Article:
Mudge, JM, Ruiz-Orera, J, Prensner, JR et al. (32 more authors) (2022) Standardized annotation of translated open reading frames. Nature Biotechnology, 40. pp. 994-999. ISSN 1087-0156

https://doi.org/10.1038/s41587-022-01369-0

Standardized annotation of translated open reading frames

Jonathan M. Mudge[1+], Jorge Ruiz-Orera[2*], John R. Prensner[3,4,5*], Marie A. Brunet[6], Ferriol Calvet Riera[1], Irwin Jungreis[3,7], Jose Manuel Gonzalez[1], Michele Magrane[1], Thomas F. Martinez[8,9], Jana Felicitas Schulz[2], Yucheng T. Yang[10,11], M. Mar Albà[12,13], Julie L. Aspden[14,15], Pavel V. Baranov[16], Ariel Bazzini[17,18], Elspeth Bruford[1,19], Maria Jesus Martin[1], Lorenzo Calviello[20,21], Anne-Ruxandra Carvunis[22,23], Jin Chen[24], Juan Pablo Couso[25], Eric W. Deutsch[26], Paul Flicek[1], Adam Frankish[1], Mark Gerstein[27,28,29,30], Norbert Hubner[2,31,32], Nicholas T. Ingolia[33], Manolis Kellis[3,7], Gerben Menschaert[34], Robert L. Moritz[26], Uwe Ohler[35,36,37], Xavier Roucou[38], Alan Saghatelian[39], Jonathan Weissman[40,41,42] & Sebastiaan van Heesch[43*]*

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK

[2]Cardiovascular and Metabolic Sciences, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), 13125 Berlin, Germany

[3]Broad Institute of MIT and Harvard, Cambridge, MA, 02142, USA

[4]Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA

[5]Division of Pediatric Hematology/Oncology, Boston Children's Hospital, Boston, MA, 02115, USA

[6]Department of Pediatrics, Medical Genetics Service, Université de Sherbrooke, Sherbrooke, Québec, Canada

[7]MIT Computer Science and Artificial Intelligence Laboratory, 32 Vassar St, Cambridge, MA 02139, USA

[8]Clayton Foundation Laboratories for Peptide Biology, Salk Institute for Biological Studies, La Jolla, CA, USA

[9]Department of Pharmaceutical Sciences, University of California, Irvine, CA, USA

[10]Program in Computational Biology & Bioinformatics, Yale University, New Haven, CT 06520, USA
[11]Department of Molecular Biophysics & Biochemistry, Yale University, New Haven, CT 06520, USA
[12]Evolutionary Genomics Group, Research Programme on Biomedical Informatics, Hospital del Mar Research Institute (IMIM) and Universitat Pompeu Fabra (UPF), Barcelona, Spain
[13]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain.
[14]School of Molecular and Cellular Biology, Faculty of Biological Sciences, University of Leeds, LS2 9JT, UK.
[15]LeedsOmics, University of Leeds, UK
[16]School of Biochemistry and Cell Biology, University College Cork, Cork, T12 XF62, Ireland
[17]Stowers Institute for Medical Research, Kansas City, MO, USA
[18]Department of Molecular and Integrative Physiology, University of Kansas Medical Center, Kansas City, KS, USA
[19]Department of Haematology, University of Cambridge School of Clinical Medicine, Cambridge CB2 0XY, UK
[20]Functional Genomics Centre, Human Technopole, Milan, Italy
[21]Computational Biology Centre, Human Technopole, Milan, Italy
[22]Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, USA
[23]Pittsburgh Center for Evolutionary Biology and Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, USA
[24]Department of Pharmacology and Cecil H. and Ida Green Center for Reproductive Biology Sciences, University of Texas Southwestern Medical Center, Dallas, TX, USA
[25]Centro Andaluz de Biologia del Desarrollo, CSIC-UPO, Seville, Spain
[26]Institute for Systems Biology, Seattle, WA 98109, United States

