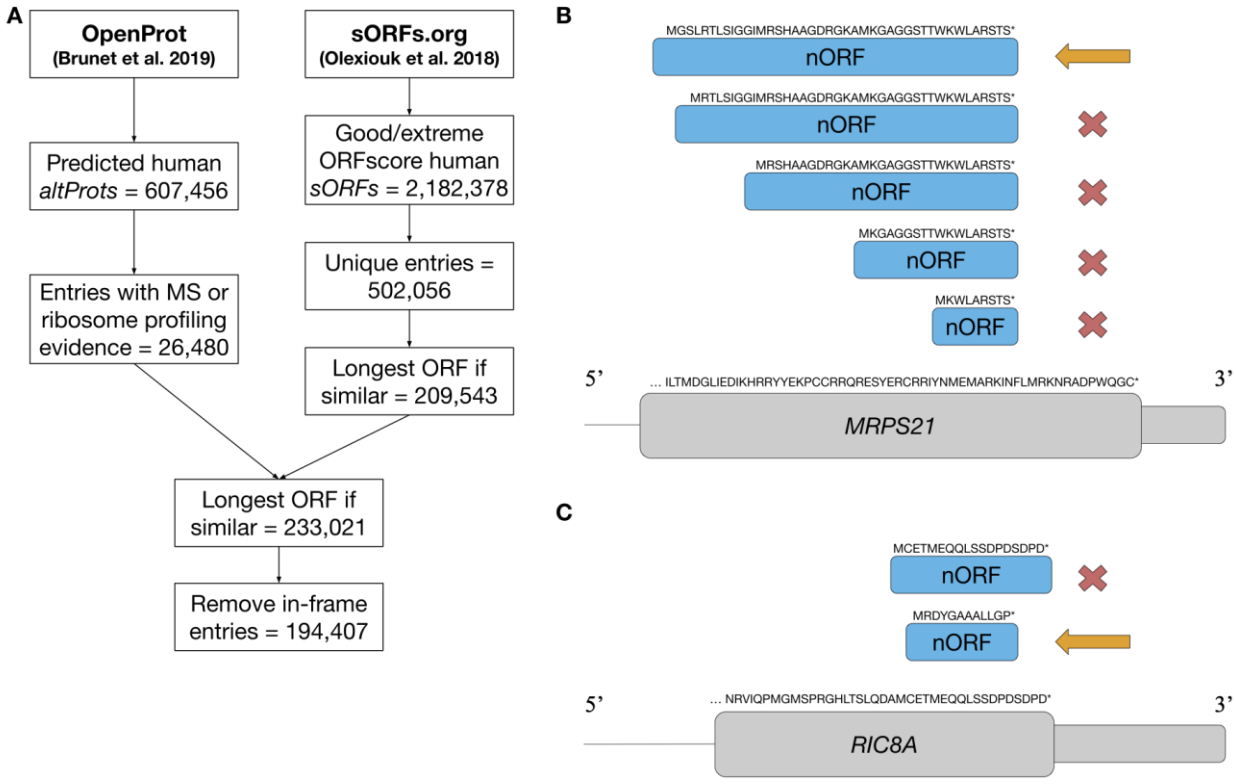

A
OpenProt
(Brunet et al. 2019)
Predicted human
altProts = 607,456
Entries with MS or
ribosome profiling
evidence = 26,480
sORFs.org
(Olexiouk et al. 2018)
Good/extreme
ORFscore human
sORFs = 2,182,378
Unique entries =
502,056
Longest ORF if
similar = 209,543
Longest ORF if
similar = 233,021
Remove in-frame
entries = 194,407
B
MGSLRTLSIGGIMRSHAAGDRGKAMKGAGGSTTWKWLARSTS*
nORF
MRTLSIGGIMRSHAAGDRGKAMKGAGGSTTWKWLARSTS*
nORF
MRSHAAGDRGKAMKGAGGSTTWKWLARSTS*
nORF
MKGAGGSTTWKWLARSTS*
nORF
MKWLARSTS*
nORF
5'
... ILTMDGLIEDIKHRRYYEKPCCRRQRESYERCRRIYNMEMARKINFLMRKNRADPWQGC*
3'
MRPS21
C
MCETMEQQLSSDPDSDPD*
nORF
MRDYGAAALLGP*
nORF
5'
... NRVIQPMGMSPRGHLTSLQDAMCETMEQQLSSDPDSDPD*
3'
RIC8A

