TYPE Original Research
PUBLISHED 08 September 2023
DOI 10.3389/fmolb.2023.1257550

A curated census of pathogenic and likely pathogenic UTR variants and evaluation of deep learning models for variant effect prediction

Emma Bohn[1‡], Tammy T. Y. Lau[1‡], Omar Wagih[1], Tehmina Masud[1*] and Daniele Merico[1,2*†]

[1]Deep Genomics Inc., Toronto, ON, Canada, [2]The Centre for Applied Genomics, Hospital for Sick Children, Toronto, ON, Canada

OPEN ACCESS

EDITED BY
Silvia Bottini,
Université Côte d'Azur, France

REVIEWED BY
Dulika Sanjeewani Sumathipala,
University of Cambridge, United Kingdom
James Stephenson,
European Bioinformatics Institute (EMBL-EBI), United Kingdom

*CORRESPONDENCE
Tehmina Masud,
tehmina@deepgenomics.com
Daniele Merico,
daniele.merico@gmail.com

†PRESENT ADDRESS
Daniele Merico,
Vevo Therapeutics, San Francisco, CA, United States

‡These authors have contributed equally to this work and share first authorship

RECEIVED 12 July 2023
ACCEPTED 28 August 2023
PUBLISHED 08 September 2023

CITATION
Bohn E, Lau TTY, Wagih O, Masud T and Merico D (2023), A curated census of pathogenic and likely pathogenic UTR variants and evaluation of deep learning models for variant effect prediction.
Front. Mol. Biosci. 10:1257550.
doi: 10.3389/fmolb.2023.1257550

Introduction: Variants in 5′ and 3′ untranslated regions (UTR) contribute to rare disease. While predictive algorithms to assist in classifying pathogenicity can potentially be highly valuable, the utility of these tools is often unclear, as it depends on carefully selected training and validation conditions. To address this, we developed a high confidence set of pathogenic (P) and likely pathogenic (LP) variants and assessed deep learning (DL) models for predicting their molecular effects.

Methods: 3′ and 5′ UTR variants documented as P or LP (P/LP) were obtained from ClinVar and refined by reviewing the annotated variant effect and reassessing evidence of pathogenicity following published guidelines. Prediction scores from sequence-based DL models were compared between three groups: P/LP variants acting though the mechanism for which the model was designed (model-matched), those operating through other mechanisms (model-mismatched), and putative benign variants. PhyloP was used to compare conservation scores between P/LP and putative benign variants.

Results: 295 3′ and 188 5′ UTR variants were obtained from ClinVar, of which 26 3′ and 68 5′ UTR variants were classified as P/LP. Predictions by DL models achieved statistically significant differences when comparing modelmatched P/LP variants to both putative benign variants and modelmismatched P/LP variants, as well as when comparing all P/LP variants to putative benign variants. PhyloP conservation scores were significantly higher among P/LP compared to putative benign variants for both the 3′ and 5′ UTR.

Discussion: In conclusion, we present a high-confidence set of P/LP 3′ and 5′ UTR variants spanning a range of mechanisms and supported by detailed pathogenicity and molecular mechanism evidence curation. Predictions from DL models further substantiate these classifications. These datasets will support further development and validation of DL algorithms designed to predict the functional impact of variants that may be implicated in rare disease.

KEYWORDS

deep learning, non-coding variation, rare disease, untranslated region (UTR), variant classification

1 Introduction

As the diagnostic utility of whole genome sequencing (WGS) in rare disease populations is increasingly documented (Stavropoulos et al., 2016; Clark et al., 2018), there is a growing appreciation for the direct implication of non-coding variation in heritable disease (French and Edwards, 2020). Clinically relevant variants have been identified in a range of functional elements residing in non-coding regions, including 5′ and 3′ untranslated regions (UTRs) (Chatterjee and Pal, 2009; Steri et al., 2018). These regions flank the coding sequence, and play an important role in mRNA stability, translation, and other mechanisms of post-transcriptional regulation. In addition, the 5′UTR overlaps the promoter region, and the corresponding DNA sequence can play a role in transcriptional regulation. Variation within the 3′ and 5′ UTR can result in functional consequences mediated through a variety of molecular mechanisms including, but not limited to, modification of RNA secondary structure, modulation of reading frame recognition by the translation machinery, and altered interaction with microRNAs (miRNAs) and RNA-binding proteins (RBPs).

As a direct result of the existence of clinically meaningful variation in UTRs, there is a need for accurate classification of pathogenicity for variants in these regions. Recent work has been conducted to develop recommendations for adapting existing variant classification guidelines to variants in non-coding contexts (Ellingford et al., 2022). Predictive algorithms have also been developed to assist in these efforts (Wells et al., 2019; Moyon et al., 2022; Petrazzini et al., 2022). However, the accuracy and consequent utility of such tools depends upon the quality of the data used for training and validation. Databases such as ClinVar (Landrum et al., 2018) and The Human Gene Mutation Database (HGMD) (Stenson et al., 2003) are frequently used as a source of training datasets. While valuable resources, conflicting classifications between submitters are common (Rehm et al., 2015) and frequent misclassifications have been documented (Shah et al., 2018; Xiang et al., 2020).

Given existing evidence of misclassification and known challenges of non-coding variant classification, we used ClinVar as a starting point for systematic curation and classification to develop a high-confidence set of pathogenic and likely pathogenic (P/LP) variants in the 3′ and 5′ UTR. Focus was placed on gathering mechanistic information, resulting in a collection of P/LP variants representing a range of proposed functional mechanisms. The census of P/LP variants presented in this work is further substantiated by findings from the application of deep learning (DL) models. These models demonstrated a distinction in prediction scores for variants acting through specific mechanisms relevant to the model, as compared to both P/LP variants acting through other mechanisms and putative benign variants. The resulting high-confidence set of curated variants will serve to support the continued development and accurate validation of DL algorithms designed to predict the functional impact of variants in the 3′ and 5′ UTR which may be implicated in rare disease.

2 Materials and methods

2.1 Variant identification and filtering

The tab delimited summary text file from the September 2022 release of ClinVar for GRCh38 was downloaded from the National Center for Biotechnology Information (NCBI) FTP site (last accessed 20 Oct 2022) (Landrum et al., 2018). Variants were further filtered to include only those with classifications of pathogenic, likely pathogenic, or with conflicting classifications of pathogenic and likely pathogenic across multiple submitters in ClinVar. Variants were annotated with allele frequencies from gnomAD v3.0 (Chen et al., 2022). Any variant with an allele frequency greater than 0.05 (5%) for any subpopulation was discarded, given this is sufficient evidence to warrant a benign classification (Richards et al., 2015). Variants were then filtered to include only those annotated in the transcript defined by ClinVar's preferred name as 3′ or 5′ UTR exonic variants, defined as those located entirely within the coordinates of the respective UTR of the transcript, which must be coding. The transcript features were defined by NCBI RefSeq v109 annotations for *Homo sapiens* (O'Leary et al., 2016).

2.2 Curation of evidence related to variant effect and pathogenicity

The first step of the curation process consisted of confirming the variant's 3/5′ UTR effect by considering the gene's principal transcript as defined by APPRIS (Rodriguez et al., 2022). We excluded variants outside of the 3/5′ UTR of the most recent version of the APPRIS-defined principal transcript. In instances where a variant had a non-UTR effect on an overlapping protein-coding gene or non-coding RNA, the literature was consulted to determine whether pathogenicity was conferred through the UTR impact. If evidence suggested that pathogenicity was mediated by the non-UTR effect, or was insufficient to reach such a conclusion, the variant was excluded. Variants were also excluded in instances where they had a non-UTR impact on a Matched Annotation from the NCBI and EMBL-EBI (MANE) Select or Plus Clinical transcript for the same gene (Morales et al., 2022), regardless of having a UTR impact on the APPRIS-defined principal transcript. We also excluded variants with a non-UTR impact on an APPRIS-defined alternative or minor transcript when evidence implicating the variant as pathogenic was in the context of this transcript or there was insufficient evidence to discern the specific transcript mediating pathogenicity. Somatic variants, repeat expansions and structural variants with impacts extending beyond the UTR (i.e., large copy number variants (CNVs) impacting multiple gene regions or multiple genes) were excluded.

Synthesis of relevant evidence pertaining to each variant involved first consulting comments and reviewing citations provided directly in ClinVar. For each variant, an ancillary search was conducted to identify additional relevant literature. Data relevant to informing classifications of pathogenicity were extracted and documented. The information extracted depended on the nature of the study. For example, extracted details relevant to functional studies included the type of assay, results and overall conclusion. For case reports, the number of cases, their respective genotypes (e.g., homozygous, compound heterozygous) and phenotypic characteristics were recorded. Information related to proposed or validated mechanisms by which variants mediate their effects in the 3′ or 5′ UTR was captured, when available.

2.3 Classification of pathogenicity

Curated evidence informed classifications of pathogenicity based on guidelines published by the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) (Richards et al., 2015). In keeping with recent recommendations for classifying non-coding variants (Ellingford et al., 2022), the following additional considerations were made: PS1 (strong criterion 1) was applied to variants impacting the same base as another variant previously established as pathogenic (e.g., a different change at the same base within a miRNA binding site). PM5 (moderate criterion 5) was applied for variants with the same predicted effect as a previously established pathogenic variant, but not at the same specific base/residue (e.g., a variant at a different site within a miRNA binding site in which another variant at a different base had been previously established as pathogenic).

2.4 Exploration of properties of genes implicated by curated variants

To investigate whether genes in which variants classified as P/LP in this study are implicated in any shared families or common pathways, we performed functional enrichment analysis using the g:GOSt tool from g:Profiler (Raudvere et al., 2019; https://biit.cs.ut.ee/gprofiler/gost; version e109_eg56_p17_1d3191d). Genes with at least one 3′ or 5′ UTR variant classified as P/LP in this study were used as the query gene set. The background was defined as any gene with at least one variant, that was not a somatic or CNV, classified as P/LP in ClinVar. The Benjamini–Hochberg false discovery rate (FDR) method was used for multiple test correction, with the significance threshold set to 0.05. The following data sources were used in the analysis: Gene Ontology (GO) molecular function, GO cellular component, GO biological process, and Reactome.