[27]Program in Computational Biology & Bioinformatics, Yale University, New Haven, CT 06520, USA
[28]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511, USA
[29]Department of Computer Science, Yale University, New Haven, CT 06511, USA
[30]Department of Statistics & Data Science, Yale University, New Haven, CT 06511, USA
[31]Charité -Universitätsmedizin, 10117 Berlin, Germany
[32]DZHK (German Centre for Cardiovascular Research), Partner Site Berlin, 13347 Berlin, Germany
[33]Department of Molecular and Cell Biology and California Institute for Quantitative Biosciences, University of California, Berkeley, Berkeley, CA, 94720, USA
[34]Biobix, Lab of Bioinformatics and Computational Genomics, Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Ghent, Belgium
[35]Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), 10115 Berlin, Germany
[36]Department of Biology, Humboldt-Universität zu Berlin, Berlin, Germany.
[37]Department of Computer Science, Humboldt-Universität zu Berlin, Berlin, Germany
[38]Department of Biochemistry and Functional Genomics, Université de Sherbrooke, Sherbrooke, Québec, Canada
[39]Clayton Foundation Laboratories for Peptide Biology, Salk Institute for Biological Studies, La Jolla, CA, USA
[40]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, 02142
[41]Whitehead Institute for Biomedical Research, Cambridge, MA, 02142
[42]Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, MA, 02142
[43]Princess Máxima Center for Pediatric Oncology, Heidelberglaan 25, 3584 CS, Utrecht, the Netherlands

[*]These authors contributed equally

*Correspondence should be addressed to J.M.M, J.R.-O., J.R.P. & S.v.H. (e-mail: jmudge@ebi.ac.uk; jorge.ruizorera@mdc-berlin.de ; prensner@broadinstitute.org; s.vanheesch@prinsesmaximacentrum.nl)

To the editor: Ribosome profiling (Ribo-seq) has extended our understanding of the translational 'vocabulary' of the human genome, discovering thousands of open reading frames (ORFs) within long non-coding RNAs (lncRNAs) and presumed untranslated regions (UTRs) of protein-coding genes. However, reference gene annotation projects have been circumspect in their incorporation of these ORFs due to uncertainties about their experimental reproducibility and physiological roles. Yet, it is clear that certain 'Ribo-seq ORFs' make stable proteins, others mediate gene regulation, and many have medical implications. Ultimately, the absence of standardized ORF annotation has created a circular problem: while Ribo-seq ORFs remain unrecognised by reference annotation databases, this lack of recognition will thwart studies examining their roles. Here, we outline a community-led effort involving Ensembl / GENCODE, HGNC, UniProtKB, HUPO/HPP and PeptideAtlas to produce a standardized catalog of 7,264 human Ribo-seq ORFs, a path to bring protein-level evidence for Ribo-seq ORFs into reference annotation databases, and a roadmap to facilitate research in the global community.

Ribo-seq[1] provides an RNA sequencing-based readout of mRNA translation by isolating ribosome-bound RNA fragments of ~30 nucleotides in length. Sequencing of these fragments offers genome-wide footprints of ribosome–RNA interactions, detecting translated ORFs with sub-codon resolution[2–8]. Although Ribo-seq circumnavigates the experimental difficulties of working with protein molecules (e.g., using mass spectrometry (MS) analytical tools) and readily finds translations missed by *in silico* evolutionary methods, it does not demonstrate actual protein existence, and most translations do not show signs of constraint as coding sequences (CDS). A wide range of 'functional' scenarios are therefore plausible for Ribo-seq ORFs (**Table 1**).