Figure 2

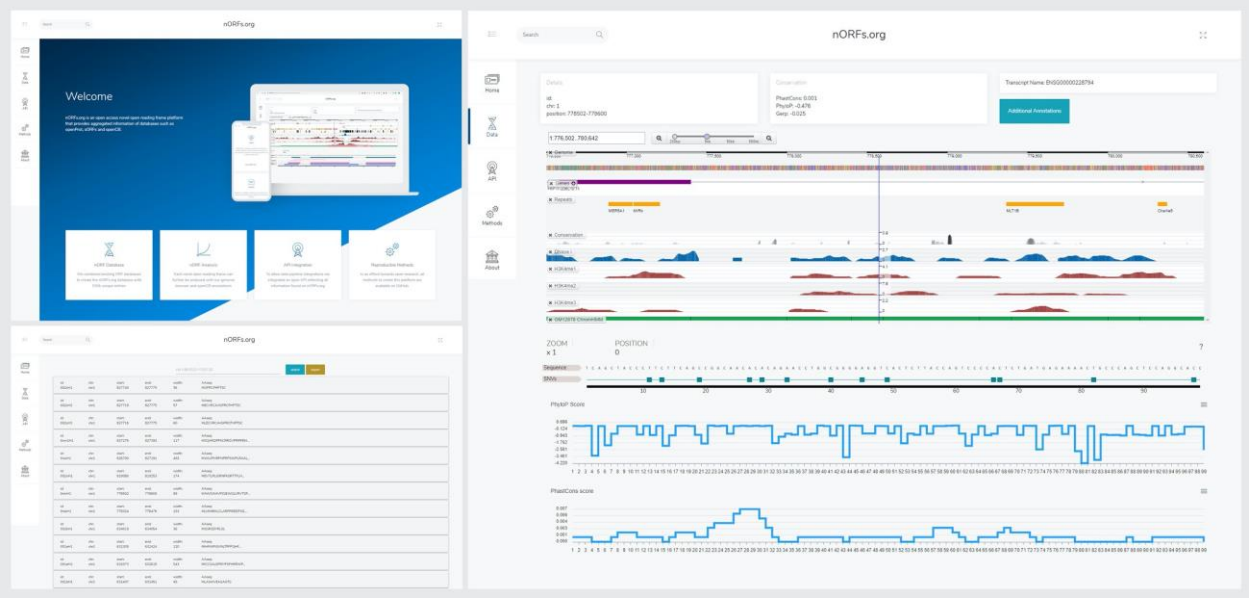


Figure 3

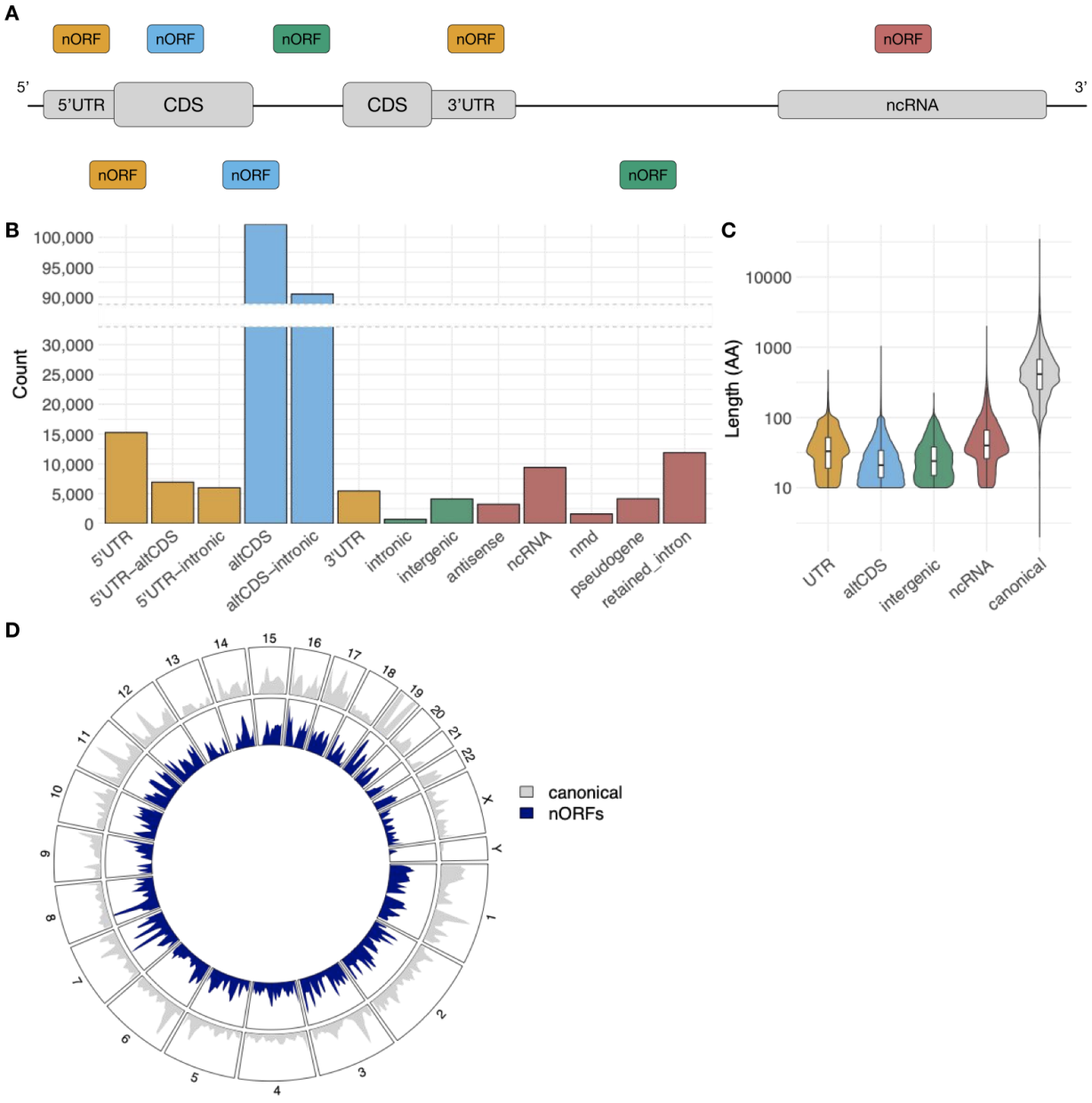

A
nORF
nORF
nORF
nORF
nORF
5'
5'UTR
CDS
CDS
3'UTR
ncRNA
3'
nORF
nORF
nORF
B
Count
100,000
95,000
90,000
30,000
25,000
20,000
15,000
10,000
5,000
0
5'UTR
5'UTR–altCDS
5'UTR–intronic
altCDS
altCDS–intronic
3'UTR
intronic
intergenic
antisense
ncRNA
nmd
pseudogene
retained_intron
C
Length (AA)
10000
1000
100
10
UTR
altCDS
intergenic
ncRNA
canonical
D
canonical
nORFs

Figure 4

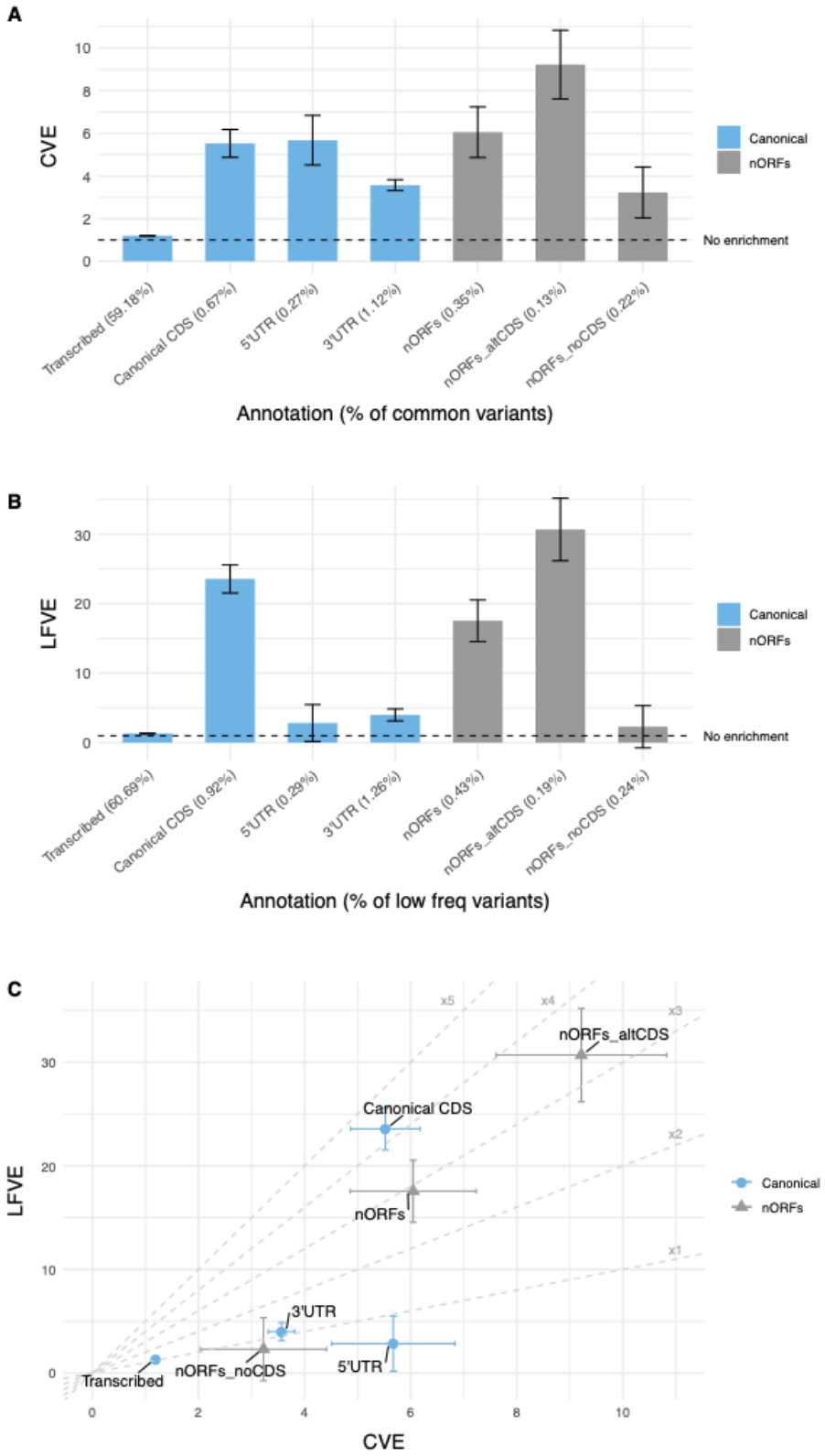

A
CVE
Canonical
nORFs
No enrichment
Transcribed (59.18%)
Canonical CDS (0.67%)
5'UTR (0.27%)
3'UTR (1.12%)
nORFs (0.35%)
nORFs_altCDS (0.13%)
nORFs_noCDS (0.22%)
Annotation (% of common variants)
B
LFVE
Canonical
nORFs
No enrichment
Transcribed (60.69%)
Canonical CDS (0.92%)
5'UTR (0.29%)
3'UTR (1.26%)
nORFs (0.43%)
nORFs_altCDS (0.19%)
nORFs_noCDS (0.24%)
Annotation (% of low freq variants)
C
LFVE
CVE
x5
x4
x3
x2
x1
nORFs_altCDS
Canonical CDS
nORFs
3'UTR
5'UTR
nORFs_noCDS
Transcribed
Canonical
nORFs