To examine the extent to which genes in which curated variants exist are known to be implicated in disease, we annotated the set of genes with at least one curated 3′ or 5′ UTR variant with genetic phenotypes documented in Online Mendelian Inheritance in Man® (OMIM; https://www.omim.org/, last accessed Apr. 18, 2023). To achieve a comprehensive scope, annotations included all values documented in the Disorders column including monogenic disorders, “non-diseases” (i.e., indicated by brackets), susceptibility to multifactorial disorders (i.e., indicated by braces), and provisional associations (i.e., indicated by a question mark). Each gene was also annotated with the number of variants classified as P/LP in ClinVar, as well as the loss-of-function (LoF) observed/expected upper bound fraction (LOEUF) from gnomAD v2.0 (Karczewski et al., 2020; last accessed 08 Aug 2023). This metric gives an impression of the extent to which genes are intolerant to LoF variants, with lower values indicating stronger intolerance. For each of these properties, a two-sided Mann-Whitney Wilcoxon test was applied to compare values between three groups: genes in which at least one 3′ or 5′ UTR variant was classified as P/LP in this study, those in which at least one VUS and no P/LP variants were identified in this study, and the background set of genes defined for the functional enrichment analysis.

2.5 Analysis of variant effects using deep learning models and conservation scores

To create the benchmark datasets for DL models, variants classified as P/LP were supplemented with putative benign variants obtained from gnomAD v3.0 (Chen et al., 2022). The datasets were created separately for the 5′ and 3′ UTR. For each transcript with a variant classified as P/LP in the corresponding UTR, the genomic coordinates of all exons within the respective UTR were used to construct an SQL query to extract variants within these regions from ‘bigquery-public-data.gnomAD.v3_genomes__chr*` tables on BigQuery. Variants were further filtered to have a total allele frequency greater than 0.01 (1%), a threshold more inclusive than that used for initial filtering of P/LP variants (0.05; see Section 2.1), to allow for a greater likelihood of including a matched putative benign variant in the same UTR for the majority of P/LP variants, resulting in a more robust benchmark. Three models were applied to the resulting datasets: FramePoolCombined (Karollus et al., 2021), Saluki (Agarwal and Kelley, 2022), and Enformer (Avsec et al., 2021). Datasets were also annotated with PhyloP conservation scores (Pollard et al., 2010).

2.5.1 FramePoolCombined predictions

The hg38 reference FASTA was downloaded from UCSC (http://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz; last accessed 11 Jan 2023). A GTF file of the hg38 NCBI RefSeq table was downloaded from UCSC (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/genes/hg38.ncbiRefSeq.gtf.gz; last accessed 15 Feb 2023). A BED file of all 5′ UTR features was created from this GTF. Predictions for 5′ UTR variants were made using the Kipoi interface provided for the FramePoolCombined model (Avsec et al., 2019; Karollus et al., 2021). Each variant was predicted one at a time in its own VCF to yield individual variant effects, as the default behavior of the model through Kipoi is to integrate all variants in the VCF into the sequence for prediction. The predicted mean ribosome load (MRL) fold change was used as the variant effect prediction. Variants were stratified into three groups for analysis: “P/LP (open reading frame (ORF) mechanism)” if they were classified as P/LP and operated through a proposed mechanism of impacting ORF recognition by the translation machinery which included impact on an existing regulatory upstream ORF (uORF) or introduction of a novel upstream start codon (“model-matched”), “P/LP (Other)” for P/LP variants operating through a different or undetermined mechanism (“model-mismatched”), and “putative benign”. A Mann-Whitney Wilcoxon test was used to compare variant effect prediction scores between groups.

2.5.2 Saluki predictions

For each 3′ UTR variant, the 6D track for the Saluki model, consisting of the one-hot encoded DNA sequence of the transcript, the coding frame, and the splice site positions was constructed for the reference and alternative sequence (Agarwal and Kelley, 2022). Predictions using all 50 cross-fold validation models provided by the authors were made for each of the reference and alternative sequences for the variant, and the average of the predictions from all models was used as the resulting score. The variant effect prediction was taken as the difference between the alternative and reference sequence scores. Variants were stratified into three groups for analysis: "P/LP (mRNA stability)" if they were classified as P/LP and proposed to impact the polyadenylation signal or mRNA stability ("model-matched"); "P/LP (Other)" for other P/LP variants operating through a different or undetermined mechanism ("model-mismatched"), and "putative benign". The scores of the variants in each group were compared using the Mann-Whitney Wilcoxon test.

2.5.3 Enformer predictions

Enformer was loaded via tfhub.dev (https://tfhub.dev/deepmind/enformer/1; last accessed 16 Feb 2023). For each 5′ UTR variant, two sequences encompassing Enformer's full context length were constructed: one centered at the reference allele and one centered at the alternative allele. For each sequence, predictions for the forward and reverse strand were made by Enformer, and the average was taken. The output was subset to the center window and two windows on either side (for a total of five windows) and only the CAGE tracks. The score for each sequence was calculated by summing over the five windows, adding a pseudocount of one, applying a log2 transformation, and finally computing the mean. The variant effect prediction was taken as the difference between the alternative and reference sequence scores. Variants were stratified into three groups: "P/LP (Transcription)" if they were classified as P/LP and operated at the level of transcription ("model-matched"); "P/LP (Other)" for all P/LP variants operating through a different or undetermined mechanism ("model-mismatched"), and "putative benign". The scores of the variants in each group were compared using the Mann-Whitney Wilcoxon test.

2.5.4 PhyloP conservation score annotations

The BigWig file of PhyloP conservation scores (hg38 100way, vertebrate alignments) was downloaded from UCSC (https://hgdownload.cse.ucsc.edu/goldenpath/hg38/phyloP100way/; last accessed 22 Feb 2023; Pollard et al., 2010). All 3′ and 5′ UTR variants were annotated based on the score at the corresponding position in the genome. If a variant's reference sequence spanned multiple nucleotides, the maximum score across the interval was taken.

3 Results

3.1 Filtering and selection of 3′ and 5′ UTR variants from ClinVar

Variants documented in ClinVar were filtered based on classification (P/LP), gnomAD allele frequency (≤5%), and location within an exon of the 3′ or 5′ UTR (see Section 2.1 for details). This filtering process yielded 295 and 188 ClinVar variants from the 3′ and 5′ UTR. Of 295 ClinVar 3′ UTR variants, there were 85 single nucleotide variants (SNV), 28 indels (≤50 nucleotides), seven structural variants (SV; >50 nucleotides), two deletion-insertions, one CNV, and 172 tandem repeat expansions. Among the 188 ClinVar 5′ UTR variants, there were 147 SNVs, 35 indels, three deletion-insertions, one CNV, and two tandem repeat expansions. Variants were assigned a confidence score as a means of initial assessment based on whether they were annotated to reside in regions outside of the 3′ or 5′ UTR on any other transcript (either an alternative transcript of the same gene or a transcript of a different gene). Higher scores reflect higher confidence that the variant's effect is mediated through its impact on the 3′ or 5′ UTR of the ClinVar preferred transcript (Table 1). A score of two was assigned to 263 (89.2%) and 117 (62.2%) variants in the 3′ and 5′ UTR, respectively (Figure 1). All variants proceeded to curation irrespective of initial confidence score.

Of the initial 295 3′ UTR variants 59 (20.0%) proceeded to the classification phase. This included 39 SNVs, 17 indels, and three SVs contained entirely within the 3′UTR (deletion sizes ranging between 92 and 473 bp). The remainder (236 variants; 80.0%) were further filtered out for a variety of reasons, the most frequent of which being omission of repeat expansions (see Section 2.2 for details; Figure 1A). All repeat expansions existed in *DMPK*, a gene associated with myotonic dystrophy 1 in which "normal alleles" consist of 5–34 CTG repeats within the 3′ UTR, typically asymptomatic "premutation alleles" between 35 and 49, and fully penetrant alleles over 50 CTG repeats (Prior, 2009). Among the 172 *DMPK* repeat expansions in our dataset, 154 comprised >50 CTG repeats, while the remaining 18 involved between 31 and 50 repeats.

Of the initial 188 5′ UTR variants, 105 (55.9%) proceeded to classification. This included 93 SNVs, 10 indels, and two deletion-insertions. Among variants excluded, the most frequent reasons were transcript-related considerations (Figure 1B). One such consideration leading to variant exclusion was where, despite impacting a UTR of the APPRIS-defined principal transcript, variants had a non-UTR impact on an alternative transcript defined as either "Select" or "Plus Clinical" (i.e., transcripts not defined as "Select", but in which known pathogenic variants have been reported) by MANE (Morales et al., 2022). Discordances of this nature were identified for seven 3′ UTR and 15 5′ UTR variants (Figures 1A, B). Three variants in *KRAS* serve as examples of the former, residing in the 3′ UTR of the APPRIS-defined principal transcript (NM_033360), but having missense impacts on the MANE Select transcript (NM_004985). *In vitro* functional evidence for one such variant (NM_004985:c.468C>G; NM_033360.4:c.*22C>G) as it resides on NM_004985 and induces a missense impact supports profound activation of the MAPK pathway (Gremer et al., 2011). In the absence of any such functional support for the variant as it resides in the 3′ UTR, its pathogenicity is more reasonably attributed to the missense impact. A collection of variants in *MOCS2* serve as examples of the latter, where despite residing in the 5′ UTR of the MANE Select and APPRIS principal transcript (NM_004531), non-UTR impacts exist on the transcript defined as MANE Plus Clinical

TABLE 1 Initial confidence score assignment of variants based on ClinVar annotations.

