Nucleic Acids Research, 2026, **54**, gkag234
https://doi.org/10.1093/nar/gkag234
Data Resources and Analyses

An expanded reference catalog of translated open reading frames for biomedical research

Sonia Chothani[1,2,†], Jorge Ruiz-Orera[3,†], Jack A.S. Tierney[4], Michal I. Swirski[5], Hakon Tjeldnes[6], Leron W. Kok[7,8], Jim Clauwaert[9,10], Eric W. Deutsch[11], M. Mar Alba[12,13], Julie L. Aspden[14,15,16], Pavel V. Baranov[17], Ariel Alejandro Bazzini[18,19], Elspeth A. Bruford[20], Marie A. Brunet[21,22,23], Tristan Cardon[24], Anne-Ruxandra Carvunis[25,26,27], Claudio Casola[28], Jyoti Sharma Choudhary[29], Kellie Dean[30], Pouya Faridi[31,32], Ivo Fierro-Monti[33,34], Isabelle Fournier[24,35], Adam Frankish[4], Mark Gerstein[36,37,38,39,40], Norbert Hubner[3,41,42,43], Yunzhe Jiang[36,37], Manolis Kellis[44,45], Thomas F. Martinez[46,47,48], Gerben Menschaert[49], Pengyu Ni[36,37], Sandra Orchard[50], Xavier Roucou[51,23], Joel Rozowsky[36,37], Michel Salzet[24,52], Mauro Siragusa[53,54], Sarah Slavoff[55], Nicola Ternette[56], Juan Antonio Vizcaino[57], Aaron Wacholder[58,59], Wei Wu[60,61], Zhi Xie[62], Yucheng T. Yang[37,63], Robert L. Moritz[11], Eivind Valen[6], Jonathan Mudge[4,*], Sebastiaan van Heesch[7,8,*], John R. Prensner[9,10,*], Owen J.L. Rackham[64,*]

[1] Genome Institute of Singapore (GIS), Agency for Science, Technology and Research (A∗STAR), 60 Biopolis Street, Genome, Singapore 138672, Republic of Singapore
[2] Centre for Computational Biology, Cardiovascular and Metabolic Disorders, Duke-NUS medical school, Singapore 169857,Republic of Singapore
[3] Cardiovascular and Metabolic Sciences, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin 13125, Germany
[4] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom
[5] Institute of Genetics and Biotechnology, University of Warsaw, 02-106 Warsaw, Poland
[6] Department of Biosciences, University of Oslo, 0316 Oslo, Norway
[7] Princess Máxima Center for Pediatric Oncology, Utrecht 3584 CS, The Netherlands
[8] Oncode Institute, Utrecht 3521 AL, The Netherlands
[9] Department of Pediatrics, Division of Pediatric Hematology/Oncology, University of Michigan Medical School, Ann Arbor, MI 48109, United States
[10] Department of Biological Chemistry, University of Michigan Medical School, Ann Arbor, MI 48109, United States
[11] Institute for Systems Biology, Seattle, WA 98109, United States
[12] Hospital del Mar Research Institute (HMRIB), Barcelona 08003, Spain
[13] Catalan Institute for Research and Advanced Studies (ICREA), Barcelona 08010, Spain
[14] School of Molecular and Cellular Biology, Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, United Kingdom
[15] Leeds Omics, University of Leeds, Leeds LS2 9JT, United Kingdom
[16] Astbury Centre for Structural Molecular Biology, University of Leeds, Leeds LS2 9JT, United Kingdom
[17] School of Biochemistry and Cell Biology, University College Cork, Cork T12 K8AF, Ireland
[18] Stowers Institute for Medical Research, 1000 E 50th Street, Kansas City, MO 64110, United States
[19] Department of Molecular and Integrative Physiology, University of Kansas School of Medicine, KS City, KS 66160, United States
[20] HUGO Gene Nomenclature Committee, Department of Haematology, University of Cambridge Clinical School, Cambridge CB2 0PT, United Kingdom
[21] Medical Genetics Service, Pediatrics Department, University of Sherbrooke, J1E 4K8 Sherbrooke, Canada
[22] Cancer Research Institute University of Sherbrooke, IRCUS, J1E 4K8 Sherbrooke, Canada
[23] Centre de Recherche du Centre hospitalier universitaire de Sherbrooke, CRCHUS, J1E 4K8 Sherbrooke, Canada
[24] Univ. Lille, Inserm, CHU Lille, U1192 - Protéomique Réponse Inflammatoire Spectrométrie de Masse - PRISM, Lille F-59000, France
[25] Department of Computational and Systems Biology, University of Pittsburgh School of Medicine, Pittsburgh PA 15260,Pennsylvania
[26] Pittsburgh Center for Evolutionary Biology and Medicine (CEBaM), Pittsburgh PA 15260,Pennsylvania
[27] Present address: Institute of Molecular Systems Biology ETH Zurich, Zurich 8093, Switzerland
[28] Department of Ecology and Conservation Biology, Texas A&M University, College Station, TX 77843,United States
[29] The Institute of Cancer Research, London SW3 6JB, UK
[30] School of Biochemistry and Cell Biology, University College Cork, Cork T12 XF62, Ireland
[31] Centre for Cancer Research, Hudson Institute of Medical Research, Clayton, VIC, Australia

Received: July 24, 2025. **Revised:** February 11, 2026. **Accepted:** February 14, 2026

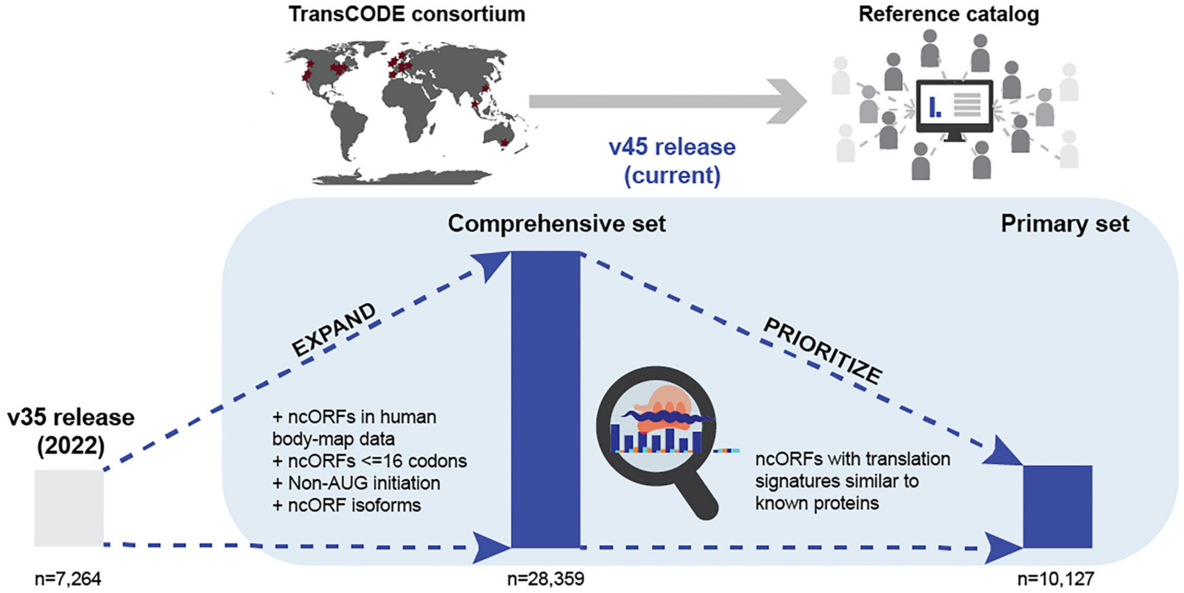
[32]Monash Proteomics & Metabolomics Platform, Department of Medicine, School of Clinical Sciences, Monash University, Clayton, VIC 3168, Australia
[33]EMBL-EBI, Wellcome Genome Campus, Cambridgeshire CB10 1SD, United Kingdom
[34]Biozentrum, University of Basel, Basel 4056, Switzerland
[35]Institut Universitaire de France, ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, 1 rue Descartes, 75231 PARIS CEDEX 05, France
[36]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, United States
[37]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, United States
[38]Department of Computer Science, Yale University, New Haven, CT 06520, United States
[39]Department of Statistics and Data Science, Yale University, New Haven, CT 06520, United States
[40]Department of Biomedical Informatics and Data Science, Yale University, New Haven, CT 06520, United States
[41]Charité-Universitätsmedizin Berlin, 10117 Berlin, Germany
[42]German Center for Cardiovascular Research (DZHK), partner site Berlin, Berlin 13347, Germany
[43]Helmholtz Institute for Translational AngioCardiosciences (HI-TAC), Max Delbrück Center for Molecular Medicine at Heidelberg University, Heidelberg 69120, Germany
[44]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139, United States
[45]Broad Institute of MIT and Harvard, Cambridge, MA 02139, United States
[46]Department of Pharmaceutical Sciences, University of California, Irvine, Irvine, CA 92617, United States
[47]Department of Biological Chemistry, University of California, Irvine, Irvine, CA 92617, United States
[48]Chao Family Comprehensive Cancer Center, University of California, Irvine, Irvine, CA 92617, United States
[49]BioBix, Lab for Bioinformatics and Computational Genomics, Faculty of Bioscience Engineering, Ghent University, Ghent 9000, Belgium
[50]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton CB10 1SD, United Kingdom
[51]Department of Biochemistry and Functional Genomics, Université de Sherbrooke, Sherbrooke, QC J1H 5N4, Canada
[52]Institut Universitaire de France (IUF), 75000 Paris, France
[53]Goethe University, Institute for Vascular Signalling, Centre for Molecular Medicine, Frankfurt am Main 60590, Germany
[54]CardioPulmonary Institute, Frankfurt am Main 60590, Germany
[55]Yale University Department of Chemistry, New Haven, CT 06520,United States
[56]University of Dundee, Dundee, Scotland DD1 5EH, United Kingdom
[57]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom
[58]Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, United States
[59]Pittsburgh Center for Evolutionary Biology and Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, United States
[60]Singapore Immunology Network (SIgN), Agency for Science, Technology and Research (A∗STAR), Singapore 138648, Republic of Singapore
[61]Department of Pharmacy & Pharmaceutical Sciences, National University of Singapore, Singapore 117543, Republic of Singapore
[62]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou 510060, China
[63]Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, 220 Handan Road, Shanghai 200433, China
[64]School of Biological Sciences, University of Southampton, Southampton SO17 1BJ, United Kingdom

[∗]To whom correspondence should be addressed. Email: jmudge@ebi.ac.uk
Correspondence may also be addressed to Sebastiaan van Heesch. Email: s.a.a.vanheesch@prinsesmaximacentrum.nl
Correspondence may also be addressed to John R. Prensner. Email: prensner@umich.edu
Correspondence may also be addressed to Owen J.L. Rackham. Email: O.J.L.Rackham@soton.ac.uk
[†]The first two authors should be regarded as Joint First Authors.