Figure 5

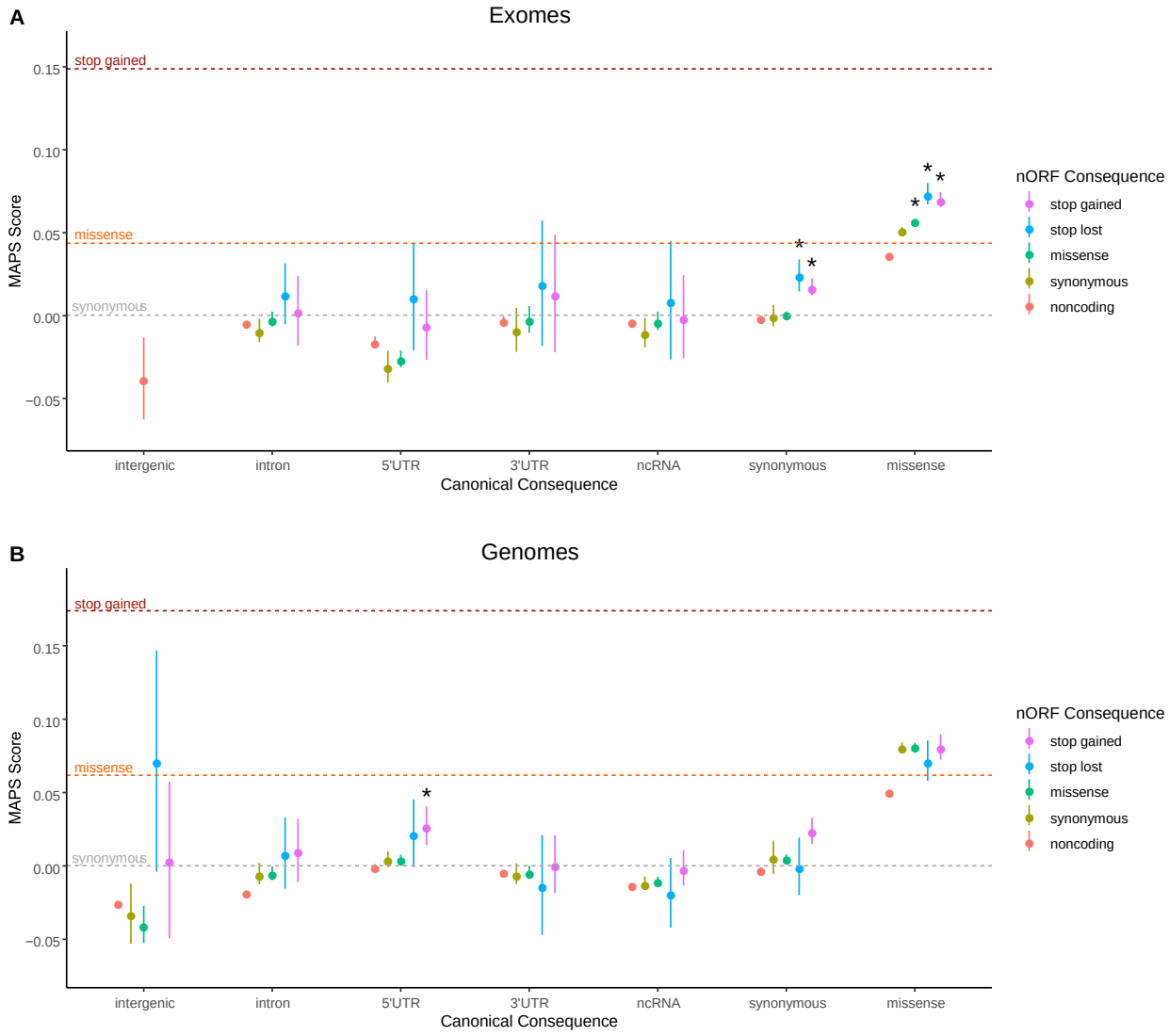
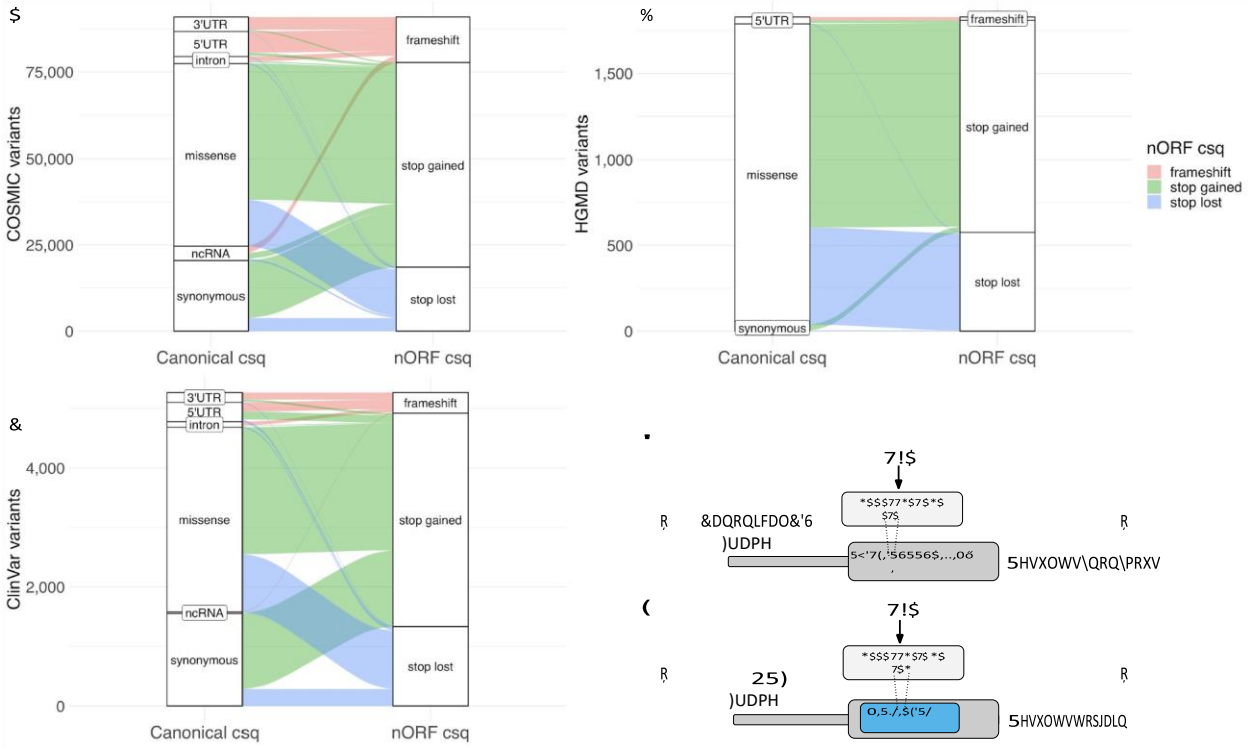