Criteria	Confidence score		
	0	1	2
3′ (or 5′) UTR exonic variant based on ClinVar preferred transcript	✓	✓	✓
Not in a coding sequence exon in another transcript or a splicing variant[a] in another transcript	-	✓	✓
Not an intronic variant in another transcript or a 5′ UTR (or 3′ UTR) variant in another transcript	-	-	✓

UTR, untranslated region.

[a]Defined as an intronic variant within 8 bp of the intron/exon boundary.

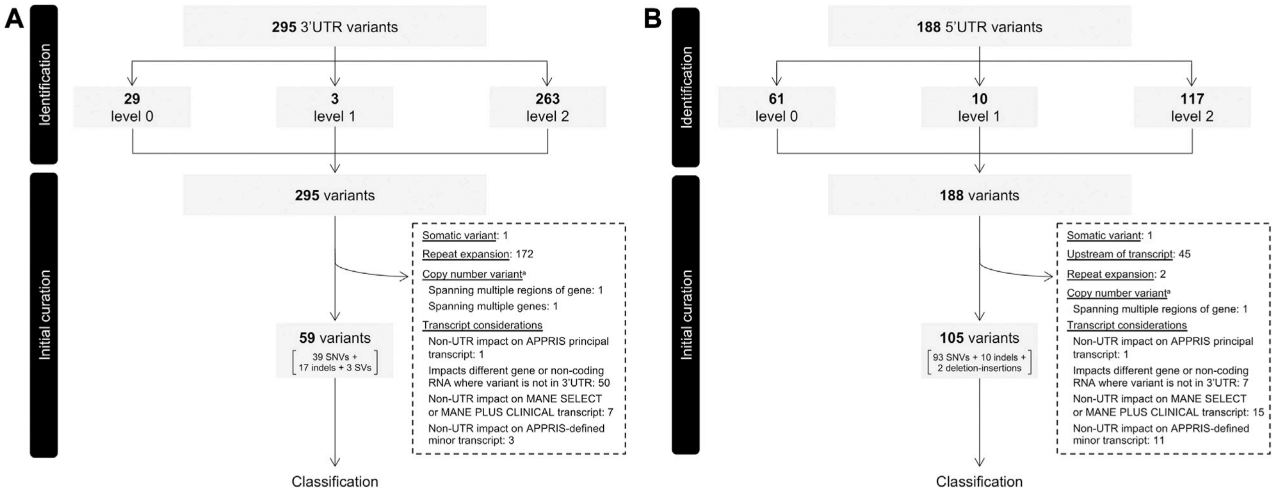
A
Identification
Initial curation
295 3'UTR variants
29
level 0
3
level 1
263
level 2
295 variants
Somatic variant: 1
Repeat expansion: 172
Copy number variant[a]
Spanning multiple regions of gene: 1
Spanning multiple genes: 1
Transcript considerations
Non-UTR impact on APPRIS principal transcript: 1
Impacts different gene or non-coding RNA where variant is not in 3'UTR: 50
Non-UTR impact on MANE SELECT or MANE PLUS CLINICAL transcript: 7
Non-UTR impact on APPRIS-defined minor transcript: 3
59 variants
39 SNVs + 17 indels + 3 SVs
Classification
B
Identification
Initial curation
188 5'UTR variants
61
level 0
10
level 1
117
level 2
188 variants
Somatic variant: 1
Upstream of transcript: 45
Repeat expansion: 2
Copy number variant[a]
Spanning multiple regions of gene: 1
Transcript considerations
Non-UTR impact on APPRIS principal transcript: 1
Impacts different gene or non-coding RNA where variant is not in 3'UTR: 7
Non-UTR impact on MANE SELECT or MANE PLUS CLINICAL transcript: 15
Non-UTR impact on APPRIS-defined minor transcript: 11
105 variants
93 SNVs + 10 indels + 2 deletion-insertions
Classification

FIGURE 1
Flow of variants through identification and initial curation phases determining eligibility for classification. **(A)** 3′ UTR variants. **(B)** 5′ UTR variants. Variants were extracted from ClinVar, assigned an initial confidence level and filtered based on several factors before proceeding to classification. UTR, untranslated region; VUS, variant of uncertain significance. [a]Copy number variants include large insertions and deletions. Identified CNVs all exceeded 1,500 bp in length.

(NM_176806). In these types of cases, and where there is no evidence providing mechanistic validation of any functional consequence of the variant in the context of a UTR of the APPRIS principal transcript, it remains challenging to ascertain whether any pathogenic impact is mediated by an effect on the UTR. These variants were therefore omitted from our dataset to limit our collection of variants to strictly those for which we are confident pathogenicity is mediated through an impact specifically on the UTR.

Of the 59 3′ UTR and 105 5′ UTR variants proceeding to classification, the majority were classified as pathogenic in ClinVar (3′ UTR: 39 (66.1%); 5′ UTR: 72 (68.6%)), with the remainder classified as either likely pathogenic or a combination of pathogenic and likely pathogenic across multiple submitters (Figure 2A). Thirty-five and 54 unique genes were represented among the 3′ and 5′ UTR variants that proceeded to classification, respectively. The distribution of variants among genes is illustrated in Supplementary Figure S1. Variants proceeding to the classification phase were classified according to guidelines published by the ACMG and AMP (Richards et al., 2015), with additional considerations made for variants in non-coding regions (Ellingford et al., 2022) (see Section 2.3 for details). All variants, classifications of pathogenicity, and supporting evidence are documented in Supplementary Tables S1, S2 for the 3′ UTR and 5′ UTR, respectively.

3.2 3′ UTR variant curation links pathogenic variants to a host of regulatory mechanisms

Of the 59 3′ UTR variants proceeding to classification, six were classified as pathogenic (10.2%), 20 likely pathogenic (33.9%) and 33 VUS (55.9%) (Figure 2B). 3′ UTR variants classified as pathogenic or likely pathogenic are provided in Table 2. There was a statistically significant difference in the number of ClinVar submissions for 3′ UTR variants classified in this study as P/LP compared to VUS, where the variants in the final P/LP group tended to have a greater number of submissions (two-sided Mann-Whitney Wilcoxon test, $p = 0.0057$; Supplementary Figure S2). There was no clear influence of the year of the most recent ClinVar submission on whether variants were assigned a final classification of P/LP or VUS (Supplementary Figure S3). Twenty-two of the 3′UTR variants classified as P/LP were supported by published functional evidence. All six variants classified as pathogenic in this work were also consistently classified as such by ClinVar submitters. Among the 20 variants classified as likely pathogenic in this

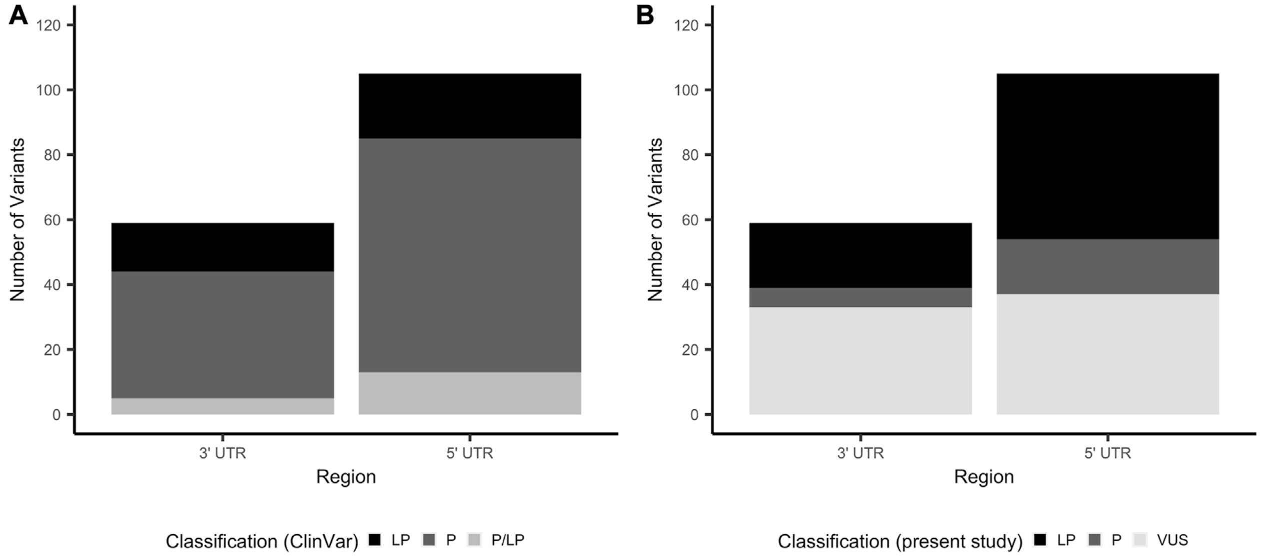

FIGURE 2
Number of 3′ and 5′ UTR variants by classification of clinical significance. **(A)** Total number of variants that proceeded to curation, stratified by clinical significance as reported by ClinVar submitters. **(B)** Total number of variants that proceeded to curation, stratified by clinical significance as classified in this study. LP, likely pathogenic; P, pathogenic; UTR, untranslated region; VUS, variant of uncertain significance.

work, 12 were pathogenic in ClinVar, four likely pathogenic, and four had conflicting classifications of pathogenic and likely pathogenic across multiple submitters.

A validated or proposed mechanism was reported for 22 variants, including disruption or introduction of a miRNA binding site (n = 7), impact on polyadenylation signal (n = 10), impact on mRNA stability (n = 3), introduction of a *de novo* splice site (n = 1), and change in secondary hairpin structure of the encoded mRNA (n = 1).