Abstract

Non-canonical (i.e. unannotated) open reading frames (ncORFs) have until recently been omitted from reference genome annotations, despite evidence of their translation, limiting their incorporation into biomedical research. To address this, in 2022, we initiated the TransCODE consortium and built the first community-driven consensus catalog of human ncORFs, which was openly distributed to the research community via Ensembl-GENCODE. While this catalog represented a starting point for reference ncORF annotation, major technical and scientific issues remained. In particular, this initial catalog had no standardized framework to judge the evidence of translation for individual ncORFs. Here, we present an expanded and refined catalog of the human reference annotation of ncORFs. By incorporating more datasets and by lifting constraints on ORF length and start codon, we define a comprehensive set of 28 359 ncORFs that is nearly four times the size of the previous catalog. Furthermore, to aid users who wish to work with ncORFs with the strongest and most reproducible signals of translation, we utilized a data-driven framework (i.e. translation signature scores) to assess the accumulated evidence for any individual ncORF. Using this approach, we derive a subset of 10 127 ncORFs with translation evidence on par with canonical protein-coding genes, which we refer to as the primary set. This set can serve as a reliable reference for downstream analyses and validation, with a particular emphasis on high quality. Overall, this update reflects continuous community-driven efforts to make ncORFs accessible and actionable to the broader research public, and further iterations of the catalog will continue to expand and refine this resource.

Graphical abstract



Introduction

Since the creation of the Ensembl-GENCODE (hereafter GENCODE) project 20 years ago [1, 2], gene annotations have been broadly divided based on whether a given gene is *protein-coding* or *non-coding*. Today, the canonical set of coding sequences (CDS) produced by GENCODE is a key resource for understanding the translated portion of the transcriptome. However, an increasing understanding of the prevalence of translation outside the set of canonical protein annotations has revealed the presence of additional functional elements. Here, perhaps no development in genome technologies has been more provocative than the adoption of ribosome profiling (Ribo-seq) [3], which has led to the identification of thousands of translated short unannotated open reading frames (typically 100 or fewer codons) [4–12]. We refer to these as translated non-canonical ORFs (ncORFs), specifically meaning ORFs with translation evidence that fall outside GENCODE's standard protein-coding annotations. Hereafter, we omit the term "translated" for brevity. In some cases, ncORFs are discovered to encode *bona fide* "microproteins" that can then be classified as canonical [13], leading to newly annotated protein-coding genes such as *MTLN* [14], *MRLN* [15], and *MIEF1-uORF/L0R8F8* [16–19]. However, many ncORFs may not meet the evolutionary constraints expected of canonical proteins [13, 20], yet potentially encode species-specific functional proteins [11, 21] or, alternatively, serve regulatory roles through their translation [22–25]. NcORF translation has been shown to occur broadly under physiological conditions, but there is great interest in the identification of ncORF translation products that may be recurrently or uniquely associated with specific cellular or pathological states [26–32]. Their widespread yet poorly understood nature highlights the need for standardized annotation to enable broader investigation by the scientific community.

Ultimately, the annotation of translated ncORFs depends on the accumulation and analysis of Ribo-seq data, the development of a new, more sophisticated classification schema for translation, and the additional inclusion of orthogonal protein evidence. To this end, in 2022, we initiated the global TransCODE Consortium, in collaboration with GENCODE [33], HGNC (HUGO Gene Nomenclature Committee) [34], UniProtKB (UniProt Knowledge Base) [35], PeptideAtlas [36], and other leading academic labs. This consortium combines expertise in annotation, Ribo-seq, gene evolution, non-coding RNA function, and mass-spectrometry proteomics. The result of Phase I of TransCODE was an initial GENCODE-supported catalog of 7264 ncORFs [20] previously identified using Ribo-seq as well as guidelines on how to interpret Ribo-seq and other experimental data nominating such ncORFs for annotation [22].

However, community-wide usage of these 7264 ncORFs has revealed both their utility as well as their limitations. It has become clear that this first catalog of ncORFs contains blind spots, making an updated catalog essential for the community. Several such blind spots resulted from the criteria used for ncORF inclusion, which were employed at that time for practical rather than biological reasons. For example, the 7264 ncORFs only include candidates that are: (i) >16 codons in length, (ii) initiated at AUG start codons, and (iii) the longest isoform in situations where multiple ncORFs share a large part of their amino acid sequence (≥90%). Yet, such filters likely hinder research on some ncORFs, particularly small ones. Indeed, it is known that ncORFs as small as two codons (i.e. "start–stop" or "minimal" ncORFs) [37] can function as regulatory elements, and recent studies in human tissues and cell lines have demonstrated widespread translation of ncORFs below our previous size cutoff [11, 12]. Moreover, translation products of ncORFs as short as eight codons can be presented on major histocompatibility complex molecules [27]. Similarly, non-AUG translation initiation events are increasingly recognized as biologically relevant, even within well-annotated protein-coding genes [38, 39, 40], with increasing evidence that they are especially common within 5′ untranslated regions (5′ UTRs) [12, 20]. Additionally, alternative splicing, the use of alternative initiation codons, or sequence variation may be specific to disease states [21, 26] or even to different stages of the cell cycle [41], making it possible to have multiple ncORF isoforms. Altogether, for the field of ncORF research to continue to grow, it is now critical to develop reference annotations of ncORFs further.

Here, we present an updated reference catalog of translated ncORFs detected by Ribo-seq based on GENCODE v45. As with the Phase 1 workflow, our goal was to use Ribo-seq data to map ncORFs to annotated transcripts; however, in this updated catalog, we incorporate additional studies made available since the production of our first catalog and remove the restrictive filters (as described above, Supplementary Fig. S1a and b). We first generate a "Comprehensive set" that substantially increases the number of ncORFs called from Ribo-seq from 7264 (Phase I) to 28 359. The new catalog adds thousands of ncORFs that were initially excluded from the first catalog, as well as ncORFs from the human body map translation project, extending the coverage of our resource to 11 additional primary human tissues and cell types [12]. We emphasize that "*Comprehensive*" refers to the inclusive approach taken in incorporating ncORFs into this set, as opposed to the sense that the catalog may be biologically complete (which remains hard to assess). As another advancement over our previous work, we specifically designate ncORFs with robust evidence of translation using two large pooled human Ribo-seq datasets into a "Primary set" of ncORFs. To do so, we utilize a standardized assessment of ncORFs using translation signature scores that act as quality-control metrics, namely, "P-sites in frame (PIF)", "Uniformity", and "Dropoff" scores (previously used in the human body map translation project [12, 42]). Based on these scores, we denote a subset of 10 127 out of 28 359 ncORFs as the "Primary" set, as these ncORFs have translation signature scores in the same range as canonical protein-coding sequences. The purpose of this Primary set (as opposed to the Comprehensive set) is to allow users to focus their analyses on ncORFs with the highest degree of translation evidence currently available. To ensure that these ncORF annotations can be widely applied across biomedical research, both sets are available at https://www.gencodegenes.org/pages/riboseq_orfs/.

Materials and methods

Collecting ncORFs from multiple datasets and mapping to GENCODE v45

We previously initiated the first catalog by selecting seven distinct Ribo-seq ncORF datasets from a range of human studies, each of which has been pivotal for genome-wide ncORF identification using Ribo-seq over the past decade [20]. Here, we included these seven datasets plus two new Ribo-seq ncORF datasets that were published in 2022 [12] and 2023 [11] (Supplementary Table S1).

When available, we retrieved exonic coordinates and sequences for these ncORFs. For datasets based on the older human genome assembly (GRCh37/hg19), we converted ORF coordinates to GRCh38/hg38 using UCSC Liftover. We compiled a total of 42 239 ncORFs. All translated ORF sequences were remapped to the Ensembl Release v.111 transcriptome (equivalent to GENCODE v45), collapsing identical genomic ncORF sequences. This resulted in 30 103 unique genomic sequences after excluding 1452 ORFs that could not be matched to any transcript and 10 684 identical ncORF regions. Our selection was limited to ncORFs found in lncRNAs, non-coding RNA transcripts within known protein-coding genes, alternative reading frames of CDS annotated within known protein-coding genes, and untranslated regions (UTRs) of coding transcripts annotated within known protein-coding genes. Therefore, we excluded 1744 ncORFs partially or totally overlapping annotated CDSs in the same frame (protein-coding or nonsense-mediated decay) or pseudogenes in any frame, as was done for the previous catalog [20]. These exclusions were necessary, as the ncORF datasets were based on older transcriptome versions, and newly annotated protein-coding sequences and pseudogenes are now available in GENCODE v45. Of 80 v35 catalog ncORFs that are now reclassified as CDSs, 49 previously overlapped incomplete protein-coding CDSs annotated with missing 5′ or 3′ regions, but we now completely classify these cases as CDS. Pseudogenes were removed because not all the original studies included only uniquely mapped reads, which could have led to false identifications of such sequences, although we acknowledge that many pseudogenes are known to be expressed and translated.