Several public resources already process and/or display Ribo-seq datasets, including sORFs.org[9], GWIPS-viz[10] and Trips-Viz[11], whereas OpenProt[12] and nORFs.org[13] incorporate Ribo-seq into

whole translatome catalogs. Meanwhile, McGillivray *et al.* have produced a catalog of upstream ORFs (uORFs) with predicted biological activity[14]. Such efforts have made important contributions in Ribo-seq ORF interpretation. Nonetheless, the global scientific community is constrained by the absence of 'reference' gene annotation, which supports most large-scale genomics projects and provides the framework for human variant interpretation (**Fig. 1a, Supplementary Fig. 1**).

The creation of Ribo-seq annotations within existing reference gene and protein databases presents specific challenges that were not faced by previous cataloging efforts[9–13]. In particular, we must consider how these annotations can be integrated into the broad range of user workflows that are already supported by global annotation resources. For such reasons, reference annotation projects are generally conservative when it comes to the incorporation of new data types. Thus, rather than attempt to describe a 'maximal' set of potential Ribo-seq translations from the outset, our strategy is to build up a comprehensive resource in stages that is reciprocally improved by input from the scientific community (**Fig. 1b**).

Here, as 'Phase I' of this work, we present a consolidated catalog of Ribo-seq ORFs from seven publications[2–8] annotated onto GENCODE v35 (**Fig. 1c; Supplementary Tables 1–9**). A detailed description of the Ribo-seq datasets, our analysis methods, and ORF characteristics is available in the **Supplementary Methods**. We removed ORFs under 16 amino acids (aa) and those translated from non-ATG ('near-cognate') initiation codons, and merged redundant sense overlapping ORFs, resulting in a collated set of 7,264 unique ORFs (**Fig. 1c**). We classified these ORFs according to their spatial relationship with existing gene annotations (**Fig. 1d**), as presented in **Table 2**. We hope community usage of this catalog will help address the key technical and biological questions necessary to move this work into 'Phase II', where we aim to create a more comprehensive resource as outlined below.

For Phase I, we investigated repeated ORF identifications between studies, observing that 3,085 of 7,264 Ribo-seq ORFs were found by more than one publication (**Supplementary Fig. 2; Supplementary Tables 2,3**). However, whereas such 'reproducibility' would demonstrate consistency in Ribo-seq signal, it neither provides insights into biological function, nor indicates

that the 4,179 non-replicated ORFs are 'false'. A major goal of Phase II will be to incorporate a greater diversity of human cell types and tissues for improved estimates of ORF reproducibility, expression patterns, and potential cell-type specificity, along with further evaluation of criteria to quantify the technical confidence in Ribo-seq ORF calls.

Furthermore, Phase I excluded many translations by restricting the consensus set to ATG-initiated 'cognate' translations of at least 16 aa in length. Although these tiny ORFs may provoke skepticism in the absence of additional evidence — the smallest annotated human protein is 24 aa — there may be no lower size limit for a functional ORF[16]. For example, the tarsal-less (*tal*) gene produces a polycistronic transcript translated into proteins as short as 11 aa in several insect species[15]. Furthermore, the inclusion of ORFs initiated with near-cognate start codons can be complicated by ambiguous predictions of initiation site positions[17]. Ribo-seq following treatment with lactimidomycin or homoharringtonine, which inhibit translation elongation and result in accumulation of sequencing reads at the putative start sites, can help to identify near-cognate start sites[17,18]. Such datasets will be leveraged by our future Phase II efforts. For our current annotation resource, we have separately aggregated the Ribo-seq ORFs with near-cognate start codons or translations shorter than 16 codons (**Supplementary Fig. 3a–c; Supplementary Tables 4,5**), rather than including them in the Phase I catalog.

A core aim of Phase II will be to identify which Ribo-seq ORFs participate in cell physiology and how they do so. One aspect is distinguishing between cellular function mediated by a stable protein versus functionality imparted at the level of translation itself. We here use 'protein' as an umbrella term for protein, peptide and polypeptide, although we recognize that the terms polypeptide, micropeptide, or microprotein are commonly used for small protein molecules (**Table 2**). Because of the challenges of protein sequencing, evolutionary analysis has played a major historical role in ORF annotation, which is based on the assumption that the evolution of translated sequences is driven by selection at the protein level. Within our **Phase I** dataset, 75 Phase I replicated Ribo-seq ORFs (2.4%) present evidence of potential protein-level constraint as measured by PhyloCSF[19] (**Supplementary Fig. 3d-f**), 10 of which have now been classified as protein coding by GENCODE (**Supplementary Table 6**).