3.3 5′ UTR variant curation identifies variants spanning multiple pathogenic mechanisms

Of the 105 5′ UTR variants proceeding to classification, 17 (16.2%) were classified as pathogenic, 51 (48.6%) likely pathogenic, and 37 VUS (35.2%; Figure 2B). 5′ UTR variants classified in this study as P/LP tended to have a greater number of ClinVar submissions compared to those classified as VUS (two-sided Mann-Whitney Wilcoxon test, p = 0.0069; Supplementary Figure S2). Final classification of 5′ UTR variants as P/LP *versus* VUS was not clearly influenced by the date of the most recent ClinVar submission (Supplementary Figure S3). Fifty-five (80.9%) of the 68 variants classified as P/LP were supported by functional evidence. Of the 17 variants classified as pathogenic in this work, 12 were classified as pathogenic in ClinVar and three had a combination of pathogenic and likely pathogenic classifications across multiple submitters. Two pathogenic variants were classified as likely pathogenic in ClinVar, despite evidence cited in the respective submissions fulfilling sufficient criteria to warrant a pathogenic classification (Richards et al., 2015). Across the 51 variants classified as likely pathogenic in this work, 38 were documented as pathogenic in ClinVar, six likely pathogenic, and seven pathogenic and likely pathogenic across multiple submitters (Table 3).

A proposed or validated mechanism was also available for 55 variants classified as P/LP (Table 3). The described mechanisms operated both at the level of transcription (i.e., in cases where the 5′ UTR overlaps the functional promoter) and translation. Translational mechanisms included impacts on an existing regulatory upstream open reading frame (uORF; n = 10), introduction of a novel upstream start codon (n = 13), and altered interaction between mRNA and RNA-binding proteins (n = 12). Mechanisms operating at the level of transcription included altered promoter activity (e.g., enhancing or repressing transcription factor interactions; n = 14), altered promoter methylation (n = 2) and splicing impacts (n = 4).

3.4 Genes harboring variants classified as P/LP in this study are involved in established gene-phenotype pairings and enriched for pathways relevant to hematological disorders

Genes with at least one curated variant in our dataset had a higher average number of variants classified as P/LP in ClinVar (Supplementary Figure S4A). For both the 3′ and 5′ UTR, there was a significantly higher number of variants classified as P/LP in ClinVar among genes in which we identified at least one P/LP 3′ or 5′ UTR variant compared to a background set of genes in ClinVar with at least one P/LP variant overall (3′ UTR: p-value = 0.015; 5′ UTR: p-value ≤0.001). There was also a statistically significant difference in ClinVar P/LP counts for genes in which we identified at least one VUS and no P/LP variants compared to the background gene set (3′ UTR; p-value ≤0.001; 5′ UTR: p-value ≤0.001). Intolerance to LoF variation, as estimated by

TABLE 2 Pathogenic and likely pathogenic 3′ UTR variants by proposed or validated mechanism (n = 26).

Variant[a]	Gene	Classification(s) (ClinVar)	Classification (present study)
Altered miRNA binding (n = 7)			
NM_000518.4:c.*32A>C	*HBB*	P	LP
NM_001845.5:c.*35C>A	*COL4A1*	P	LP
NM_001845.5:c.*32G>T	*COL4A1*	P	P
NM_001845.5:c.*32G>A	*COL4A1*	P	P
NM_001845.5:c.*31G>T	*COL4A1*	P	LP
NM_001281503.1:c.*689G>A	*SLITRK1*	P	P
NM_006044.3:c.*282A>T	*HDAC6*	P	LP
Impact on polyadenylation signal (n = 10)			
NM_014017.3:c.*23C>A	*LAMTOR2*	P	LP
NM_000518.4:c.*110_*114del	*HBB*	P	LP
NM_000518.4:c.*113A>G	*HBB*	P/LP	LP
NM_000518.4:c.*110T>C	*HBB*	P	P
NM_000518.4:c.*110T>A	*HBB*	P	LP
NM_000518.4:c.*93_*105del	*HBB*	P	LP
NM_000517.4:c.*92A>G	*HBA2*	P	LP
NM_000517.4:c.*94A>G	*HBA2*	P	P
NM_003491.3:c.*43A>G	*NAA10*	LP	LP
NM_003491.3:c.*39A>G	*NAA10*	P/LP	LP
Impact on mRNA stability (n = 3)			
NM_000207.2:c.*59A>G	*INS*	P	LP
NM_003073.4:c.*82C>T	*SMARCB1*	P/LP	LP
NM_001017980.3:c.*13_*104del	*VMA21*	P	LP
Altered splicing (n = 1)			
NM_000132.3:c.*56G>T	*F8*	LP	LP
Impact on secondary structure (n = 1)			
NM_206926.1:c.*1107T>C	*SELENON*	LP	LP
Undetermined mechanism (n = 4)			
NM_000210.3:c.*94_*96del	*ITGA6*	LP	LP
NM_000518.4:c.*6C>G	*HBB*	P	P
NM_001017980.3:c.*6A>G	*VMA21*	P	LP
NM_000444.5:c.*231A>G	*PHEX*	P/LP	LP

[a]Variant HGVS, nomenclature is provided for the APPRIS-defined principal transcript.
LP, likely pathogenic; P, pathogenic; P/LP, P and LP, classifications across multiple submitters; uORF, upstream open reading frame; UTR, untranslated region.

LOEUF scores (Karczewski et al., 2020), was not statistically significantly different between any of the three groups examined (i.e., genes with at least one P/LP variant in our dataset, genes with VUS only in our dataset, background gene set; Supplementary Figure S4B).

To examine the extent to which genes in which curated variants exist are known to be implicated in disease, we annotated all genes with at least one curated 3′ or 5′ UTR variant with genetic phenotypes documented in OMIM (Supplementary Table S5). All of the 49 unique genes involving variants classified in this study as P/LP had associated genetic phenotypes documented in OMIM. Of these, only four were not associated with well-established monogenic disorders and instead involved either susceptibility to

TABLE 3 Pathogenic and likely pathogenic 5′ UTR variants by proposed or validated mechanism (n = 68).

Variant[a]	Gene	Classifications(s) (ClinVar)	Classification (present study)
Translation: Impact on existing regulatory uORF (n = 10)			
NM_000460.3:c.-31G>T	*THPO*	P	LP
NM_000460.3:c.-47del	*THPO*	P	LP
NM_005144.4:c.-218A>G	*HR*	P	LP
NM_005144.4:c.-249C>G	*HR*	P	LP
NM_005144.4:c.-315C>T	*HR*	P	LP
NM_005144.4:c.-320T>C	*HR*	P	P
NM_005144.4:c.-320T>A	*HR*	LP	LP
NM_000280.4:c.-118_-117del	*PAX6*	P/LP	P
NM_000280.4:c.-122dup	*PAX6*	LP	P
NM_004064.4:c.-454_-451del	*CDKN1B*	P	LP
Translation: introduction of novel upstream start codon (n = 13)			
NM_006516.2:c.-107G>A	*SLC2A1*	LP	P
NM_001204.7:c.-947_-946delinsAT	*BMPR2*	P	LP
NM_000939.3:c.-11C>A	*POMC*	P	LP
NM_000313.3:c.-39C>T	*PROS1*	P	LP
NM_054027.5:c.-11C>T	*ANKH*	P	LP
NM_001131005.2:c.-8C>T	*MEF2C*	P	LP
NM_001131005.2:c.-26C>T	*MEF2C*	P	LP
NM_001131005.2:c.-66A>T	*MEF2C*	P	LP
NM_001114753.2:c.-127C>T	*ENG*	P/LP	LP
NM_012203.2:c.-4_-3delinsAT	*GRHPR*	P	LP
NM_001025295.2:c.-14C>T	*IFITM5*	P	P
NM_000518.4:c.-29G>A	*HBB*	P	LP
NM_006767.3:c.-38T>A	*LZTR1*	P/LP	LP
Translation: altered mRNA-protein interaction (n = 12)			
NM_002032.2:c.-164A>T	*FTH1*	P	LP
NM_000146.3:c.-168G>A	*FTL*	P	P
NM_000146.3:c.-168G>C	*FTL*	P	P
NM_000146.3:c.-168G>T	*FTL*	P	P
NM_000146.3:c.-167C>T	*FTL*	P	P
NM_000146.3:c.-164C>A	*FTL*	P	LP
NM_000146.3:c.-164C>T	*FTL*	P	P
NM_000146.3:c.-161C>G	*FTL*	P	LP
NM_000146.3:c.-161C>T	*FTL*	P	P
NM_000146.3:c.-160A>G	*FTL*	P	LP
NM_000146.3:c.-157G>A	*FTL*	P	P
NM_000146.3:c.-149G>C	*FTL*	P	LP

(Continued on following page)

TABLE 3 (*Continued*) Pathogenic and likely pathogenic 5′ UTR variants by proposed or validated mechanism (n = 68).