In total, we generated a Comprehensive set of 28 359 ncORFs. Of these, 7092 ncORFs were already part of the first catalog, and 21 267 were newly added. Unlike the first catalog, we did not apply any minimum length filter or exclude non-AUG start codons. Non-AUG ncORFs were predicted in five of the nine considered studies, with the exception of those using RiboTaper or ORFquant, which only annotated AUG-initiated ncORFs [4, 6, 9, 11]. Notably, in [7], ncORFs without an identifiable AUG start codon were defined from stop codon to stop codon, facilitating the inclusion of non-AUG-initiated ncORFs. Because of this, we acknowledge that some of the non-AUG ncORFs presented here may have inaccurate initiation site annotations.

Transcript assignment and classification of identified ncORFs

The vast majority of ncORFs overlapped several transcript models within a given gene and could not be uniquely mapped to a single host transcript. To resolve these ambiguities, we assigned the ncORF to the main isoform selected using MANE [43]. For the rest of the cases that could not be assigned to the MANE Select isoform, we chose the isoform with the highest APPRIS [44] score as the most probable isoform translating the ORF. In cases where multiple transcripts had comparable APPRIS scores, we further examined the Ensembl transcript support level scores and selected the one with the highest support, prioritizing protein-coding transcripts over non-coding ones. Of note, transcripts annotated as readthrough were given the lowest level of support and were assigned to an ncORF only when no other compatible isoform was available. All potential Ensembl transcript and gene IDs associated with each ORF, as well as the selected host transcripts, are detailed in Supplementary Table S2. While we made sure that the ncORFs do not partially or totally overlap the amino acid sequences of any annotated CDSs in the same region, it is possible that future releases of transcript annotations will include new models, and at least some ncORFs may be reannotated as new CDS extensions resulting from alternative splicing [45].

NcORFs were categorized into seven distinct types exactly as defined for the first catalog [20], i.e. based on the biotype of the assigned host isoform and the ORF position relative to known canonical protein-coding sequences.

lncRNA-ORFs: ncORFs found on long non-coding RNA genes.

PT-ORFs: ncORFs encoded by non-coding transcripts from protein-coding genes.

Upstream ORFs or uORFs: encoded within 5′ UTR sequences.

Upstream overlapping ORFs or uoORFs: encoded within 5′ UTR sequences and partially overlapping an already annotated CDS downstream in an alternative frame.

Internal ORFs or intORFs: completely overlapping within an already annotated CDS in an alternative frame.

Downstream ORFs or dORFs: encoded within 3′ UTR sequences.

Downstream overlapping ORFs or doORFs: encoded within 3′ UTR sequences and partially overlapping an already annotated CDS in an alternative frame.

Lastly, as a further advancement in this catalog, we now modify our GENCODE naming of these ORFs. We have decided not to proceed with the dual system used previously, e.g. c1riboseqorf1/c2norep2. This decision was made because we consider that ORFs should not be named according to their reproducibility, e.g. c2norep2, as this could potentially change. Therefore, for this catalog, we now use a simplified system based around only "c1riboseqorf1" and "c1riboseqorf2". We keep the names of ncORFs from the previous catalog where appropriate, i.e. for those that were called as replicated, and subsequently continuing numbering "upwards" to assign (i) new names for ncORFs found in the previous catalog that were not replicated (i.e. previously annotated as "norep") and (ii) names for novel ncORFs not included in the first catalog. The "norep" names that were used in the first catalog are also listed in the new datafile, as "legacy_names_v35" to allow comparison with the previous catalog. We anticipate that a standardized nomenclature system classifying ncORFs will be devised in due course.

Creating test datasets by pooling Ribo-seq data for primary set selection

In this study, we used two pooled Ribo-seq data resources. First, we obtained the previously published human body-map dataset [12], generated using the workflow as previously described. Read alignment was carried out using STAR [46] and only uniquely mapped reads were retained. From these, only read lengths between 27 and 30 nucleotides that showed >60% overall 3-nt periodicity when assessed in annotated protein-coding genes were selected to generate the P-site read files. P-site files from individual samples were then combined to create the pooled body map dataset, which was used to quantify translation signature scores. This resulted in a 1.3 billion P-site dataset across 11 cell types and tissues with an overall 3-nt periodicity of 85%.

Second, we used the Ribocrypt data repository. Metadata for publicly available ribosome profiling datasets were obtained from the Ribosome Data Portal [47]. Sequencing libraries corresponding to ribosome profiling samples were downloaded and processed with massiveNGSpipe (https://github.com/rc-biotech/massiveNGSpipe), as outlined in Swirski *et al.* 2025 (manuscript in submission). Briefly, raw FASTQ files were downloaded from the Amazon AWS NCBI SRA mirror (https://registry.opendata.aws/ncbi-sra). Adapters and barcodes were detected with a custom function. Their removal and length trimming were performed with fastp [48], using a minimum read-length cutoff of 20 nt. Identical reads were then collapsed, with copy number encoded in the read name, to reduce computational burden. Subsequently, reads were aligned to the human reference genome (hg38) with STAR [46], using the following non-default parameters: min.length = 20, mismatches = 3, trim.front (5′) = 0, max.multimap = 10, alignment.type = "Local". P-site positions were inferred with the shiftFootprintsPerExperiment function from ORFik [49]. For each library and read length, this procedure estimates the distance between the 5′ end of the read and the P-site. Only read lengths exhibiting clear triplet periodicity, as detected by a Fast Fourier Transform (FFT)-based algorithm, were subjected to the P-site offset detection algorithm described in detail by Tjeldnes *et al.* [49]. The human merged track was then created using massiveNGSpipe::pipeline_merge_org function. Only uniquely aligned reads were used to construct the merged track.

Ribo-seq coverage profiles across all included genes can be visually inspected under the link:

https://ribocrypt.org/?dff=all_merged-Homo_sapiens_modalities&frames_type=columns&kmer=1&colors=Color_blind&unique_align=TRUE.

In total, for Ribocrypt data, 3,892 human ribosome profiling libraries make up the merged track, amounting to 218 billion individual reads and 30.42 billion bona-fide footprints (unique alignments from periodic read lengths that do not map to rRNA).

Quantifying translation signature scores for ncORFs

We quantified translation signature scores for each ncORF in the Comprehensive set to test for evidence of translation in two test datasets generated by pooling Ribo-seq resources as described above.

Translation signature scores include three metrics as previously described ([12], [42]), namely (i) PIF, (ii) Uniformity, and (iii) Drop-off. P-sites in frame were calculated as the proportion of inferred P-site reads in the translating frame of the ncORF to the total inferred P-site reads in the ncORF. For the Uniformity calculation, each codon was tested for the proportion of P-site reads in Frame 1 to the total reads in the codon. The number of codons that had >33% of reads in Frame 1 were considered as having evidence for translation, and the proportion of codons that shows translation to the total number of codons was considered as the Uniformity. Last, for Drop-off score, a 15 bp transcript flanking window on either side of the stop codon was selected. P-site reads before and after were quantified, and Drop-off was calculated as the ratio of reads before the stop codon with respect to the total reads in both before and after flanking regions. These three scores were calculated for each ncORF in the Comprehensive set independently in each of the test datasets. These scores were also quantified previously [12] for known protein-coding ORFs to identify ncORFs with similar translation signatures. A normal distribution was fitted to the score distributions of annotated protein-coding ORFs, and thresholds were defined based on the 95th percentile for PIF and Uniformity, and the mean value was used for Drop-off score. For the human body-map dataset, the thresholds were determined as 75.89% for PIF, 71.3% for Uniformity, and 92% for Drop-off. For the Ribocrypt dataset, the thresholds were determined as 51.12% for PIF, 87.79% for Uniformity, and 88% for Drop-off scores, respectively. The ncORFs in the Comprehensive set that passed these thresholds in

Table 1. Comparison of v35 and v45 ncORF catalog

	v35 catalog (2022)	v45 catalog (2025)	
		Comprehensive set	Primary set
Total ncORFs	7264	28 359	10 127
Number of studies included	7	9	9
Number of Ribo-seq samples included	139	226	226
Overlap of v35 ncORFs	-	7092	1999
Length restriction	Yes (>16 codons)	No	No
Start-codon restriction	Yes (AUG only)	No	No
Isoforms included	Partially (clustered isoforms ≥90% shared codon sequence)	Yes	Yes
Standardized metrics applied (translation signature scores) using pooled human-body map or RiboCrypt data	No	No	Yes

either of the test datasets were considered to have high evidence for translation and selected for the Primary set. The ncORFs in the Primary set and the Comprehensive set were ranked for the Uniformity scores and the top 10 ncORFs from Primary set and the bottom 10 ncORFs from Comprehensive set were selected for visualization (P-sites per million or PPM <1 were skipped in the selection. PPM calculation described in the next section). The scores were able to select for individual ncORFs, which showed clear periodicity throughout the length of the ncORFs, thus discriminating ncORFs from ncORFs with just a single read stack or discontinuous periodicity or periodicity in the wrong frame (Supplementary Fig. S2 and Supplementary File S1).

Isoforms of ncORFs and overlap across studies

We compiled ncORFs from various studies, each using different datasets and methodologies. While identical ncORFs were treated as unique instances, we also identified many cases of ncORFs sharing part of their codon sequences. Therefore, ncORFs sharing any codon sequence were classified as ncORF isoforms. All isoforms were included in both the Comprehensive and Primary sets. A description of all and primary ncORF-overlapping isoforms is available at Supplementary Tables S5 and S6, respectively.