A
Exomes
stop gained
missense
synonymous
MAPS Score
0.15
0.10
0.05
0.00
−0.05
intergenic
intron
5'UTR
3'UTR
ncRNA
synonymous
missense
Canonical Consequence
nORF Consequence
stop gained
stop lost
missense
synonymous
noncoding
B
Genomes
stop gained
missense
synonymous
MAPS Score
0.15
0.10
0.05
0.00
−0.05
intergenic
intron
5'UTR
3'UTR
ncRNA
synonymous
missense
Canonical Consequence
nORF Consequence
stop gained
stop lost
missense
synonymous
noncoding

Figure 6


$
COSMIC variants
75,000
50,000
25,000
0
3'UTR
5'UTR
intron
missense
ncRNA
synonymous
frameshift
stop gained
stop lost
Canonical csq
nORF csq
%
HGMD variants
1,500
1,000
500
0
5'UTR
missense
synonymous
frameshift
stop gained
stop lost
Canonical csq
nORF csq
nORF csq
frameshift
stop gained
stop lost
&
ClinVar variants
4,000
2,000
0
3'UTR
5'UTR
intron
missense
ncRNA
synonymous
frameshift
stop gained
stop lost
Canonical csq
nORF csq
.
7!$
*$$$77*7*$
7
R
&DQRQLFDO&'6
)UDPH
5HVXOWV\QRQ\PRXV
(
7!$
*$$$77*7 *$
7$*
R
25)
)UDPH
5HVXOWVWRSJDLQ

SUPPLEMENTARY MATERIAL

A platform for curated products from novel Open Reading Frames (nORFs) prompts reinterpretation of disease variants.

Matthew DC Neville[1†], Robin Kohze[1†], Chaitanya Erady[1], Narendra Meena[2], Matthew Hayden[3], David N. Cooper[3], Matthew Mort[3], Sudhakaran Prabakaran[1,2,4*]

[1]Department of Genetics, University of Cambridge, Downing Site, CB2 3EH, UK

[2]Department of Biology, Indian Institute of Science Education and Research, Pune, Maharashtra, 411008, India

[3]Institute of Medical Genetics, Cardiff University, Heath Park, Cardiff CF14 4XN, UK

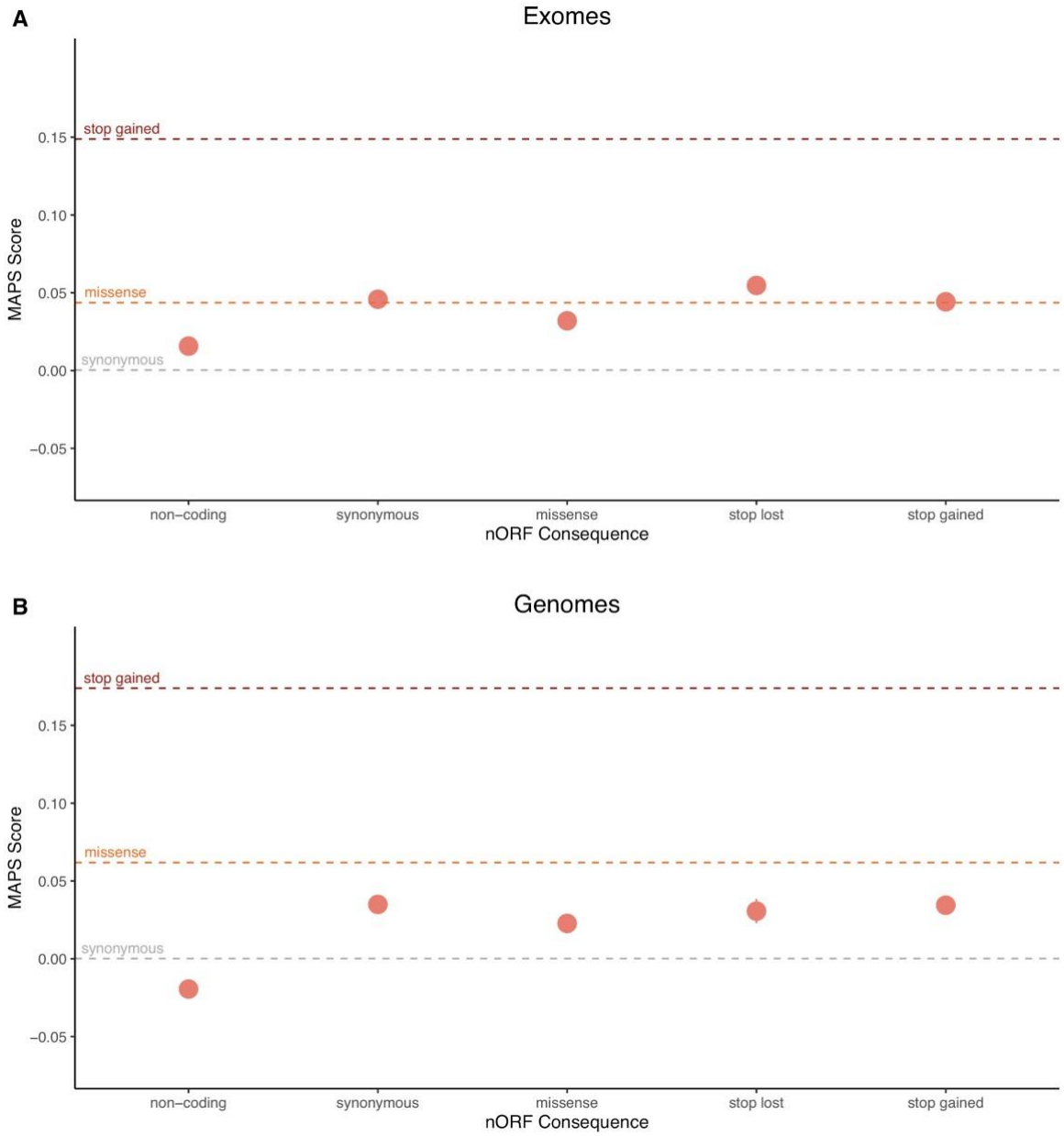
[4]St Edmund's College, University of Cambridge, CB3 0BN, UK

*Corresponding author, email: sp339@cam.ac.uk.