Variant[a]	Gene	Classifications(s) (ClinVar)	Classification (present study)
Transcription: altered promoter activity (n = 14)			
NM_003051.3:c.-202G>A	*SLC16A1*	P	LP
NM_005105.4:c.-21G>A	*RBM8A*	P/LP	P
NM_000551.3:c.-75_-55del	*VHL*	LP	LP
NM_000037.3:c.-73_-72del	*ANK1*	P	LP
NM_000375.2:c.-203T>C	*UROS*	P	LP
NM_014915.2:c.-127A>T	*ANKRD26*	P/LP	P
NM_014915.2:c.-127A>G	*ANKRD26*	P	LP
NM_014915.2:c.-127A>C	*ANKRD26*	LP	LP
NM_014915.2:c.-128G>A	*ANKRD26*	P	P
NM_000518.4:c.-18C>G	*HBB*	P	LP
NM_001814.5:c.-55C>A	*CTSC*	P	LP
NM_017671.4:c.-20A>G	*FERMT1*	P	LP
NM_000026.3:c.-49T>C	*ADSL*	LP	LP
NM_000133.3:c.-17A>G	*F9*	P	LP
Transcription: altered promoter methylation (n = 2)			
NM_007294.3:c.-107A>T	*BRCA1*	P	LP
NM_000249.3:c.-27C>A	*MLH1*	P/LP	LP
Transcription: splicing (n = 4)			
NM_021067.4:c.-60A>G	*GINS1*	P	LP
NM_021067.4:c.-48C>G	*GINS1*	P	P
NM_000451.3:c.-19G>A	*SHOX*	P/LP	LP
NM_000166.5:c.-17G>A	*GJB1*	P/LP	LP
Undetermined mechanism (n = 13)			
NM_005105.4:c.-19G>A	*RBM8A*	P	LP
NM_022787.3:c.-69C>T	*NMNAT1*	P	LP
NM_173546.2:c.-158C>T	*KLHDC8B*	P	P
NM_133433.4:c.-321_-320delinsA	*NIPBL*	P	LP
NM_014915.2:c.-126T>G	*ANKRD26*	P	LP
NM_014915.2:c.-126T>C	*ANKRD26*	P/LP	LP
NM_014915.2:c.-128G>T	*ANKRD26*	LP	LP
NM_014915.2:c.-128G>C	*ANKRD26*	LP	LP
NM_014915.2:c.-134G>A	*ANKRD26*	P/LP	LP
NM_000518.4:c.-50A>C	*HBB*	P	LP
NM_000133.3:c.-22T>C	*F9*	P	LP
NM_001551.3:c.-57_-55delinsAA	*IGBP1*	P	LP
NM_000166.5:c.-103C>T	*GJB1*	P	LP

[a]Variant HGVS, nomenclature is provided for the APPRIS-defined principal transcript.
LP, likely pathogenic; P, pathogenic; P/LP, P and LP, classifications across multiple submitters; uORF, upstream open reading frame; UTR, untranslated region.

a multifactorial disorder or had provisional associations. Functional enrichment analysis was used to investigate shared functions and common pathways between genes harboring at least one variant classified as P/LP in this study (Supplementary Table S6). Top enriched GO cellular components included ferritin complex (GO: 0070288; genes in analyzed set: *FTH1, FTL*; Benjamini–Hochberg adjusted (adj.) *p*-value = 0.010), germ cell nucleus (GO:0043073; *SMARCB1, SLC2A1, MLH1, BRCA1*; adj. *p*-value = 0.010), intracellular ferritin complex (GO:0008043; *FTH1, FTL*; adj. *p*-value = 0.010), and blood microparticle (GO:0072562; *HBB, HBA2, SLC2A1, PROS1, ENG*; adj. *p*-value = 0.014). Several Reactome pathways enriched among genes involving variants classified as P/LP in this study were related to hemoglobin and coagulation including "defective factor IX causes thrombophilia" (REAC:R-HSA-9672383; *F8, F9*; adj. *p*-value = 0.021), "intrinsic Pathway of Fibrin Clot Formation" (REAC:R-HSA-140837; *F8, PROS1, F9*; adj. *p*-value = 0.039), and "heme signaling" (REAC: R-HSA-9707616; *HBB, HBA2, MEF2C*; adj. *p*-value = 0.042). There was no statistically significant enrichment across GO molecular functions or GO biological processes.

3.5 Deep learning models capture the impact of curated pathogenic UTR variants

To further validate the curated set of P/LP 3′ and 5′ UTR variants, we investigated whether DL models trained for select UTR mechanisms could successfully capture the effects of the curated set of UTR variants in Tables 2, 3. These models and their respective mechanisms are as follows: transcriptional effects in the 5′ UTR (Enformer; Avsec et al., 2021), impacting ORF recognition by the translation machinery (FramePoolCombined; Karollus et al., 2021), and mRNA stability in the 3′ UTR (Saluki; Agarwal and Kelley, 2022). For each of the 5′ and 3′ UTR, a dataset was constructed by selecting putative benign variants observed in gnomAD v3.0 that were in the corresponding UTR for any transcripts with a P/LP variant and had a total allele frequency above 1% (see Section 2.4 for details; Supplementary Tables S3, S4). Though we tried to have at least one benign variant in the same UTR for every transcript, this was not always possible. This resulted in 68 P/LP variants and 24 putative benign variants in the 5′ UTR, and 26 P/LP and 67 putative benign variants in the 3′ UTR datasets.

Variant effect predictions were made by each model for the relevant UTR dataset. There was a significant difference (*p*-values ≤0.05 two-sided Mann-Whitney Wilcoxon test) between the absolute value of scores between all P/LP variants in their respective UTR combined and putative benign variants with all three DL models (Supplementary Figure S5). A further delineation in scores was obtained when stratifying by mechanism, with model-matched P/LP variants having significantly higher scores when compared to model-mismatched P/LP variants (FramePoolCombined *p*-value ≤0.001; Saluki *p*-value ≤0.005; Enformer *p*-value ≤0.005), in addition to putative benign variants (Figures 3A–C). This finding highlights that the additional consideration of variant mechanisms is informative. Furthermore, considering the absolute value of model prediction scores was superior to considering the original predictions in differentiating model-matched, model-mismatched, and putative benign variants, particularly for Enformer (Supplementary Figure S6). We anticipate that this may be due to variant pathogenicity being conferred through diverse mechanisms, including both expression increase and decrease. As such, the magnitude rather than the directionality of expression change, is more informative for stratifying pathogenic variants that operate through transcription as predicted by Enformer.

The UTR datasets were then used as benchmarks to evaluate the performance of DL models on the binary classification task of differentiating between putative benign variants and model-matched P/LP variants. All models achieved greater than random performance, with area under the curve (AUC) values for the receiver operator curve (ROC) ranging from 0.66 to 0.84 (Supplementary Figure S7). The performance was further improved when using the absolute values for all models' scores, again suggesting that the magnitude of effect has greater predictive power for pathogenicity when applied to the select mechanisms for which DL models are trained to predict (Supplementary Figure S8).

The availability of curated mechanistic data for this analysis provided valuable information for validating DL model predictions. For example, for P/LP variants that impact the ORF, variants introducing a novel upstream ORF (orange points; Supplementary Figure S6A) tended to have lower MRL predictions from FramePoolCombined, likely due to the majority of novel uORFs resulting in out-of-frame translation (Supplementary Table S2). In contrast, the two variants (NM_005144.5(HR):c.-218A>G, NM_000460.4(THPO):c.-47del) with the highest MRL predictions result in less inhibitory activity from an uORF, and greater translation of the main physiological ORF due to the disruption of an uORF, respectively. Further, there were nine P/LP variants in ANKRD26, of which eight were predicted by Enformer to result in increased expression (Supplementary Figure S6C; Supplementary Table S4). Though there was insufficient evidence to propose a definitive mechanism for all variants, curated evidence supported a mechanism in which variants likely result in loss of inhibitory regulatory action by RUNX1 and FLI1 (Supplementary Table S2). This mechanism aligns with what would be expected of thrombocytopenia-associated ANKRD26 variants, which have been documented to operate through a gain-of-function mechanism (Pippucci et al., 2011).

Lastly, to supplement the support provided by the models, we also turned to conservation-based methods which have been used to identify potential regulatory regions in the non-coding genome before the recent advances of DL models. Each variant in the constructed datasets had its position annotated with PhyloP conservation scores (Pollard et al., 2010). There was a significant difference in scores between all P/LP variants and the putative benign variants within each of the UTR regions (Figure 3D), with PhyloP ≥1.6 offering the best separation between all P/LP and putative benign variants, when combining variants in both UTRs. To gain insight into the conservation of certain non-coding regulation mechanisms, we looked at the conservation scores stratified by mechanism. For variants in the 5′ UTR, variants that impacted translation tended to occur at nucleotide positions that were more conserved (Supplementary Figure S9A, top panel).