Testing limitations in Primary set

P-sites per million (PPM) values were calculated by normalizing ORF P-site counts by the ORF length and scaling to per-million total normalized counts. PPM values were quantified for each ncORF and annotated protein-coding ORF, and PPM >1 values were considered as a threshold for expression. NcORFs in the Comprehensive set and Primary set were split into length bins of 20 nucleotides (nt) each until 500 nt and into ncORF type categories such as uORF, dORF, uoORF, doORF, lncRNA-ORF, PT-ORF, and scores for such length bins and categories were presented using boxplots. Proportion of ncORFs in each bin and category was quantified and a chi-square test was carried out to test if the proportion of ncORFs that passed the PIF and Uniformity scores changed for ncORFs that are less than or more than 100 nt. PIF and Uniformity scores were tested in annotated ORFs across the same length bins as ncORFs for reference (Supplementary Fig. S4f). Genes with length <100 amino acids were selected, and distributions of PIF, Uniformity and Drop-off were plotted for a randomly selected transcript per gene. P-site profiles were plotted for known microproteins using human body map, and RiboCrypt data resource and translation signature scores were calculated for each microprotein (Supplementary Fig. S5).

Results

An expansion of the ncORF catalog

The new catalog of ncORFs is mapped to GENCODE v45 and is now referred to as the "v45" catalog. It builds upon our initial release in July 2022, which was mapped to GENCODE v35 [20] (see Table 1). The v45 catalog incorporates datasets from the same seven published studies as used for the first catalog, and according to the same selection principles (see Methods for details) includes two additional human Ribo-seq ncORF datasets that were published in 2022 [12] and 2023 [11] (Supplementary Table S1), leading to a 62% increase in ncORF-calling samples and adding 11 cell-type or tissue coverage. First, Chothani *et al.* [12], carried out an extensive ncORF human translation body map study, generating more than a billion inferred P-site reads using 167 Ribo-seq samples. These samples covered six primary human cell types (atrial fibroblasts, coronary artery endothelial cells, umbilical vein endothelial cells, hepatocytes, vascular smooth muscle cells, embryonic stem cells) and five human tissues (brain, adipose tissue, heart, skeletal muscle, kidney), resulting in the identification of 7767 ncORFs. Second, we incorporated 221 experimentally interrogated ncORFs reported by Sandmann *et al.* [11], previously not included, as each was 16 codons or fewer in length. The v45 catalog also takes a more inclusive approach to ncORF identification; we now include ORFs of 16 codons or less, those initiated by non-AUG codons, and all identified isoforms of each ncORF (Supplementary Fig. S1a and b).

In total, we identified 28 359 ncORFs that we refer to as the "Comprehensive set" of our v45 catalog (Fig. 1a and Supplementary Table S2). We retained most of the ncORFs (7092 out of 7264) from the first catalog, excluding 172 prior ncORFs because (i) they no longer mapped to GENCODE transcripts in v45 due to changes in transcript annotations, which now place these ncORFs partially or entirely in intergenic or intronic regions ($n = 65$), (ii) are now annotated as known proteins [13, 20] ($n = 25$), (iii) have been reassessed as extensions of known proteins ($n = 80$), or (iv) map to pseudogenes ($n = 2$) (Supplementary Fig. S1c and Supplementary Table S3). Building upon this, the Comprehensive set also includes: (i) 3124 additional isoforms (due

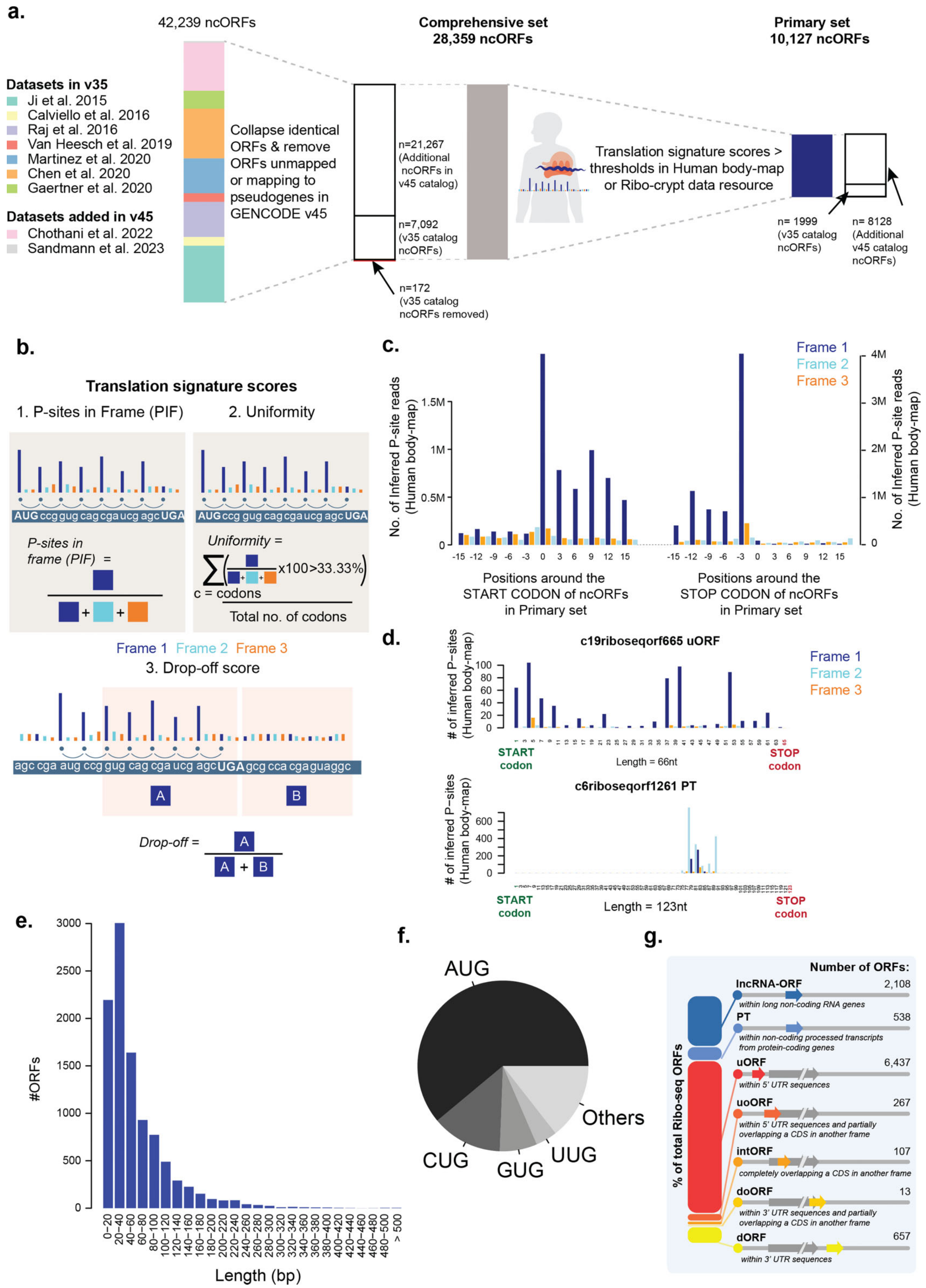


Figure 1. A Phase 2 GENCODE catalog for ncORFs using translation signatures. (**a**) Sankey plot based flow chart showing number of ncORFs in comprehensive set and selection to primary set. NcORFs were collected from nine studies [4–12] listed as Datasets, without any hard filters on length or start-codon, followed by obtaining a comprehensive set by collapsing identical ncORFs and removing ncORFs not mapping to GENCODE v45 or mapping to pseudogenes. Translation signature scores applied on each ncORF using human body-map or RiboCrypt to obtain *Primary* set. (**b**) Translation signature scores and its quantification. This included three metrics to define translation signature presence in each ncORF. *PIF* quantifies the proportion of Inferred P-sites in the translating frame with respect to the total inferred P-sites in the ncORF. *Uniformity* quantifies the percentage of codons in the ncORF that have > 33.33% inferred P-sites in the translating frame. *Drop-off score* quantifies the proportion of inferred P-sites before the stop-codon with respect to the total inferred P-sites in the translating frame in a 30bp window around the stop codon. (**c**) A bar plot showing an aggregated P-site profile (using the human body-map) around the start- and stop-codons of the *Primary* set of ncORFs (**d**) A bar plot showing P-site profile of two selected Ribo-seq ncORFs with high and low translation signature scores. (**e**) A histogram showing the length distribution of *Primary* set ncORFs. (**f**) A pie chart showing the distribution of start-codons identified for the *Primary* set ncORFs. (**g**) A stacked bar chart showing numbers of *Primary* set ncORFs across different ncORF types.

to alternative splicing or alternative translation initiation) of ncORFs that were present in the first catalog and (ii) a further 18 143 ncORFs that do not overlap with any of the ncORFs from the first catalog (Supplementary Fig. S1d).

Data-driven discrimination of a Primary set of ncORFs for improved end-user usability

In gene annotation, filtered subsets help users navigate large transcript collections, such as the 19 252 Matched Annotation from NCBI and EBI (MANE) transcripts within the 252 989 total transcripts in GENCODE v45 [43]. While MANE prioritizes functionally relevant, well-supported models, ncORFs currently lack equivalent functional information. Therefore, we rely on data-driven metrics to identify a filtered set of ncORFs. In our initial catalog (v35) [20], we tested for replication of ncORFs across multiple studies as a proxy for detection reliability (Supplementary Fig. S1e-g). However, this approach is constrained by substantial variability in both the quality of the individual Ribo-seq dataset as well as the nature and performance of computational workflows deployed. These issues are further confounded by the relative sparseness of Ribo-seq reads across an ncORF when using single samples or smaller Ribo-seq datasets [13, 20, 50]. To address these inconsistencies, here we define a Primary set by using two large pooled Ribo-seq resources [The Human Bodymap [12] and RiboCrypt (https://ribocrypt.org/)] as a way to assess the translational signatures, using three data-driven metrics. By ensuring that a primary ncORF satisfies these metrics, irrespective of the method of identification, we can ensure a more robust and consistent filtering strategy.