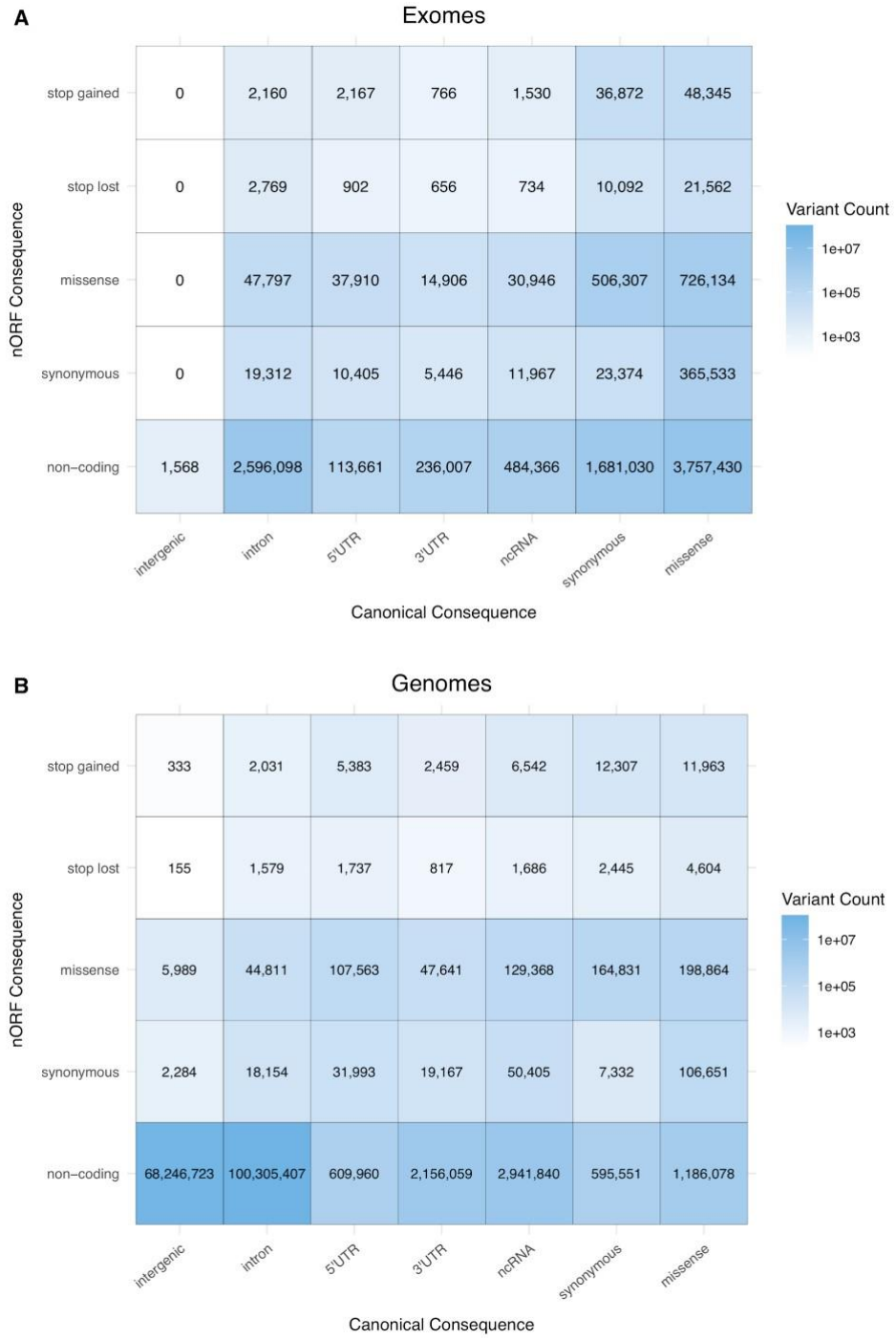
[†]These authors contributed equally to the work.

Table of Contents

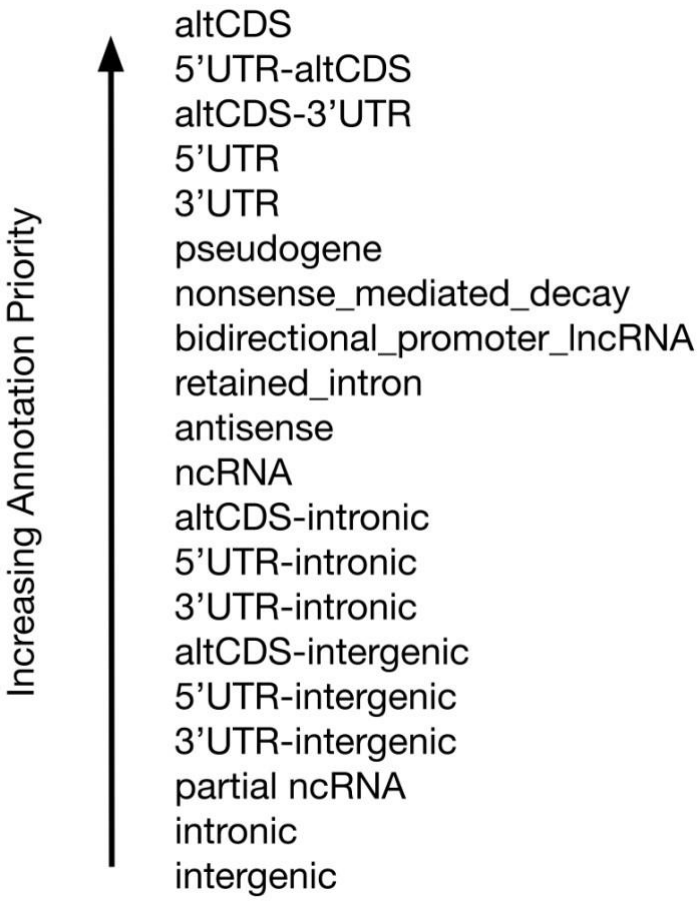
Supplemental Figures



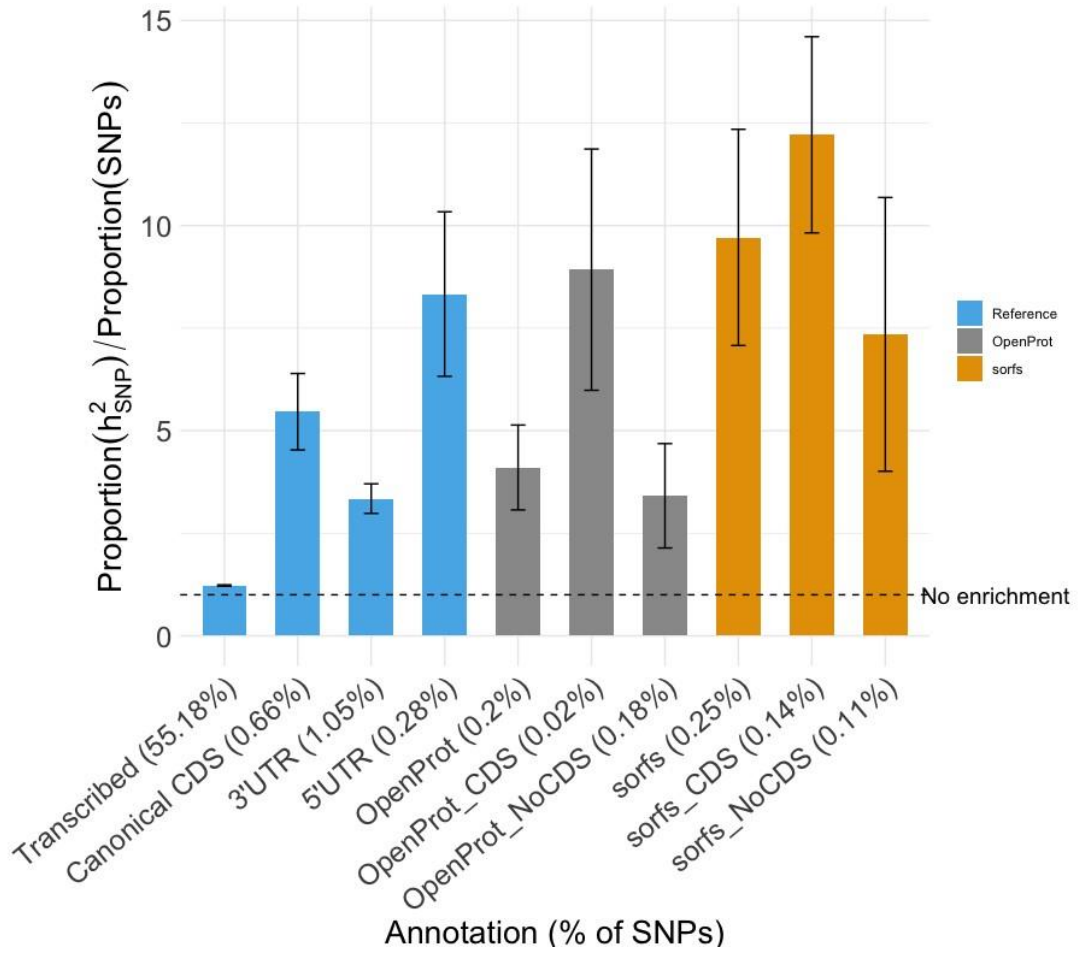
Supplemental Figure S1. Unclear selection signatures when only considering nORF consequences. The mutability-adjusted proportion of singletons (MAPS) is shown across functional categories for SNVs in gnomAD (A) exomes and (B) genomes. Each functional category is subdivided by variant annotation in nORFs. Dotted lines correspond to results from bins of only canonical annotations previously reported (Karczewski et al. 2020). Higher values indicate an enrichment of lower frequency variants, suggesting negative selection.