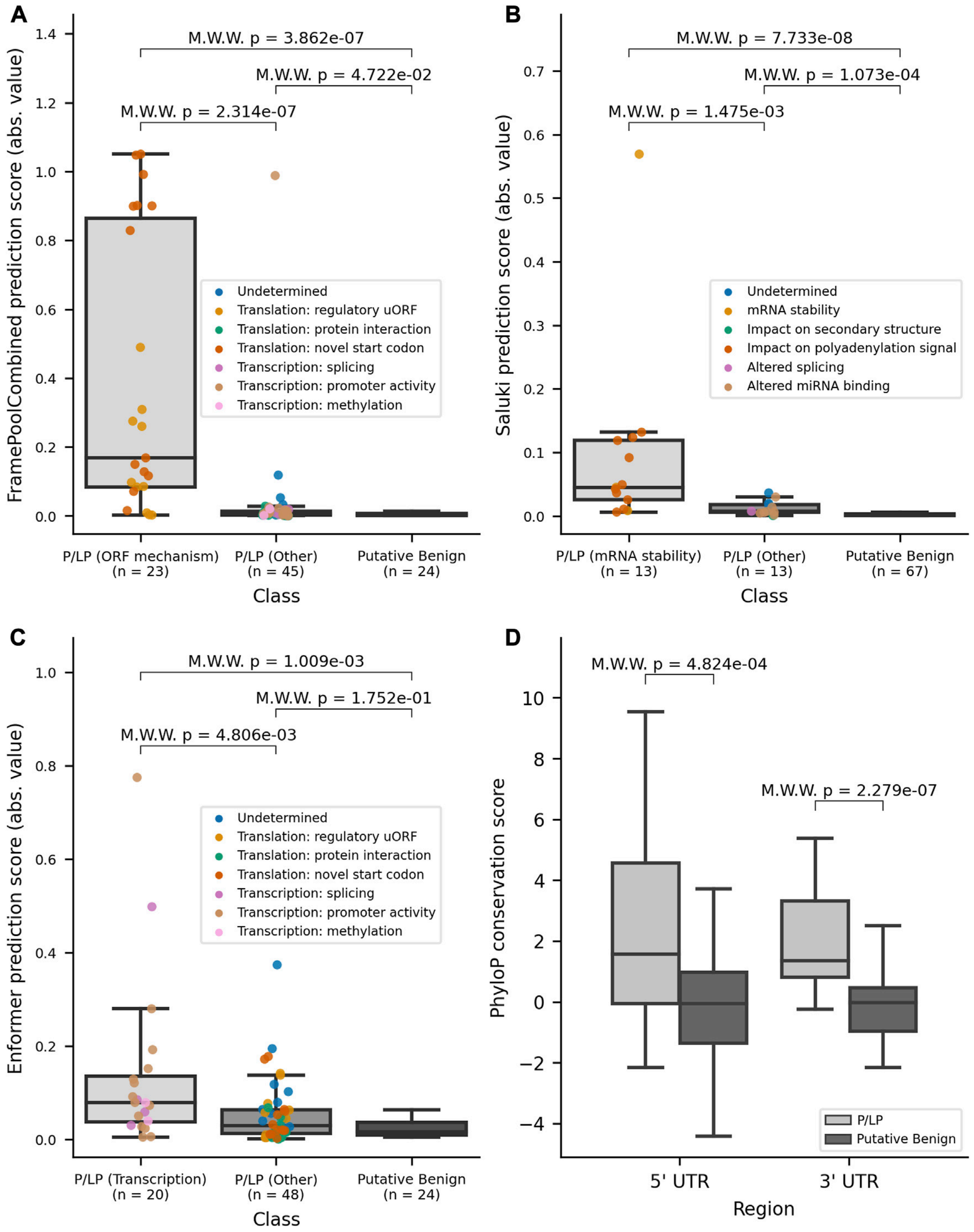

FIGURE 3
Variant effect scores (absolute values) from DL models for model-matched and model-mismatched variants and PhyloP conservation scores. **(A)** Distribution of the absolute values of scores for 5′ UTR variants as predicted by FramePoolCombined. **(B)** Distribution of the absolute values of scores for 3′ UTR variants as predicted by Saluki. **(C)** Distribution of the absolute values of scores for 5′ UTR variants as predicted by Enformer. **(D)** Distribution of PhyloP conservation scores for variants in the 5′ and 3′ UTR, stratified by pathogenicity. The boxplots have a center line for the median, the box limits are at the 25th and 75th percentiles, and the whiskers extend to 1.5x the 25th and 75th percentile values. Variants in the P/LP group have been colored based on their curated mechanism. M.W.W, Mann-Whitney Wilcoxon 2-sided test.

Variants that alter mRNA-protein interactions had the highest median PhyloP score, although this may be biased by the fact that all variants, except for one, are in *FTL* and occur in an iron response element (IRE), which is a conserved stem-loop (Addess et al., 1997). Variants that create novel upstream start codons were less conserved than those that impact existing uORFs, which is in line with existing findings that start and stop codons in existing uORFs tend to be highly conserved (Whiffin et al., 2020; Lee et al., 2021). Variants that affect transcription through altered promoter activity were located at sites that ranged between both high divergence and conservation, despite the importance of transcription factor binding motifs. For variants in the 3′ UTR, the ability to draw conclusions was limited by small sample size for most mechanisms. However, the mechanisms with the largest sample sizes, altered miRNA binding and polyadenylation signaling, both had median PhyloP scores greater than 1.5 (Supplementary Figure S9B, top panel), consistent with these mechanisms relying on binding motifs and consensus sequences. For both 3′ and 5' UTR, the B/LB variants have median PhyloP scores around zero. Contrastingly, the model-matched and model-mismatched P/LP variants have median PhyloP scores that are greater than zero (Supplementary Figure S9A, bottom panel; Supplementary Figure S9B, bottom panel).

4 Discussion

In this study, we aim to present a curated census of variants in the 3′ and 5′ UTR, with a focus on the mechanisms through which they confer pathogenicity. We curated a dataset of 26 and 68 variants classified as pathogenic or likely pathogenic in the 3′ and 5′ UTR, using a systematic approach and detailed evidence curation. Our analysis uncovered a considerable proportion of variants with interpretations of P/LP in ClinVar that lacked sufficient evidence to warrant such classifications. Reclassification of these variants highlights the need to exercise caution in utilizing public repositories without consideration of the level of evidence provided when evaluating variant pathogenicity. This finding is aligned with the ClinGen Sequence Variant Interpretation Working Group's recommendation to omit "reputable source criteria" (i.e., PP5 (supporting criterion 5), BP6 (supporting benign criterion 6)) from clinical variant classification practices, citing the strong preference for primary data over expert opinion (Biesecker and Harrison, 2018).

Beyond the variant sets themselves, our work provides insights into key challenges and considerations when classifying variants in non-coding regions. In their recommendations for adapting ACMG/AMP guidelines to non-coding variants, Ellingford et al. emphasize several important considerations, including classification in the context of the most clinically relevant transcript. Multiple methods have been developed to define biologically relevant transcripts, including the MANE collaboration's approach, based on evidence including evolutionary conservation and expression level (Morales et al., 2022), and APPRIS, which defines principality based on cross-species conservation and information related to protein structure and function (Rodriguez et al., 2022). While we classified variants in the context of the APPRIS-defined principal, recent work by Pozo and colleagues reported a high level of concordance between MANE and APPRIS approaches, with agreement documented in over 94% of genes evaluated (Pozo et al., 2022). In our experience, MANE Select and APPRIS principal transcripts were concordant in most cases. We identified a few instances (3′ UTR: n = 7 variants; 5′ UTR: n = 15 variants) where, despite residing in the UTR of the APPRIS-defined principal transcript, variants had a non-UTR impact on a transcript defined as either 'Select' or 'Plus Clinical' (i.e., transcripts not defined as 'Select', but in which known pathogenic variants have been reported) by MANE (Morales et al., 2022). These findings highlight the importance of defining and classifying variants in the context of the most biologically relevant transcript, as classifications in alternative transcripts may yield different conclusions with respect to pathogenicity.

There was an increasing number of variants in our curated set over time when considering the year of the most recent ClinVar submission for variants proceeding to curation in this study (Supplementary Figure S3A). This finding aligns with the increasing availability and implementation of WGS in recent years and consequent shift in focus towards the clinical relevance of variants in non-coding regions (Stavropoulos et al., 2016; Lionel et al., 2018; Kim et al., 2023). However, there was no clear trend in the time at which variants were most recently evaluated and whether they retained their P/LP status following curation and reclassification in this study (Supplementary Figure S3B). Classifications of both P/LP and VUS were assigned to variants with submissions as recent as 2022, highlighting persistent challenges stemming from a historical lack of interpretation guidelines tailored towards non-coding variants and lack of consistency in the stringency of assertion criteria application by ClinVar submitters. We anticipate that future implementation of tailored assertion criteria developed to expand the breadth of noncoding variants that can be meaningfully interpreted will improve the rigor with which variants are classified and submitted to public repositories.

The genes implicated by variants classified as P/LP in this study are largely involved in established gene-phenotype pairings, with all of the 49 unique genes having associated genetic phenotypes documented in OMIM, 45 of which (91.8%) involving a well-established monogenic condition (Supplementary Table S5). While the substrate available to perform functional enrichment analysis was limited by the relatively small number of genes implicated by variants classified in this study as P/LP, we did find significant enrichment for multiple GO cellular components and Reactome pathways relevant to hematological conditions. This finding aligns with the notion that such related conditions as hemoglobinopathies are among the most prevalent inherited disorders, with genetic mechanisms well studied as a result (Harteveld et al., 2022; Kountouris et al., 2022). Pathogenic variant spectra underlying hereditary bleeding disorders including hemophilia and von Willebrand disease have also been studied in detail (Pezeshkpoor et al., 2022), likely contributing to an overrepresentation of implicated genes in our dataset.

In addition to providing classifications as informed by applicable criteria, our datasets serve as a rich repository of mechanistic information accompanied by detailed summaries of functional evidence, where available. A diverse array of proposed mechanisms is represented across variants, operating at both the level of transcription and translation, albeit with varying levels of evidentiary support. Mechanistic information is relevant not only from the pathophysiological perspective, providing insight into functional underpinnings contributing to disease, but also with respect to the development and validation of mechanism-specific DL predictors of variant effects. Application of existing DL models to the curated 3′ and 5′ UTR datasets revealed a clear distinction in prediction scores when applied to variants operating through the mechanism for which the given predictor was designed. Furthermore, particularly in the case of Enformer, the absolute value of the scores was more informative in differentiating model-matched P/LP variants from the two other two groups. This finding reflects that pathogenicity can be conferred through both overexpression and lack of expression of a gene, as regulated through transcription. As such, the magnitude of change should be considered when evaluating the effect of multiple variants in aggregate, which may operate through both expression increase and decrease. Lastly, by combining PhyloP scores with this mechanistic information, we were able to provide insights into which non-coding mechanisms may be more greatly conserved in humans. Overall, our findings support the value of detailed and systematic curation of non-coding variant effects for the development, validation, and use of DL variant effect predictors in a mechanism-specific manner.

The datasets compiled and reported herein are not intended as an exhaustive resource of all P/LP 3′ and 5′ UTR variants documented in the literature, but rather a high-confidence set generated using a systematic approach consisting of both upstream bioinformatic processing and downstream curation and classification. Variants documented as P/LP in ClinVar were used as a reasonable starting point, although we recognize that non-coding variants are under-ascertained clinically and therefore under-reported in this database. Further, it is possible that true P/LP variants are currently documented in ClinVar as VUS, given historical challenges and lack of availability of classification guidelines tailored to variants in these regions. Increased implementation of WGS in the clinical setting and utilization of guidelines developed specifically for non-coding variants (Ellingford et al., 2022) will in theory increase the number of UTR variants documented as P/LP in ClinVar. These efforts can add to the breadth of the substrate available to contribute to the development of datasets similar to ours in the future. Also, towards our effort to develop and implement a standardized and reproducible approach, we considered only variants for which a UTR impact was mediated through the variant's effect on the APPRIS-defined principal transcript. It is therefore possible that variants with true pathogenic impacts mediated through UTRs on alternative transcripts, such as those residing within transcripts defined as MANE Plus Clinical, exist and are not represented in our datasets. Lastly, one limitation of our benchmarks for variant effect prediction is that not all P/LP variants could be matched to a putative benign variant in the same UTR of the same gene. As large-scale sequencing consortiums continue to focus on WGS, benchmarks for DL models can be improved upon to be more balanced and representative.