To construct the Primary set, we apply filters based on features derived from Ribo-seq data, selecting ncORFs whose translation signatures closely resemble those of canonical protein-coding genes. As a result, to standardize the evidence for ncORFs obtained from different studies, we independently assessed each ncORF to identify a subset of ORFs featuring the same degree of experimental evidence as observed for canonical CDSs. Specifically, we used Ribo-seq data-derived P-site profiles using the pooled human body-map data [12, 42] and pooled RiboCrypt data (https://ribocrypt.org/) to assess each individual ncORF in the Comprehensive set using previously designed metrics [12, 42]: (i) “PIF”, which determines the proportion of inferred P-site reads that are found in the translating frame; (ii) “Uniformity”, which quantifies the proportion of codons within the ncORF that have >33.33% reads in the translating frame; and (iii) the “Drop-off score”, which quantifies the proportion of P-sites in-frame before the stop codon to the total P-sites in-frame within the 30 bp window before and after the stop codon (Fig. 1b; see the “Materials and methods” section for details). An ncORF from the Comprehensive set was included in the Primary set if it passed the determined threshold in either the human body map or RiboCrypt data. Thresholds were determined for RiboCrypt dataset using annotated CDS with the same methodology as previously determined for the human body-map dataset [12] to identify ncORFs with similar translation signatures to annotated CDSs. As a result, we compiled a Primary set of 10 127 ncORFs (from the 28 359 in the Comprehensive set) (Fig. 1a and Supplementary Table S2; comparison to v35 release shown in Fig. 1a, Table 1, and Supplementary Fig. S1h and i). Upon aggregating P-sites around the start and stop of the ncORFs, we found clearer translation signatures for the Primary set as compared to the remainder of the Comprehensive set (Fig. 1c and d, Supplementary Figs S1j and S2, and Supplementary File S1).

ncORFs in the Primary set with alternative features

Of the 10 127 ncORFs in the Primary set, 5981 ncORFs (59.0%) are ≤16 codons in length (Fig. 1e and Supplementary Fig. S1h and i), and 3943 ncORFs (38.9%) start from non-AUG initiation codons (Fig. 1f), two thresholds applied in our initial v35 catalog. In the Primary set, we found that almost two-thirds of the identified ncORFs are upstream ORFs (uORFs, $n = 6437$, 63.6%, Fig. 1g) followed by ORFs in long non-coding RNA (lncRNA) genes (lncRNA-ORFs, $n = 2108$, 20.8%) (Potential explanations of these observations are discussed below in the Limitations section). For example, the Primary set includes the 4-codon-long uORF c22riboseqorf449 in *ATF4*, which is known to influence downstream CDS translation upon stress or disease [51]. The Primary set also includes 114 “minimal ORFs” that only contain a start and a stop codon. Such minimal ORFs could serve as regulatory elements [37] and are thus an important inclusion for reference annotations [52]. Within the Primary set, we found that 61.1% of ncORFs initiate at AUG codons, 13.3% at CUG, and 7.0% at GUG, similar to what has been previously reported by others [53] (Fig. 1f).

Next, we sought to define ncORF isoforms in the Primary set that can arise via alternative splicing and/or alternative initiation. These processes have historically been challenging to resolve in Ribo-seq data due to the sparsity of Ribo-seq coverage throughout the length of individual ncORFs [22], as well as complexities in the underlying transcript models. Furthermore, it is also expected that both processes can be subjected to differential expression, i.e. across cell types and different biological conditions. Within the Primary set, we identified 3303 ncORF isoforms that partially share the amino acid sequences of other ncORFs within the same set (Supplementary Fig. S3a), representing 32.6% of the total set. Of these, 185 and 2921 ncORFs shared identical start and stop positions, respectively, with other ncORFs (Supplementary Fig. S3a), while an additional 168 exhibited overlaps at either the start or stop codon of other ncORFs. Notably, only 332 ncORFs exhibited an overlap of ≥90% with other ncORFs, indicating that most ncORF isoforms contributed substantial stretches of unique codon sequences. We also identified 1259 non-AUG-initiating ncORFs as isoforms of AUG ncORFs and 3281 non-AUG-initiating ncORFs that do not have other AUG-initiating ncORFs included in the Primary set. For instance, *LMBRD2* encodes two overlapping uORFs with partially shared codon sequences that are translated into two small ncORF products of 39 and 42 codons, respectively, in the complete absence of a nearby AUG (Supplementary Fig. S3b). As another example, the *MIR1915HG* lncRNA contains two different ncORF variants starting with two alternative non-AUG and AUG codons that each display evidence of translation throughout their length (Supplementary Fig. S3c–e).

Assessment of gaps in the Primary set

In this resource, we have separated the process of ncORF identification (a process that is handled in the source manuscript of ncORF) and the creation of translation signatures (which are computed in this resource). The reason for this choice is that without a community-agreed pipeline for ncORF calling, we believe that independently scoring each ncORF using the same metrics applied to the same extensive Ribo-seq datasets

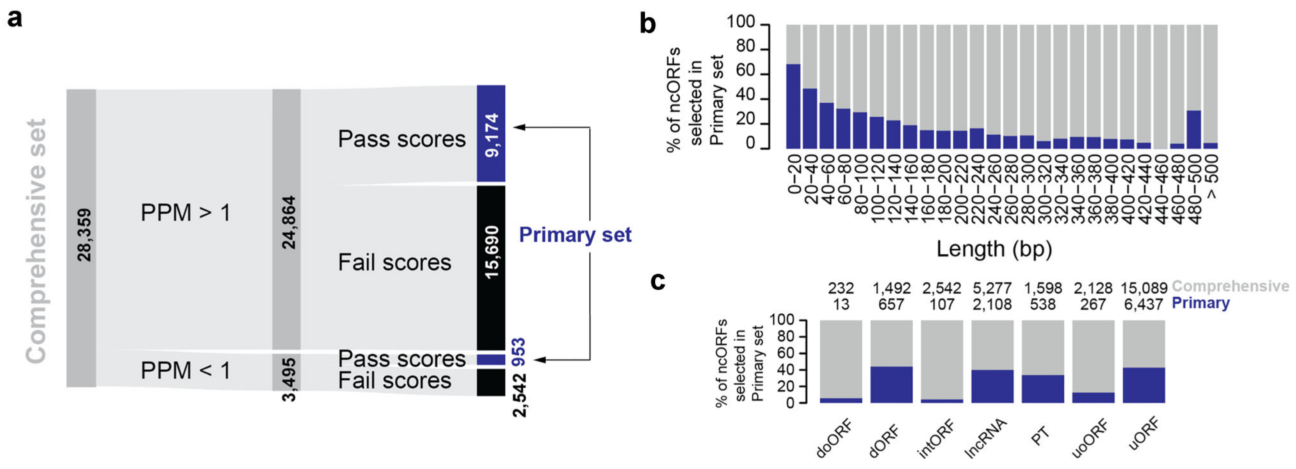


Figure 2. Coverage of translation signatures in Comprehensive vs Primary set. (**a**) A Sankey plot showing the number of ncORFs in the comprehensive set and corresponding ORFs that have P-sites per million (PPM) >1 in the pooled human body-map Ribo-seq data followed by the number of ncORFs that pass the translation signature scores. (b, c) Stacked barplot showing the proportion of ncORFs passing thresholds across length bins (**b**) and within each ncORF type (**c**). Gray: Comprehensive set ncORFs, blue: Primary set.

provides a reliable way to get a standardized Primary set. However, the mismatch between the samples from which ncORFs were called and the samples from which the scores were calculated might create a potential blind spot for the resource. To address this potential blind spot in our workflow, we tested whether the human body map or Ribocrypt data provided sufficient coverage to test all ncORFs, regardless of their source tissue or cell type. Evaluating each ncORF for their expression levels, we found that 87.7% of ncORFs in the Comprehensive set (24 864 out of 28 359) had a PPM >1 in either of the two resources, enough to confidently allow for their analysis. In contrast, 3495 ncORFs had low P-site coverage (PPM <1); despite their low expression, 27.3% still passed the translation signature scores and met the criteria for inclusion in the Primary set (Fig. 2a). However, it is important to note that some of the remaining ncORFs with low coverage that did not meet the criteria may exhibit high translation scores in other tissues or cell lines that were not analyzed in this study.

Furthermore, we tested whether ncORF length and type (Fig. 1e and g; see the "Materials and methods" section for details) influenced their chances of being included in the Primary set using the abovementioned Ribo-seq quality metrics. We did so because we reasoned that metrics like Uniformity could be biased in favor of short ncORFs, where ribosome footprint coverage across a small number of codons would still result in good uniformity. Similarly, we anticipated that ncORF types that overlap other ncORFs in alternative reading frames would score poorly on each of the three metrics. Indeed, we found that smaller ncORFs perform relatively well compared to longer ncORFs, indicating a length-dependent reduction in ncORF performance for the PIF and Uniformity metrics (Fig. 2b and Supplementary Fig. S4a and b). For example, 63.48% (12 303 out of 19 382) of ncORFs <100 nucleotides in length had a PIF score above the threshold in either of the resources, compared to 37.22% (3341 out of 8977) of ncORFs >100 nucleotides (Chi-square test, P-value 2.87×10^{-117}). Second, we found that ncORFs overlapping with CDS sequences in a different reading frame, such as uoORFs, doORFs, or intORFs, had limited representation in the Primary set as the scores assume a single ORF is translated in the given region (Fig. 2c and Supplementary Fig. S4 c, d, and e). Together, this shows that longer ncORFs, overlapping ORFs, and internal ORFs may be underrepresented in the Primary set, and investigators interested in these ncORF types can source these from the Comprehensive set instead.