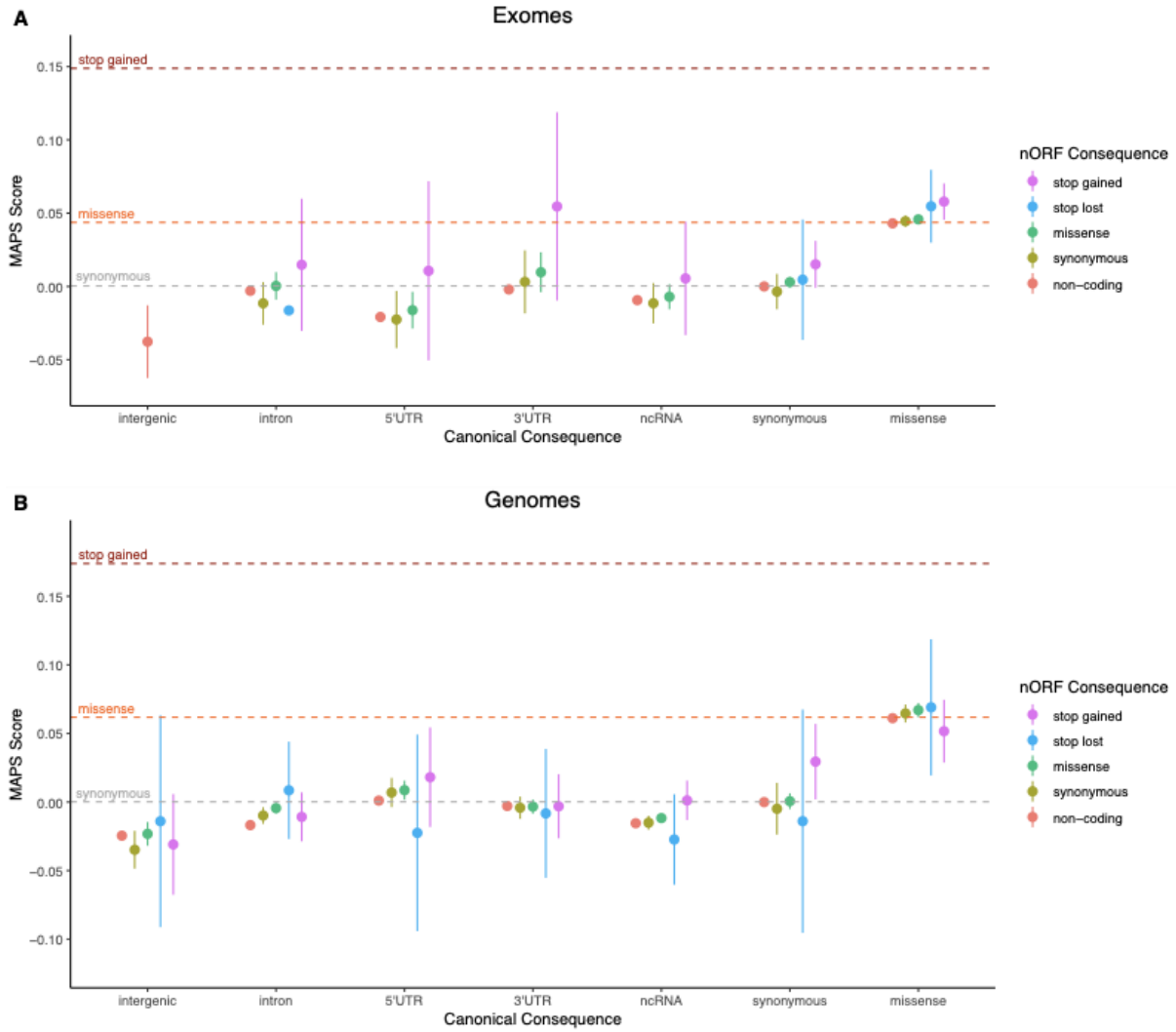
Supplemental Figure S2. Variant bin counts for gnomAD MAPS analysis. The 14.9 million and 229.9 million high-quality variants from gnomAD (A) exomes and (B) genomes respectively were binned based on their worst consequence in the context of nORFs and canonical annotations. Non-coding refers to variants annotated as non-coding by VEP (intergenic + upstream gene + downstream gene) in the context of nORFs. Bins are colored based on count on a logarithmic scale.



Supplemental Figure S3. Annotation Prioritization Priority. Genomic location annotation of nORFs was prioritized by first selecting overlaps with canonical CDS in an alternative frame (altCDS), altCDS-UTR combinations, and full UTR overlaps. This was followed by full overlaps with ncRNA transcripts, then by partial overlaps with protein coding transcripts, partial overlap with ncRNA transcripts, and finally intronic or intergenic regions.



Supplemental Figure S4 Averaged heritability partitioned across 11 UK Biobank traits for nORF regions. Heritability enrichment was compared for canonical gene annotation from GENCODE vs nORF annotation for common variation enrichment (CVE), defined as the proportion of common variant heritability explained by the annotation divided by the proportion of common variants in the annotation.



Supplemental Figure S5 MAPS calculated using only OpenProt entries. The mutability-adjusted proportion of singletons (MAPS) was calculated for 35 variant bins of SNVs from gnomAD (A) exomes and (B) genomes. The canonical annotation of the bin is indicated along the x-axis, while the nORF annotation is indicated by colour. Dotted lines correspond to results from bins of only canonical annotations previously reported (Karczewski et al. 2020). Higher values indicate an enrichment of lower frequency variants, suggesting negative selection. Bins with fewer than 100 variants are excluded.

Supplemental Methods

Selection of sources for evidence of nORFs

Three existing databases with entries that qualify as nORFs were considered for inclusion in the nORFs dataset: OpenProt (Brunet et al. 2019), sORFs.org (Olexiouk et al. 2018), and SmProt (Hao et al. 2018). SmProt was not used due to inconsistencies in data (e.g. incorrect genomic coordinate annotations) and lack of details in their methods to reanalyse the data, specifically in regards to their MS evidence (Olexiouk et al. 2018). By contrast, OpenProt and sORFs.org have shown commitment to providing consistent, verifiable, and maintained data, and were therefore used as the main sources for the nORFs dataset.