In conclusion, we present a high-confidence set of P/LP 3′ and 5′ UTR variants spanning a range of mechanisms and supported by detailed evidence curation and a systematic approach to classification. We anticipate these datasets to serve as valuable resources, given that variants documented in existing public repositories are not systematically curated in a consistent manner, and the reliability of such sources depends upon individual submitters. In generating this resource, we highlight key challenges and important considerations when conducting variant classification specifically in non-coding regions, including considering the potential for differential mechanistic impacts and classifications depending on the transcript considered. Findings from DL models further substantiate our classifications, with a distinction in scores supporting the relevance of mechanism-informed use of such predictors. These datasets will serve to support continued efforts towards developing and validating DL models designed to predict the impact of genetic variants residing in the 3′ and 5′ UTR.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

EB: Data curation, Formal Analysis, Investigation, Methodology, Visualization, Writing–original draft, Writing–review and editing. TL: Formal Analysis, Investigation, Methodology, Software, Visualization, Writing–original draft, Writing–review and editing. OW: Methodology, Writing–review and editing. TM: Conceptualization, Methodology, Project administration, Supervision, Writing–review and editing. DM: Conceptualization, Methodology, Writing–review and editing.

Funding

The authors declare financial support was received for the research, authorship, and/or publication of this article. This study was funded by Deep Genomics Inc.

Acknowledgments

We would like to thank Nasim Monfared for providing input on methodological aspects related to evidence curation, as well as Shreshth Gandhi and Albi Celaj for their input regarding the usage of deep learning models. We would also like to thank Brendan Frey, Laurence MacPhie, and Tom Peters for reviewing

the manuscript. An earlier version of this manuscript has been published on medRxiv (Bohn et al., 2023).

Conflict of interest

EB, TL, OW, and TM are employed by Deep Genomics Inc. DM was employed by Deep Genomics Inc. and is currently employed by Vevo Therapeutics. This study received funding from Deep Genomics in the form of salary support and covering computational costs. The funder was involved in the decision to submit for publication. All authors are entitled to a Deep Genomics Inc. stock option or are stockholders. DM is entitled to a Vevo Therapeutics stock option.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2023.1257550/full#supplementary-material

SUPPLEMENTARY FIGURE S1

Distribution of variants across genes. **(A)** Genes represented across 59 3′ UTR variants proceeding to classification. **(B)** Genes represented across 105 5′ UTR variants proceeding to classification.

SUPPLEMENTARY FIGURE S2

Number of ClinVar submissions for variants classified as P/LP and VUS in this study. This includes 26 P/LP variants and 33 VUS for 3′ UTR and 68 P/LP variants and 37 VUS for 5′ UTR. Boxplots have a center line for the median, box limits at the 25th and 75th percentiles, and the whiskers extended to 1.5x the 25th and 75th percentile values. The y-axis is log2 scaled. LP, likely pathogenic; M.W.W, Mann-Whitney Wilcoxon 2-sided test; P, pathogenic; VUS, variant of uncertain significance.

SUPPLEMENTARY FIGURE S3

Overview of variants classified as P/LP or VUS in this study based on the year of their most recent submission. **(A)** The distribution of variants classified as P/LP or VUS in this study based on the year of their most recent submission. In each facet, the values on the y-axis sum to one for each classification group (P/LP or VUS). **(B)** The proportion of variants classified as P/LP or VUS in our study based on the year of their most recent submission. Variants are plotted along the x-axis in terms of the year of their most recent submission (when available) and the bar is colored based on the proportion of variants out of the total for that year that were classified as P/LP or VUS LP, likely pathogenic; P, pathogenic; UTR, untranslated region; VUS, variant of uncertain significance.

SUPPLEMENTARY FIGURE S4

Comparison of the number of classified P/LP variants from ClinVar and Loss-of-Function Observed/Expected Upper bound Fraction (LOEUF) scores of genes with at least one P/LP variant classified in this study, genes with at least one VUS classified in this study (but no P/LP classified variants) and a background set of genes in ClinVar. The background set of genes are defined as all genes in ClinVar with at least one P/LP classified variant as per ClinVar that is not a somatic or copy number variant. **(A)** For genes in each of the 3′ and 5′ UTR dataset, those that had at least one classified P/LP or VUS variant have a significantly greater number of P/LP classified variants in ClinVar as compared to the background set of genes. Boxplots have a center line for the median, box limits at the 25th and 75th percentiles, and the whiskers extended to 1.5x the 25th and 75th percentile values. The y-axis is log10 scaled. **(B)** The LOEUF of genes in each of the sets is plotted. The constraint of genes that had P/LP or VUS curated variants in the UTRs are not significantly different from the background set of variants in ClinVar. LP, likely pathogenic; M.W.W, Mann-Whitney Wilcoxon 2-sided test; P, pathogenic; VUS, variant of uncertain significance.

SUPPLEMENTARY FIGURE S5

Variant effect scores (absolute values) from DL models for P/LP variants and PhyloP conservation scores, with variants stratified by pathogenicity. **(A)** Distribution of the absolute value of scores for 5′ UTR variants as predicted by FramePoolCombined. **(B)** Distribution of the absolute value of scores for 3′ UTR variants as predicted by Saluki. **(C)** Distribution of the absolute value of scores for 5′ UTR variants as predicted by Enformer. **(D)** Distribution of PhyloP conservation scores for variants in the 5′ and 3′ UTR, stratified by pathogenicity. The boxplots have a center line for the median, the box limits are at the 25th and 75th percentiles, and the whiskers extend to 1.5x the 25th and 75th percentile values. Variants in the P/LP group have been colored based on their curated mechanism. M.W.W, Mann-Whitney Wilcoxon 2-sided test.

SUPPLEMENTARY FIGURE S6

Variant effect scores (original predictions) from DL models for model-matched and model-mismatched variants and PhyloP conservation scores. **(A)** Distribution of scores for 5′ UTR variants as predicted by FramePoolCombined. **(B)** Distribution of scores for 3′ UTR variants as predicted by Saluki. **(C)** Distribution of scores for 5′ UTR variants as predicted by Enformer. **(D)** Distribution of PhyloP conservation scores for variants in the 5′ and 3′ UTR, stratified by pathogenicity. The boxplots have a center line for the median, the box limits are at the 25th and 75th percentiles, and the whiskers extend to 1.5x the 25th and 75th percentile values. Variants in the P/LP group have been colored based on their curated mechanism. M.W.W, Mann-Whitney Wilcoxon 2-sided test.

SUPPLEMENTARY FIGURE S7

Receiver operator (ROC) and precision-recall (PRC) curves showing the performance of deep learning models in classifying model-matched P/LP variants and putative benign variants. **(A)** The performance of FramePoolCombined on classifying P/LP variants that affect the ORF (n = 23) and putative benign variants in the 5′ UTR (n = 24). **(B)** The performance of Saluki on classifying P/LP variants that affect mRNA stability (n = 13) and putative benign variants in the 3′ UTR (n = 67). **(C)** The performance of Enformer on classifying P/LP variants that affect transcription (n = 20) and putative benign variants in the 5′ UTR (n = 24). AUC, area under the curve; TPR, true positive rate; FPR, false positive rate.

SUPPLEMENTARY FIGURE S8

Receiver operator (ROC) and precision-recall (PRC) curves showing the performance of deep learning models in classifying model-matched P/LP variants and putative benign variants, using absolute values of the scores. **(A)** The performance of FramePoolCombined on classifying P/LP variants that affect the ORF (n = 23) and putative benign variants in the 5′ UTR (n = 24). **(B)** The performance of Saluki on classifying P/LP variants that affect mRNA stability (n = 13) and putative benign variants in the 3′ UTR (n = 67). **(C)** The performance of Enformer on classifying P/LP variants that affect transcription (n = 20) and putative benign variants in the 5′ UTR (n = 24). AUC, area under the curve; TPR, true positive rate, FPR, false positive rate.

SUPPLEMENTARY FIGURE S9

PhyloP conservation scores for P/LP variants stratified by mechanism and putative B/LB variants. **(A)** Top: 5′ UTR P/LP variants grouped by their mechanism Bottom: 5′ UTR P/LP variants grouped into whether they were in a model-matched group (for any model) or have a mechanism in which there was no model available (model-mismatched or unknown) and 5′ UTR putative benign variants. **(B)** Top: 3′ UTR P/LP variants grouped by their mechanism. Bottom: 3′ UTR P/LP variants grouped into whether they were in a model-matched group (for any model) or have a mechanism in which there was no model available (model-mismatch or unknown) and 3′ UTR putative benign variants. The boxplots have a center line for the median, the box limits are at the 25th and 75th percentiles, and the whiskers extend to 1.5x the 25th and 75th percentile values. The dashed line represents a

PhyloP conservation score of 0. B, benign; LB, likely benign; LP, likely pathogenic; P, pathogenic.

SUPPLEMENTARY TABLE S1
3′ UTR variants: annotations, classifications and supporting evidence.

SUPPLEMENTARY TABLE S2
5′ UTR variants: annotations, classifications and supporting evidence.

SUPPLEMENTARY TABLE S3
The 3′ UTR benchmark of pathogenic and likely pathogenic variants (hg38) and putative benign variants from gnomAD v3.0.

SUPPLEMENTARY TABLE S4
The 5′ UTR benchmark of pathogenic and likely pathogenic variants (hg38) and putative benign variants from gnomAD v3.0.