Availability of the new GENCODE catalog

The Comprehensive and Primary sets are available at https://www.gencodegenes.org/pages/riboseq_orfs/. A track hub is available at https://ftp.ebi.ac.uk/pub/databases/ensembl/riboseq/TransCODE_GENCODEv45_riboseqORF_catalog/TransCODE-GENCODE_v45_RiboSeqORFs.hub.txt. A UCSC Browser session with appropriate Gencode comprehensive annotation is available at https://genome-euro.ucsc.edu/s/jackt/Phase2%2DTransCODE. Regardless of their current level of supporting data, we emphasize that every ncORF in the Comprehensive set has been published as being translated in at least one of the source studies and thus serves as a valuable reference for tracking and comparison in ncORF discovery, even if absent from the Primary set. We anticipate that some Comprehensive set ncORFs may gain additional supporting data over time, allowing them to be reclassified into the Primary set in future updates.

Discussion

Despite significant advances in our understanding of the human genome, gene and ORF annotations remain a work in progress. Redundancy in independent efforts for generating ncORF catalogs, as well as low overlap across published sets, has highlighted the need for a unified reference annotation. This need has been further amplified by ongoing efforts to characterize potential protein-coding genes from ncORF catalogs [13, 54]. Here, we have assembled an updated catalog of human ncORFs, seeking to address several current gaps in ncORF annotation. We now provide an expanded Comprehensive set of 28 359 ncORFs, aggregated from the several large-scale ribosome profiling studies, without any length or start-codon filters (considering all possible start codons reported in their respective studies), intended for efforts aiming to assess the ncORF search space in an inclusive manner. Additionally, we denote a Primary set of

10 127 ncORFs that have clear translation signatures in at least one large pooled human dataset, using a combination of PIF, Uniformity, and Drop-off scores. This subset may be useful for analyses where limiting false positives is critical, and we expect these highly supported ncORFs will serve as a "gold standard" set. We view the value of this updated catalog as threefold: (i) a unified resource for researchers; (ii) a high-quality Primary set that may serve as a field reference; and (iii) the development of standardized metrics to quantify translation evidence. These features may prove valuable for worldwide efforts to characterize the biological function of ncORFs [8, 26, 55–57]; the protein-coding potential of ncORFs [13], including initiatives such as the Understudied Proteins Initiative [58] and iMOP [59]; the evolutionary dynamics of ncORFs [11, 21, 60]; the tissue- and disease-specificity of ncORFs; and the potential for nucleotide variants to impact ncORFs in human health and disease.

Potential users of this updated ncORF catalog should note that the Primary set has many very short ncORFs; 59.0% of all annotated ncORFs are 16 codons or shorter. In part, this is to be expected from both a technical and biological perspective. Shorter ncORFs can achieve high PIF and Uniformity scores with fewer base pairs tested, and detection of shorter ncORFs may also be expected, given that longer ncORFs are more likely to trigger nonsense-mediated decay [61]. Many of these short ncORFs are found in the 5′ UTRs of mRNAs, and because mRNAs have higher transcript expression overall compared to lncRNAs, this may increase detection sensitivity for uORFs and uoORFs [62]. Additionally, 5′ UTRs are known hotspots for the emergence of microproteins and play a central role in translational regulation, including canonical protein expression control [63, 64]. An inverse relationship between uORF length and translational reinitiation efficiency has also been observed [23]; and recent population genetics studies provide further support for the functional relevance of these elements through evidence of sequence constraint [65]. Therefore, the enrichment of short uORFs in the Primary set is likely driven by biological factors, such as the overall higher level of translation for uORFs compared to lncRNA-ORFs, dORFs, and others [26], rather than false positives from the original ORF callers or artifacts from experimental protocols, as the included datasets are not biased to incorporate cells treated with antibiotics (e.g. homoharringtonine, lactimidomycin) that stall ribosomes in the 5′ UTR [66, 67]. Another key feature in this catalog is the inclusion of various isoforms of a given ncORF, such as different splicing isoforms or alternative start or stop positions. We included these for three main reasons. First, they likely reflect genuine transcriptional and translational complexity rather than technical artifacts. Second, comprehensive annotation supports efforts to characterize translation and enables tracking of isoforms that may be differentially detected across studies, because different RNA isoforms can expose alternatively translated ncORFs whose expression varies across cell types, tissues, or conditions. Third, given the current limitations in assigning function, excluding isoforms prematurely may overlook biologically relevant candidates not captured in the Primary set.

Nevertheless, we acknowledge the limitations of this work. First, our usage of published ncORF calls from a limited number of publications may have a bias toward ncORFs identified in specific samples, by specific research groups, or by certain Ribo-Seq analysis tools/protocols. Equally, the use of two large ribosome profiling resources used in the creation of our translational signatures may result in the exclusion of known (or real) ncORFs because of the relative under-representation of the cell types or tissues in which they are found. Second, trusting the ncORF calls made by other publications necessitates accepting decisions made about pre-processing and post-processing steps of the data, which may differ between research efforts. Notably, recent machine learning approaches employ no preprocessing steps for the data [68] or use sequence information to improve the prediction of Ribo-seq signal [69–71]. Third, scoring metrics such as PIF and Uniformity are biased toward certain ncORFs, particularly small ones (as noted above) and biased against those overlapping with other translated reading frames or CDSs. In overlapping ncORFs, periodic ribosome footprint signals become mixed, making their independent evaluation difficult and leading to an underestimation of their presence [22]. Fourth, start codon determination and assigning the correct RNA isoform to each ncORF remain challenging. Fifth, as GENCODE does not provide annotation for circRNAs, ncORFs originating from them are not incorporated. Last, here, we refer to these translation regions as ncORFs, noting that this is not a codified term in our databases and we anticipate a new ontology system will be developed in due course. For all of these reasons, we regard reference annotation to be an ongoing and iterative process and hope to be able to better address these potential concerns in future catalog iterations as more Ribo-seq datasets are incorporated and methods for ncORF detection improve.

Although the Primary set is unlikely to represent a complete survey of ncORFs with biological relevance, we expect the filtering approach taken here increases the rate of accurate ncORF detection and provides a subset of ncORFs with the highest utility for the life sciences and medical communities. Moreover, our resource provides a comprehensive overview of Ribo-seq evidence for all ncORFs included in both the Primary and Comprehensive sets. Researchers can customize their analyses by selecting their own thresholds for PIF, Uniformity, and Drop-off scores or by filtering based on additional metadata—such as expression levels or ORF length (Supplementary Table S2 and S4). This flexibility enables optimization for various downstream applications, including mass spectrometry, immunopeptidomics [13], evolutionary analyses [11, 13], testing for degradation signals [72, 73], interaction network mapping [74], and other strategies aimed at identifying functional candidates aligned with specific research goals.

In summary, we present here an expanded catalog of ncORFs aligned with the GENCODE v45 resource and support a Primary set of 10 127 ncORFs that have high-quality evidence of their translation. We believe that this work will be widely applicable to future biomedical research on ncORFs in diverse settings, including human health and disease.

Acknowledgements

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Ensembl is a registered trademark of EMBL. J.R.P. is the Ben and Catherine Ivy Foundation Clinical Investigator of the Damon Runyon Cancer Research Foundation [CI-127-24].

Author contributions: Sonia Chothani: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Visualization. Writing—original draft, Writing—review & editing. Jorge Ruiz-Orera:

Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Visualization, Writing—original draft, Writing—review & editing. Jack A.S. Tierney: Data curation, Investigation, Resources, Visualization, Writing—review & editing. Michal I Swirski: Data curation, Formal analysis, Investigation, Methodology,Writing—review & editing. Hakon Tjeldnes: Data curation, Formal analysis, Investigation, Methodology, Writing—review & editing. Leron W. Kok: Formal analysis, Investigation, Software, Writing—review & editing. Jim Clauwaert: Formal analysis, Investigation, Software, Writing—review & editing. Eric W. Deutsch: Data curation, Formal analysis, Investigation, Writing—review & editing. M. Mar Alba: Investigation, Writing—review & editing. Julie L. Aspden: Investigation, Writing—review & editing. Pavel V. Baranov: Investigation, Writing—review & editing. Ariel Alejandro Bazzini: Investigation, Writing—review & editing. Elspeth A. Bruford: Investigation, Writing—review & editing. Marie A. Brunet: Investigation, Writing—review & editing. Tristan Cardon: Investigation, Writing—review & editing. Anne-Ruxandra Carvunis: Investigation, Writing—review & editing. Claudio Casola: Investigation, Writing—review & editing. Jyoti Sharma Choudhary: Investigation, Writing—review & editing. Kellie Dean: Investigation, Writing—review & editing. Pouya Faridi: Investigation, Writing—review & editing. Ivo Fierro-Monti: Investigation, Writing—review & editing. Isabelle Fournier: Investigation, Writing—review & editing. Adam Frankish: Investigation, Writing—review & editing. Mark Gerstein: Investigation, Writing—review & editing. Norbert Hubner: Investigation, Writing—review & editing. Yunzhe Jiang: Investigation, Writing—review & editing. Manolis Kellis: Investigation, Writing—review & editing. Thomas F Martinez: Investigation, Writing—review & editing. Gerben Menschaert: Investigation, Writing—review & editing. Pengyu Ni: Investigation, Writing—review & editing. Sandra Orchard: Investigation, Writing—review & editing. Xavier Roucou: Investigation, Writing—review & editing. Joel Rozowsky: Investigation, Writing—review & editing. Michel Salzet: Investigation, Writing—review & editing.

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

J.R.P. has received research honoraria from Novartis Biosciences and Quantum-Si, and is on the scientific advisory board for, and receives research funding from, ProFound Therapeutics. J.L.A. is an advisor to Microneedle Solutions. G.M. is co-founder and CSO of OHMX.bio. P.F. is a member of the scientific advisory board of Infinitopes. A.-R. C. is a member of the advisory board of ProFound Therapeutics. P.V.B. is a cofounder and shareholder of Eirnabio Ltd. O.J.L.R is the founder, shareholder and scientific advisory board member of Mogrify Limited.