OpenProt (Release 1.3) predicts all possible ORFs with an ATG start codon and a minimum length of 30 codons that map to an Ensembl (Zerbino et al. 2018) or Refseq (O'Leary et al. 2016) transcript. They identified 607,456 alternate ORFs (altORFs) that are neither canonical ORFs, nor an isoform of those ORFs, but in non-coding regions or an alternate frame to canonical CDS. Although OpenProt maps to both Ensembl and Refseq transcripts, we focus exclusively on the Ensembl annotations for compatibility with the sORFs.org dataset and other downstream analyses. From the altORFs mapped to Ensembl transcripts, we consider the 26,480 altORFs with translation evidence from MS (21,708), ribosome profiling (5,059), or both (398).

The sORFs.org database (downloaded April 30, 2019) uses notably different inclusion criteria, annotating 'sORFs' with translation evidence from 43 human ribosome profiling experiments, then adding MS evidence found in publicly available datasets. The sORFs are defined as ORFs between 10 and 100 codons using any of four start codons: 'ATG', 'CTG', 'TTG', or 'GTG', and are not restricted to known transcripts.

These sORFs are identified through a translation initiation site (TIS) detection pipeline with a noise filtering step to limit false positive detection events. This process is fully described in

their database paper (Olexiouk et al. 2018). Briefly, they identify all start sites genome-wide using the four most common start codons: 'ATG', 'CTG', 'TTG' and 'GTG'. They then scan all start sites for an in-frame stop codon within 300nt (thereby limiting detected sORFs to 100 codons), both with and without considering transcript splicing data. They filter out unlikely translation events by implementing a threshold of at least 10% in-frame coverage and 10 ribosome profiling fragments (RPFs). As a noise filter to detect and remove false positives, they convert sORF transcripts into binary arrays of positions covered by ribosomes (1) and positions not covered (0). Then, they shuffle the array and recalculate in-frame coverage 10,000 times, allowing a probability calculation for the likelihood of a non-random translation event.

Curation of nORFs

The curation steps we performed to create a nORF dataset are detailed in **Fig. 1**. The final dataset that we created a) contains only nORFs with translation evidence from either MS or ribosome profiling b) contains no duplicate or highly similar entries and c) contains only ORFs clearly distinct from currently annotated canonical proteins.

We used 607,456 predicted altORFs from OpenProt and filtered to the 26,480 entries with MS or ribosome profiling evidence of translation. From over 2.1 million sORFs.org entries with 'good' or 'extreme' ORFscore (Bazzini et al. 2014), 502,056 entries with unique genomic mappings were extracted **(Fig 1A)**. The next step involved processing similar entries in the sORFs.org dataset that shared the same stop site and amino acid sequences up to differing start sites. A characteristic example is shown in **Fig 1B** where in an alternative frame of the final coding exon of the *MRPS21* gene, sORFs.org provides evidence for five small ORFs sharing the same end site and differing only by their start site. This is common in the sORFs.org dataset because of the ambiguity in ribosome profiling experiments to identify the

correct translation start site, unless specifically using methods that search for them (e.g. ribosome profiling with antibiotics used to trap newly initiated ribosomes at start codons) (Olexiouk et al. 2018; Weaver et al. 2019). Although ideally the correct start site(s) would be identified through experiments, this data is not currently available. For consistency and simplicity, we have selected the longest ORF in these cases, which may not always represent the true translated ORF, but will always encompass all ORFs identified at these sites. We emphasize this ambiguity in the correct start site as an important limitation to be kept in mind when using the dataset. In all, the selection of the longest ORF at ambiguous start sites further reduced extracted sORFs.org entries to 209,543.

Next, the OpenProt and sORFs.org datasets were merged, 1,028 redundant entries between the datasets were removed, and 1,976 cases of ambiguous start sites between the two datasets were resolved by again taking the longest ORF, resulting in a merged total of 233,021 entries. The small number of overlapping or similar entries between the two datasets can be partly attributed to different inclusion criteria for ORFs between the databases (i.e. ORF length, start codon, transcript requirement) and the main source of entries (sORFs from ribosome profiling and OpenProt predominantly from MS).

Finally, we separated all entries that were in-frame with canonical CDS, as the translation evidence from these entries cannot be unambiguously resolved as to whether they are from a canonical protein product or an independent nORF embedded within a canonical protein. We identified 38,614 such entries and removed them, leaving a total of 194,407 entries in the final nORFs dataset. An example case is shown in **Fig. 1C** where two small ORFs overlap the CDS of the *RICA* gene. One of these ORFs is in the same frame as the *RICA* CDS and was therefore filtered out, whereas the second ORF is in a different frame and retained in the dataset. Following this final curation step all entries in the nORF dataset that overlap canonical CDS are in a different frame from and do not share amino acid sequence with that CDS.

Annotation of nORFs

We annotated each nORF with reference to human GENCODE (v30) gene annotations (Frankish et al. 2019). The annotation categories included nORFs mapping to UTRs or CDS of protein coding transcripts, ncRNAs, or intergenic regions. When multiple annotations were possible, due to multiple transcripts in a region, annotations were prioritized by first selecting full overlaps with protein coding transcripts, particularly those that overlap canonical CDS in an alternative reading frame (altCDS), followed by full overlaps with ncRNA transcripts, then by partial transcript overlaps, and finally intronic or intergenic regions. Our detailed prioritization summary is shown in **Supplemental Fig. 3**.

Using GENCODE 34 (latest version) our pipeline identifies 194,291 rather than 194,407 nORFs, meaning that between releases 30 and 34, 116 nORFs became part of canonical CDS as newly identified genes or as part of new coding transcripts of existing genes. We find it encouraging that some nORFs are becoming canonical CDS and plan to regularly update our GENCODE reference in future iterations of the nORFs database.

Database and web platform

To reduce the threshold of accessibility, databases need to be accessible with minimal requirements of tools or prior knowledge. We therefore built an online platform with Representational State Transfer (REST) application programming interface (API) functionality. This online platform acts as an entry and lookup point for individual entries, while the REST API is feature compatible with existing bioinformatics pipelines. We made the curated and annotated GRCH38 raw dataset available in .bed and .gtf format as well as downloadable nORFs.org UCSC track.

The norfs.org web platform was built with JavaScript ES6, webpack 4.5.0 and Facebook's react.js framework in version 16.4.1. Furthermore, Gogo react 2.04 provides the CSS3

elements for a flexible dashboard layout. Google's Firebase cloud service was utilized to host a distributed NoSQL equivalent database with REST API access.

On top of these base protocols, npm packages such as react-router (4.4.2), feature-viewer (0.1.44) and biodalliance (0.13) (Down et al. 2011) were embedded to create a professional and highly customizable layout. Specifically, biodalliance allowed the creation of a genome browser with optional additional feature tracks, and the feature-viewer was used to annotate the peptide itself with structure and potential variant annotations. Considering reproducible research guidelines, we used git as a versioning tool and uploaded the repository to GitHub under an MIT license (https://github.com/PrabakaranGroup/nORFs.org).