SUPPLEMENTARY TABLE S5
OMIM genetic phenotypes, ClinVar P/LP variant counts, and gnomAD LOEUF scores for genes implicated by curated 3′ and 5′ UTR variants.

SUPPLEMENTARY TABLE S6
Pathways significantly enriched among genes with P/LP variants as classified in this study.

References

Addess, K. J., Basilion, J. P., Klausner, R. D., Rouault, T. A., and Pardi, A. (1997). Structure and dynamics of the iron responsive element RNA: implications for binding of the RNA by iron regulatory binding proteins. *J. Mol. Biol.* 274, 72–83. doi:10.1006/jmbi.1997.1377

Agarwal, V., and Kelley, D. R. (2022). The genetic and biochemical determinants of mRNA degradation rates in mammals. *Genome Biol.* 23, 245. doi:10.1186/s13059-022-02811-x

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., et al. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* 18, 1196–1203. doi:10.1038/s41592-021-01252-x

Avsec, Ž., Kreuzhuber, R., Israeli, J., Xu, N., Cheng, J., Shrikumar, A., et al. (2019). The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nat. Biotechnol.* 37, 592–600. doi:10.1038/s41587-019-0140-0

Biesecker, L. G., Harrison, S. M., and ClinGen Sequence Variant Interpretation Working Group (2018). The ACMG/AMP reputable source criteria for the interpretation of sequence variants. *Genet. Med.* 20, 1687–1688. doi:10.1038/gim.2018.42

Bohn, E., Lau, T., Wagih, O., Masud, T., and Merico, D. (2023). *A curated census of pathogenic and likely pathogenic UTR variants and evaluation of deep learning models for variant effect prediction*. medRxiv. doi:10.1101/2023.07.10.23292474

Chatterjee, S., and Pal, J. K. (2009). Role of 5′-and 3′-untranslated regions of mRNAs in human diseases. *Biol. Cell.* 101, 251–262. doi:10.1042/BC20080104

Chen, S., Francioli, L. C., Goodrich, J. K., Collins, R. L., Kanai, M., Wang, Q., et al. (2022). *A genome-wide mutational constraint map quantified from variation in 76,156 human genomes*. bioRxiv. doi:10.1101/2022.03.20.485034

Clark, M. M., Stark, Z., Farnaes, L., Tan, T. Y., White, S. M., Dimmock, D., et al. (2018). Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. *NPJ Genom. Med.* 3, 16. doi:10.1038/s41525-018-0053-8

Ellingford, J. M., Ahn, J. W., Bagnall, R. D., Baralle, D., Barton, S., Campbell, C., et al. (2022). Recommendations for clinical interpretation of variants found in non-coding regions of the genome. *Genome Med.* 1, 73. doi:10.1186/s13073-022-01073-3

French, J. D., and Edwards, S. L. (2020). The role of noncoding variants in heritable disease. *Trends Genet.* 36, 880–891. doi:10.1016/j.tig.2020.07.004

Gremer, L., Merbitz-Zahradnik, T., Dvorsky, R., Cirstea, I. C., Kratz, C. P., Zenker, M., et al. (2011). Germline KRAS mutations cause aberrant biochemical and physical properties leading to developmental disorders. *Hum. Mutat.* 32, 33–43. doi:10.1002/humu.21377

Harteveld, C. L., Achour, A., Arkesteijn, S. J., Ter Huurne, J., Verschuren, M., Bhagwandien-Bisoen, S., et al. (2022). The hemoglobinopathies, molecular disease mechanisms and diagnostics. *Int. J. Lab. Hematol.* 44, 28–36. doi:10.1111/ijlh.13885

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443. doi:10.1038/s41586-020-2308-7

Karollus, A., Avsec, Ž., and Gagneur, J. (2021). Predicting mean ribosome load for 5′ UTR of any length using deep learning. *PLoS Comput. Biol.* 17, e1008982. doi:10.1371/journal.pcbi.1008982

Kim, J., Woo, S., de Gusmao, C. M., Zhao, B., Chin, D. H., DiDonato, R. L., et al. (2023). A framework for individualized splice-switching oligonucleotide therapy. *Nature* 12, 828–836. doi:10.1038/s41586-023-06277-0

Kountouris, P., Stephanou, C., Lederer, C. W., Traeger-Synodinos, J., Bento, C., Harteveld, C. L., et al. (2022). Adapting the ACMG/AMP variant classification framework: A perspective from the ClinGen hemoglobinopathy variant curation expert panel. *Hum. Mut.* 43, 1089–1096. doi:10.1002/humu.24280

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 46, D1062-D1067–7. doi:10.1093/nar/gkx1153

Lee, D. S., Park, J., Kromer, A., Baras, A., Rader, D. J., Ritchie, M. D., et al. (2021). Disrupting upstream translation in mRNAs is associated with human disease. *Nat. Commun.* 12, 1515. doi:10.1038/s41467-021-21812-1

Lionel, A. C., Costain, G., Monfared, N., Walker, S., Reuter, M. S., Hosseini, S. M., et al. (2018). Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. *Genet. Med.* 20, 435–443. doi:10.1038/gim.2017.119

Morales, J., Pujar, S., Loveland, J. E., Astashyn, A., Bennett, R., Berry, A., et al. (2022). A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* 604, 310–315. doi:10.1038/s41586-022-04558-8

Moyon, L., Berthelot, C., Louis, A., Nguyen, N. T., and Roest Crollius, H. (2022). Classification of non-coding variants with high pathogenic impact. *PLoS Genet.* 18, e1010191. doi:10.1371/journal.pgen.1010191

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733–D745. doi:10.1093/nar/gkv1189

Petrazzini, B. O., Lopez-Bello, F., and Naya, H. (2022). *Clinical prediction of pathogenic variants in non-coding regions of the human genome*. medRxiv. doi:10.1101/2022.02.25.22271514

Pezeshkpoor, B., Oldenburg, J., and Pavlova, A. (2022). Insights into the molecular genetic of hemophilia A and hemophilia B: the relevance of genetic testing in routine clinical practice. *Hamostaseologie* 42, 390–399. doi:10.1055/a-1945-9429

Pippucci, T., Savoia, A., Perrotta, S., Pujol-Moix, N., Noris, P., Castegnaro, G., et al. (2011). Mutations in the 5′ UTR of ANKRD26, the ankirin repeat domain 26 gene, cause an autosomal-dominant form of inherited thrombocytopenia, THC2. *THC2. Am. J. Hum. Genet.* 88, 115–120. doi:10.1016/j.ajhg.2010.12.006

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* 20, 110–121. doi:10.1101/gr.097857.109

Pozo, F., Rodriguez, J. M., Martínez Gómez, L., Vázquez, J., and Tress, M. L. (2022). APPRIS principal isoforms and MANE select transcripts define reference splice variants. *Bioinformatics* 38, ii89–94. doi:10.1093/bioinformatics/btac473

Prior, T. W., and American College of Medical Genetics ACMG Laboratory Quality Assurance Committee (2009). Technical standards and guidelines for myotonic dystrophy type 1 testing. *Genet. Med.* 11, 552–555. doi:10.1097/GIM.0b013e3181abce0f

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., et al. (2019). g: profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* 47, W191-W198. doi:10.1093/nar/gkz369

Rehm, H. L., Berg, J. S., Brooks, L. D., Bustamante, C. D., Evans, J. P., Landrum, M. J., et al. (2015). ClinGen—The clinical genome resource. *N. Engl. J. Med.* 372, 2235–2242. doi:10.1056/NEJMsr1406261

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., et al. (2015). Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of medical genetics and genomics and the association for molecular Pathology. *Genet. Med.* 17, 405–424. doi:10.1038/gim.2015.30

Rodriguez, J. M., Pozo, F., Cerdán-Vélez, D., Di Domenico, T., Vázquez, J., and Tress, M. L. (2022). Appris: selecting functionally important isoforms. *Nucleic Acids Res.* 50, D54–D59. doi:10.1093/nar/gkab1058

Shah, N., Hou, Y. C., Yu, H. C., Sainger, R., Caskey, C. T., Venter, J. C., et al. (2018). Identification of misclassified ClinVar variants via disease population prevalence. *Am. J. Hum. Genet.* 102, 609–619. doi:10.1016/j.ajhg.2018.02.019

Stavropoulos, D. J., Merico, D., Jobling, R., Bowdin, S., Monfared, N., Thiruvahindrapuram, B., et al. (2016). Whole genome sequencing expands diagnostic utility and improves clinical management in pediatric medicine. *NPJ Genom. Med.* 1, 15012-–9. doi:10.1038/npjgenmed.2015.12

Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S., et al. (2003). Human gene mutation database (HGMD): 2003 update. *Hum. Mutat.* 2, 577–581. doi:10.1002/humu.10212

Steri, M., Idda, M. L., Whalen, M. B., and Orrù, V. (2018). Genetic variants in mRNA untranslated regions. *Wiley Interdiscip. Rev. RNA* 9, e1474. doi:10.1002/wrna.1474

Wells, A., Heckerman, D., Torkamani, A., Yin, L., Sebat, J., Ren, B., et al. (2019). Ranking of non-coding pathogenic variants and putative essential regions of the human genome. *Nat. Commun.* 10, 5241–5249. doi:10.1038/s41467-019-13212-3

Whiffin, N., Karczewski, K. J., Zhang, X., Chothani, S., Smith, M. J., Evans, D. G., et al. (2020). Characterising the loss-of-function impact of 5'untranslated region variants in 15,708 individuals. *Nat. Commun.* 11, 2523. doi:10.1038/s41467-019-10717-9

Xiang, J., Yang, J., Chen, L., Chen, Q., Yang, H., Sun, C., et al. (2020). Reinterpretation of common pathogenic variants in ClinVar revealed a high proportion of downgrades. *Sci. Rep.* 10, 331–335. doi:10.1038/s41598-019-57335-5