Funding

GENCODE (J.M.M., J.R.P., A.F.) are supported by the National Human Genome Research Institute grant (2U24HG007234-13). S.C. was supported by Khoo postdoctoral fellowship award, SingHealth and Singapore Ministry of Health's National Medical Research Council under Open Fund—Young Individual Research Grant (OFYIRG23jan-0034/MOH-001340). E.W.D and R.L.M were supported in part by the National Institutes of Health grants R24 GM148372 (E.W.D.), R01 GM087221 (E.W.D., R.L.M.), S10 OD026936 (R.L.M.), and by National Science Foundation grants DBI-2324882 (E.W.D.), DBI-1933311 (E.W.D.), and MRI-1920268 (R.L.M.). N.H. was supported by an ERC-advanced grant under the European Union Horizon 2020 Research and Innovation Program (AdG788970), the Deutsche Forschungsgemeinschaft (German Research Foundation—DFG; CRC/SFB-1470–B03), the Pathfinder Cardiogenomics Programme of the European Innovation Council of the European Union DCM-NEXT, and in part by a grant from the Chan Zuckerberg Foundation. E.B. acknowledges funding from the US National Institutes of Health (NIH) National Human Genome Research Institute (NHGRI) under award number U24HG003345. J.A.V. was supported by funding from Wellcome (grant number 223745/Z/21/Z), and from EMBL core funding. J.M.M. and A.F were supported by the Wellcome Trust (grant number 108749/Z/15/Z) and the European Molecular Biology Laboratory (EMBL). S.v.H. acknowledges funding from Fonds Cancers (FOCA, Belgium), Stichting Reggeborgh (the Netherlands), and Villa Joep. This publication is part of the project "Evolutionarily young microproteins in childhood brain cancer" (with project number VI.Vidi.223.022) of the research programme NWO talent programme Vidi, which is (partly) financed by the Dutch Research Council (NWO), awarded to S.v.H. Research reported in this publication was supported by Oncode Accelerator, a Dutch National Growth Fund project under grant number NGFOP2201, awarded to S.v.H. This work was co-financed by Oncode Institute, which is partly funded by the Dutch Cancer Society. J.R.P. acknowledges funding from the National Institutes of Health/National Cancer Institute (K08-CA263552-01A1); the V Foundation for Cancer Research (V2024-013); Tough2gether Foundation; Hyundai Hope on Wheels Foundation; the Yuvaan Tiwari Foundation; DIPG/DMG Research Funding Alliance; Book for Hope Foundation; CureSearch for Cancer Research; Morgan Adams Foundation; the Lindonlight Collective (GR-24-008 and GR-24-012); the Chad Carr Pediatric Brain Tumor Center; and the University of Michigan Rogel Cancer Center.

Data availability

The data underlying this article are available in https://www.gencodegenes.org/pages/riboseq_orfs/ as well as in the online supplementary material. Human body-map raw data can be downloaded from GEO superseries GEO: GSE182377. Ribocrypt data can be accessed at https://ribocrypt.org/. The ncORF lists from individual lists can be found in their respective publications https://doi.org/10.7554/eLife.08890, https://doi.org/10.1038/nmeth.3688, https://doi.org/10.7554/eLife.13328, https://doi.org/10.1016/j.cell.2019.05.010, https://doi.org/10.1038/s41589-019-0425-0, https://doi.org/10.1126/science.aay0262, https://doi.org/10.7554/eLife.58659.

References

1. Guigó R, Flicek P, Abril JF *et al.* EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biol* 2006;7:S2.1–31.

2. Harrow J, Denoeud F, Frankish A *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol* 2006;7:S4.1–9.
3. Ingolia NT, Ghaemmaghami S, Newman JRS *et al.* Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* 2009;324:218–23. https://doi.org/10.1126/science.1168978
4. van Heesch S, Witte F, Schneider-Lunitz V *et al.* The translational landscape of the human heart. *Cell* 2019;178:242. https://doi.org/10.1016/j.cell.2019.05.010
5. Ji Z, Song R, Regev A *et al.* Many lncRNAs, 5′ UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife* 2015;4:e08890. https://doi.org/10.7554/eLife.08890
6. Calviello L, Mukherjee N, Wyler E *et al.* Detecting actively translated open reading frames in ribosome profiling data. *Nat Methods* 2016;13.165. https://doi.org/10.1038/nmeth.3688
7. Martinez TF, Chu Q, Donaldson C *et al.* Accurate annotation of human protein-coding small open reading frames. *Nat Chem Biol* 2020;16.458. https://doi.org/10.1038/s41589-019-0425-0
8. Chen J, Brunner AD, Cogan JZ *et al.* Pervasive functional translation of noncanonical human open reading frames. *Science* 2020;367.1140. https://doi.org/10.1126/science.aay0262
9. Gaertner B, van Heesch S, Schneider-Lunitz V *et al.* A human ESC-based screen identifies a role for the translated lncRNA LINC00261 in pancreatic endocrine differentiation. *eLife* 2020;9:e58659. https://doi.org/10.7554/eLife.58659
10. Raj A, Wang SH, Shim H *et al.* Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *eLife* 2016;5:e13328. https://doi.org/10.7554/eLife.13328
11. Sandmann C-L, Schulz JF, Ruiz-Orera J *et al.* Evolutionary origins and interactomes of human, young microproteins and small peptides translated from short open reading frames. *Mol Cell* 2023;83:994. https://doi.org/10.1016/j.molcel.2023.01.023
12. Chothani SP, Adami E, Widjaja AA *et al.* A high-resolution map of human RNA translation. *Mol Cell* 2022;82:2885–99. https://doi.org/10.1016/j.molcel.2022.06.023
13. Deutsch EW, Kok LW, Mudge JM *et al.* High-quality peptide evidence for annotating non-canonical open reading frames as human proteins. bioRxiv, https://doi.org/10.1101/2024.09.09.612016, 9 September 2024, preprint: not peer reviewed.
14. Stein CS, Jadiya P, Zhang X *et al.* . Mitoregulin: a lncRNA-encoded microprotein that supports mitochondrial supercomplexes and respiratory efficiency. *Cell Rep* 2018;23:3710–20. https://doi.org/10.1016/j.celrep.2018.06.002
15. Anderson DM, Anderson KM, Chang CL *et al.* A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell* 2015;160:595. https://doi.org/10.1016/j.cell.2015.01.009
16. Rathore A, Chu Q, Tan D *et al.* MIEF1 Microprotein Regulates Mitochondrial Translation. *Biochemistry* 2018;57:5564–75. https://doi.org/10.1021/acs.biochem.8b00726
17. Brown A, Rathore S, Kimanius D *et al.* Structures of the human mitochondrial ribosome in native states of assembly. *Nat Struct Mol Biol* 2017;24:866. https://doi.org/10.1038/nsmb.3464
18. Andreev DE, O"Connor PB, Fahey C *et al.* Translation of 5′ leaders is pervasive in genes resistant to eIF2 repression. *eLife* 2015;4:e03971. https://doi.org/10.7554/eLife.03971
19. Delcourt V, Brunelle M, Roy AV *et al.* The protein coded by a short open reading frame, not by the annotated coding sequence, is the main gene product of the dual-coding gene MIEF1. *Mol Cell Proteomics* 2018;17:2402. https://doi.org/10.1074/mcp.RA118.000593
20. Mudge JM, Ruiz-Orera J, Prensner JR *et al.* Standardized annotation of translated open reading frames. *Nat Biotechnol* 2022;40:994–9. https://doi.org/10.1038/s41587-022-01369-0
21. Ruiz-Orera J, Miller DC, Greiner J *et al.* Evolution of translational control and the emergence of genes and open reading frames in human and non-human primate hearts. *Nat Cardiovasc Res* 2024;3:1217–35. https://doi.org/10.1038/s44161-024-00544-7
22. Prensner JR, Abelin JG, Kok LW *et al.* What can Ribo-seq, immunopeptidomics, and proteomics tell us about the noncanonical proteome? *Mol Cell Proteomics* 2023;22:100631. https://doi.org/10.1016/j.mcpro.2023.100631
23. Dever TE, Ivanov IP, Hinnebusch AG. Translational regulation by uORFs and start codon selection stringency. *Genes Dev* 2023;37:474–89. https://doi.org/10.1101/gad.350752.123
24. Dever TE, Ivanov IP, Sachs MS. Conserved upstream open reading frame nascent peptides that control translation. *Annu Rev Genet.* 2020;54.237. https://doi.org/10.1146/annurev-genet-112618-043822
25. Hinnebusch AG, Ivanov IP, Sonenberg N. Translational control by 5′-untranslated regions of eukaryotic mRNAs. *Science* 2016;352.1413https://doi.org/10.1126/science.aad9868
26. Hofman DA, Ruiz-Orera J, Yannuzzi I *et al.* Translation of non-canonical open reading frames as a cancer cell survival mechanism in childhood medulloblastoma. *Mol Cell* 2024;84:261–76.e18. https://doi.org/10.1016/j.molcel.2023.12.003
27. Ouspenskaia T, Law T, Clauser KR *et al.* Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. *Nat Biotechnol* 2022;40:209–17. https://doi.org/10.1038/s41587-021-01021-3
28. Camarena ME, Theunissen P, Ruiz M *et al.* Microproteins encoded by noncanonical ORFs are a major source of tumor-specific antigens in a liver cancer patient meta-cohort. *Sci Adv* 2024;10:eadn3628. https://doi.org/10.1126/sciadv.adn3628
29. Duhamel M, Drelich L, Wisztorski M *et al.* Spatial analysis of the glioblastoma proteome reveals specific molecular signatures and markers of survival. *Nat Commun* 2022;13:6665. https://doi.org/10.1038/s41467-022-34208-6
30. Cardon T, Franck J, Coyaud E *et al.* Alternative proteins are functional regulators in cell reprogramming by PKA activation. *Nucleic Acids Res* 2020;48:7864. https://doi.org/10.1093/nar/gkaa277
31. Cardon T, Fournier I, Salzet M. Unveiling a ghost proteome in the glioblastoma non-coding RNAs. *Front Cell Dev Biol* 2021;9:703583. https://doi.org/10.3389/fcell.2021.703583
32. Garcia-Del Rio DF, Derhourhi M, Bonnefond A *et al.* Deciphering the ghost proteome in ovarian cancer cells by deep proteogenomic characterization. *Cell Death Dis* 2024;15:712. https://doi.org/10.1038/s41419-024-07046-1
33. Mudge JM, Carbonell-Sala S, Diekhans M *et al.* GENCODE 2025: reference gene annotation for human and mouse. *Nucleic Acids Res* 2025;53:D966–75. https://doi.org/10.1093/nar/gkae1078
34. Seal RL, Braschi B, Gray K *et al.* Genenames.Org: the HGNC resources in 2023. *Nucleic Acids Res.* 2023;51:D1003–9. https://doi.org/10.1093/nar/gkac888
35. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res* 2025;53:D609–17. https://doi.org/10.1093/nar/gkae1010
36. Desiere F, Deutsch EW, King NL *et al.* The PeptideAtlas project. *Nucleic Acids Res* 2006;34:D655–8. https://doi.org/10.1093/nar/gkj040
37. Miyake T, Inoue Y, Shao X *et al.* Minimal upstream open reading frame of Per2 mediates phase fitness of the circadian clock to day/night physiological body temperature rhythm. *Cell Rep* 2023;42:112157. https://doi.org/10.1016/j.celrep.2023.112157
38. Fedorova AD, Kiniry SJ, Andreev DE *et al.* Thousands of human non-AUG extended proteoforms lack evidence of evolutionary selection among mammals. *Nat Commun* 2022;13:7910. https://doi.org/10.1038/s41467-022-35595-6
39. Pancsa R, Andreev DE, Dean K. The implication of non-AUG-initiated N-terminally extended proteoforms in cancer. *RNA Biol* 2025;22:1–18. https://doi.org/10.1080/15476286.2025.2498203