Stratified LD score regression (S-LDSC) heritability analysis

Heritability is a statistical concept used to describe how much of the observed variation for a phenotype is due to genetic variation (Visscher et al. 2008). S-LDSC (Finucane et al. 2015; Gazal et al. 2017) enables the estimation of heritability enrichment for functional annotations in human traits and diseases. Although it was originally restricted to partitioning heritability explained by common variants (h^2_C), it was recently extended with the baseline-LF model (Gazal et al. 2018) to also allow the partitioning of heritability explained by low frequency

(0.5%≤ MAF <5%) variants (h^2_{LF}). We used S-LDSC with the baseline-LF model to estimate the heritability enrichment of nORF annotations for both common and low frequency heritability. To achieve this we had to first generate the two required inputs to S-LDSC: genome-wide association study (GWAS) summary statistics for traits of interest and an external LD reference panel with ancestry matching the GWAS population. As applied previously (Gazal et al. 2018), we obtained summary statistics for 40 heritable, complex UK Biobank (Bycroft et al. 2018) traits (downloaded from

https://data.broadinstitute.org/alkesgroup/UKBB/UKBB_409K/) that were restricted to 409 K individuals with UK ancestry. We then generated an LD reference panel for UK ancestry to match the summary statistics with 3,567 UK10K (The UK10K Consortium 2015, 10) wholegenome sequencing (WGS) samples from the ALSPAC and TWINSUK cohorts.

With these inputs, we analyzed a total of 177 genomic annotations, each corresponding to a defined set of variants, for their heritability enrichment. Of the 177, 163 are together known as the previously described baseline-LF model (Gazal et al. 2018). Briefly, the baseline-LF model is made up of MAF bins, LD-related annotations, and 33 main binary annotations for both lowfrequency and common variants. These main binary annotations include a number of gene related, regulatory, and conservation based annotations. We added to the analysis 14 custom annotations, from seven functional annotations doubled for common variants and low frequency variants. Of these seven, three custom annotations were nORF related: one for all nORFs, and 2 in which nORFs were split at the variant level to those regions which overlap canonical CDS (norfs_altCDS), and those which do not (nORFs_noCDS). The remaining 4 were canonical annotations from GENCODE: transcribed regions, CDS, 5'UTRs, and 3'UTRs. It should be noted that similar annotations appear to be already present in the baseline-LF model, but they were generated from a different reference set than our nORFs (UCSC 2013) (Gusev et al. 2014) and their 'Coding' annotation contains UTRs, which our custom annotation does not.

For the baseline-LD functional annotations and our custom annotations, we calculated common variant enrichment (CVE) and low frequency variant enrichment (LFVE) for each of the 40 UK Biobank traits. CVE is the proportion of common heritability (h^2_C) divided by the proportion of common single nucleotide polymorphisms (SNPs) in the annotation, while LFVE is proportion of low-frequency heritability (h^2_{LF}) divided by the proportion of low frequency SNPs in the annotation:

$$CVE = Prop(h^{2}_{C})/Prop(common\ SNPs)$$

$$LFVE = Prop(h^{2}_{LF})/Prop(low\ frequency\ SNPs)$$

Meta-analysis of results was conducted using random-effects meta-analyses in the rmeta package on 27 independent traits (Gazal et al. 2018), indicated in **Table S1**. All standard errors were computed using a block jackknife procedure (Bulik-Sullivan et al. 2015). Results for all traits separately are available at https://github.com/PrabakaranGroup/nORF-data-prep/.

Mutability adjusted proportion of singletons (MAPS) analysis

In addition to its importance for association studies and measures of heritability, genetic variation is also critical to evaluating natural selection at the variant level. Natural selection is an essential mechanism of evolution and acts over time to eliminate deleterious variants from populations. MAPS allows the estimation of negative selection, a proxy for functional importance, of variant classes in the genome. In this study, we apply MAPS to nORF variant classes (e.g. missense, stop lost, stop gained) to infer possible negative selection and potential signals of function.

We calculated the MAPS score for classes of variants based on their consequence in nORFs and canonical annotations to infer selection levels against these variants. MAPS was first described in the release of the Exome Aggregation Consortium (ExAC) dataset (Lek et al. 2016) and then updated with the release of gnomAD (Karczewski et al. 2019).

We calculated MAPS with gnomAD genomes and exomes by using publicly available code at https://github.com/macarthur-lab/gnomad_lof. We modified the code to include variant bins based on both nORF consequences and canonical consequences, rather than only canonical consequences. We selected five nORF consequences of interest: missense, synonymous,

stop lost, stop gained, and non-coding (intergenic + upstream gene + downstream gene) and 7 canonical consequences of interest: missense, synonymous, ncRNA, 5'UTR, 3'UTR, intronic and intergenic. For each of these 35 (5x7) bins, MAPS calibrated expected variant frequencies to account for 1 surrounding base of context and CpG methylation, two factors known to influence the mutability of base pairs (Lek et al. 2016). The transformation between variant frequencies and the expected proportion of singletons was regressed against the observed proportion of synonymous variants in canonical proteins. As the MAPS score given to variant classes is a relative metric, this means that synonymous variants in canonical proteins were set as 0 and higher scores reflected more negative selection. We reported MAPS scores for bins with at least 100 variants in the gnomAD exomes or genomes dataset respectively.

P-values were calculated using a bootstrapping approach as applied previously (Whiffin et al. 2020). For a given bin with *n* variants, *n* variants were randomly sampled with replacement and used to calculate MAPS for two bins or interest: bin A and bin B. This was repeated over 10,000 permutations with the P-value being the proportion of permutations where MAPS of bin B was less than MAPS of bin A. P-values were considered significant if they passed Bonferroni correction of 18 tests for exomes and 21 for genomes.

Variant annotation

A major application of variant class importance is to the biological interpretation of specific mutations, particularly in disease contexts. A number of databases exist to catalogue disease mutations such as HGMD (Stenson et al. 2017) and ClinVar (Landrum et al. 2018) for inherited mutations and COSMIC (Tate et al. 2019) for acquired cancer mutations. For these disease mutations, the primary method of mechanistic interpretation is annotating variants for their impact in canonical protein coding genes (Gloss and Dinger 2018). In particular, loss-offunction mutations, such as nonsense, frameshift or essential splice variants, are thought to

be common mechanisms of pathogenicity (MacArthur et al. 2012). However, the mechanism of many disease mutations cannot currently be explained by their impact on canonical proteins alone (Gloss and Dinger 2018). This is unsurprising given that non-coding regions make up over 98% of the genome and have diverse regulatory functions (ENCODE Project Consortium 2012). Specific examples of non-coding mechanisms include effects on gene expression from mutations at gene promoters (Fredriksson et al. 2014) or epigenetic imprinting loci (Chuang et al. 2017) and RNA stability from 5' and 3' UTR mutations (Zeraati et al. 2017).

Variant annotation was carried out using version 96 of VEP (McLaren et al. 2016) to investigate the consequences of variants in the context of canonical frames and nORFs. Variant sets were obtained for annotation as VCFs. These included gnomAD genomes and exomes (release 2.1.1) (Karczewski et al. 2019), HGMD (pro release 2019.2) (Stenson et al. 2017), ClinVar (release 2019 0708) (Landrum et al. 2018), and COSMIC coding and non-coding mutations (v89) (Tate et al. 2019). Each set of variants was annotated for their most severe consequence as defined by VEP with respect to a) canonical gene annotations, corresponding to GENCODE 30 in GRCh38 or GENCODE 30 lifted over to GRCh37 and b) nORF annotations provided as a custom GTF in the appropriate genome assembly.

When examining possible disease mutations that could be explained by nORF consequences, we first filtered variants from the disease mutations databases (COSMIC, HGMD, and ClinVar) to remove those with strongly deleterious annotations in canonical proteins (i.e. essential splice, frameshift, stop gained, stop lost, start lost). We then further filtered these variant sets to those with possible pathogenic consequences in nORFs (stop lost, stop gained, and frameshift).