40. Menschaert G, Van Criekinge W, Notelaers T *et al.* Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events*. *Mol Cell Proteomics* 2013;12:1780–90. https://doi.org/10.1074/mcp.M113.027540

41. Ly J, Xiang K, Su K-C *et al.* Nuclear release of eIF1 restricts start-codon selection during mitosis. *Nature* 2024;635:490–8. https://doi.org/10.1038/s41586-024-08088-3

42. Menon D, Rackham O, Chothani S. Translation signature scores: data-driven approach to assess evidence for active translation. *Methods Mol Biol* 2026;2992:91–112.

43. Morales J, Pujar S, Loveland JE *et al.* A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* 2022;604:310–5. https://doi.org/10.1038/s41586-022-04558-8

44. Rodriguez JM, Rodriguez-Rivas J, Di Domenico T *et al.* APPRIS 2017: principal isoforms for multiple gene sets. *Nucleic Acids Res* 2018;46:D213–7. https://doi.org/10.1093/nar/gkx997

45. Ji HJ, Salzberg SL. Upstream open reading frames may contain hundreds of novel human exons. *PLoS Comput Biol* 2024;20:e1012543. https://doi.org/10.1371/journal.pcbi.1012543

46. Dobin A, Davis CA, Schlesinger F *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2012;29:15–21. https://doi.org/10.1093/bioinformatics/bts635

47. Tierney JAS, Świrski MI, Tjeldnes H *et al.* RiboSeq.Org: an integrated suite of resources for ribosome profiling data analysis and visualization. *Nucleic Acids Res* 2025;53:D268–74. https://doi.org/10.1093/nar/gkae1020

48. Chen S, Zhou Y, Chen Y *et al.* fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;34:i884–90. https://doi.org/10.1093/bioinformatics/bty560

49. Tjeldnes H, Labun K, Torres Cleuren Y *et al.* ORFik: a comprehensive R toolkit for the analysis of translation. *BMC Bioinformatics* 2021;22:336. https://doi.org/10.1186/s12859-021-04254-w

50. Chothani S, Ho L, Schafer S *et al.* Discovering microproteins: making the most of ribosome profiling data. *RNA Biol* 2023;20:943–54. https://doi.org/10.1080/15476286.2023.2279845

51. Vattem KM, Wek RC. Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. *Proc Natl Acad Sci USA* 2004;101:11269–74. https://doi.org/10.1073/pnas.0400541101

52. Tanaka M, Sotta N, Yamazumi Y *et al.* The minimum open reading frame, AUG-stop, induces boron-dependent ribosome stalling and mRNA degradation. *Plant Cell* 2016;28:2830–49. https://doi.org/10.1105/tpc.16.00481

53. Fritsch C, Herrmann A, Nothnagel M *et al.* Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. *Genome Res.* 2012;22:2208–18. https://doi.org/10.1101/gr.139568.112

54. Wacholder A, Deutsch EW, Kok LW *et al.* Community benchmarking and evaluation of human unannotated microprotein detection by mass spectrometry based proteomics. *Nat Commun* 2026;17:1241. https://doi.org/10.1038/s41467-025-68002-x

55. Prensner JR, Enache OM, Luria V *et al.* Noncanonical open reading frames encode functional proteins essential for cancer cell survival. *Nat Biotechnol* 2021;39:697–704. https://doi.org/10.1038/s41587-020-00806-2

56. Zheng C, Wei Y, Zhang P *et al.* CRISPR/Cas9 screen uncovers functional translation of cryptic lncRNA-encoded open reading frames in human cancer. *J Clin Invest* 2023;133:e159940. https://doi.org/10.1172/JCI159940.

57. Shi C, Liu F, Su X *et al.* Comprehensive discovery and functional characterization of the noncanonical proteome. *Cell Res* 2025;35:186–204.

58. Kustatscher G, Collins T, Gingras A-C *et al.* Understudied proteins: opportunities and challenges for functional proteomics. *Nat Methods* 2022;19:774–9. https://doi.org/10.1038/s41592-022-01454-x

59. Armengaud J, Cardon T, Cristobal S *et al.* Novel model organisms and proteomics for a better biological understanding. *J Proteomics* 2025;316:105441. https://doi.org/10.1016/j.jprot.2025.105441

60. Vakirlis N, Vance Z, Duggan KM *et al. De novo* birth of functional microproteins in the human lineage. *Cell Rep* 2022;41:111808. https://doi.org/10.1016/j.celrep.2022.111808

61. May GE, Akirtava C, Agar-Johnson M *et al.* Unraveling the influences of sequence and position on yeast uORF activity using massively parallel reporter systems and machine learning. *eLife* 2023;12:e69611. https://doi.org/10.7554/eLife.69611

62. Derrien T, Johnson R, Bussotti G *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* 2012;22:1775–89. https://doi.org/10.1101/gr.132159.111

63. Barbosa C, Peixeiro I, Romão L. Gene expression regulation by upstream open reading frames and human disease. *PLoS Genet* 2013;9:e1003529. https://doi.org/10.1371/journal.pgen.1003529

64. Morris DR, Geballe AP. Upstream open reading frames as regulators of mRNA translation. *Mol Cell Biol* 2000;20:8635. https://doi.org/10.1128/MCB.20.23.8635-8642.2000

65. Whiffin N, Karczewski KJ, Zhang X *et al.* Characterising the loss-of-function impact of 5′ untranslated region variants in 15,708 individuals. *Nat Commun* 2020;11:2523. https://doi.org/10.1038/s41467-019-10717-9

66. Ingolia NT, Lareau LF, Weissman JS. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 2011;147:789–802. https://doi.org/10.1016/j.cell.2011.10.002

67. Lee S, Liu B, Lee S *et al.* Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc Natl Acad Sci USA* 2012;109:E2424–32. https://doi.org/10.1073/pnas.1207846109

68. Clauwaert J, McVey Z, Gupta R *et al.* Deep learning to decode sites of RNA translation in normal and cancerous tissues. *Nat Commun* 2025;16:1275. https://doi.org/10.1038/s41467-025-56543-0

69. He J, Xiong L, Shi S *et al.* Deep learning prediction of ribosome profiling with Translatomer reveals translational regulation and interprets disease variants. *Nat Mach Intell* 2024;6:1314–29. https://doi.org/10.1038/s42256-024-00915-6

70. Clauwaert J, McVey Z, Gupta R *et al.* TIS Transformer: remapping the human proteome using deep learning. *NAR Genom Bioinform* 2023;5:lqad021. https://doi.org/10.1093/nargab/lqad021

71. Shao B, Yan J, Zhang J *et al.* Riboformer: a deep learning framework for predicting context-dependent translation dynamics. *Nat Commun* 2024;15;2011.

72. Kesner JS, Chen Z, Shi P *et al.* Noncoding translation mitigation. *Nature* 2023;617:395–402. https://doi.org/10.1038/s41586-023-05946-4

73. Yang H, Li Q, Stroup EK *et al.* Widespread stable noncanonical peptides identified by integrated analyses of ribosome profiling and ORF features. *Nat Commun* 2024;15:1932. https://doi.org/10.1038/s41467-024-46240-9

74. Garcia-Del Rio DF, Cardon T, Eyckerman S *et al.* . Employing non-targeted interactomics approach and subcellular fractionation to increase our understanding of the ghost proteome. *iScience* 2023;26:105943. https://doi.org/10.1016/j.isci.2023.105943

Received: July 24, 2025. **Revised:** February 11, 2026. **Accepted:** February 14, 2026