However, the evolutionary profile of many Phase I Ribo-seq ORFs remains hard to interpret. In part, this is because distinguishing ORF selection at the protein and DNA levels can be especially difficult for very small regions, noting that Ribo-seq ORFs are typically much smaller than known annotated proteins (**Supplementary Fig. 3g-j**). A second drawback is that evolutionary analysis cannot infer the protein-coding or regulatory potential of evolutionarily 'young' de novo Ribo-seq ORFs[20]. Reference annotation projects remain skeptical on the existence of proteins that are not deeply conserved, despite the fact that some young proteins clearly do participate in cellular physiology[20,21]. Furthermore, there is a substantial knowledge gap on the mode and tempo of regulatory ORF evolution. Here, genetic variation within human populations may provide insights. For example, Whiffin *et al*[22] recently used the gnomAD human variation dataset to identify 3,191 genes where uORF-perturbing variants are likely to be deleterious, thereby inferring the physiological importance of these translations. Meanwhile Neville *et al.*[23] used the same dataset to find aggregate evidence of selective pressure against deleterious variants in their nORFs.org catalog[13], especially pronounced for STOP-gain variants in uORFs. In prostate cancer, a recent analysis of 5′ UTR variants found regulatory roles for several uORFs[23].

While Ribo-seq ORFs may have regulatory roles irrespective of an encoded protein, the first step in confirming a protein-level physiological role for a Ribo-seq ORF is to demonstrate the existence of the protein in the cell. Mass spectrometry (MS) is a widely-accepted approach to catalogue the proteome, and its utility will be an important area of investigation for Phase II. At present, 609 of 7,264 Ribo-seq ORFs were reported to have support by published MS datasets (**Supplementary Table 10**). However, different groups use distinct methodologies and parameters for MS, and for Phase I these findings are simply reported in **Supplementary Tables 2 and 3** without further investigation. Reference annotation projects have historically favoured high stringency MS approaches, and the Human Proteome Organization (HUPO) / Human Proteome Project (HPP) — which aims to produce a full annotation of the human proteome — has published guidelines to standardize the nature of MS evidence required to annotate a human protein[24]. As one facet of our development of an MS workflow, these Ribo-seq ORFs have been added to the PeptideAtlas analytical pipeline, as used by HUPO. For Phase II, our projects will jointly examine the question of how best to use MS data to define which Ribo-seq ORFs produce

proteins. For reference annotation, we see two aspects to this: first, how to set standards for accepting and reporting potential MS support for a prospective Ribo-seq ORF protein; and second, how to define the point at which the body of evidence supports protein-coding annotation.

These aspects are illustrated by a preliminary analysis, which took advantage of the fact that 333 of our Ribo-seq ORFs are present in sequences previously queried by the PeptideAtlas workflow (**Supplementary Methods**). We find single-mapping peptide-spectrum matches (PSMs) for 13 Ribo-seq ORFs (**Supplementary Table 11**); all except one is supported by a single PSM, whereas most of the peptides identified are not fully tryptic (two examples are presented in **Supplementary Fig. 4**). The majority of observed PSMs derive from human leukocyte antigen (HLA) peptidome datasets, which is consistent with prior proteomic analyses demonstrating enrichment for peptides mapping to Ribo-seq ORFs in immunopeptidome data[25–27]. We emphasise that this preliminary analysis was not a full remapping of MS data and contained a fraction of the Ribo-seq ORFs; a larger, focused effort will be forthcoming.

There are multiple aspects as to why Ribo-seq ORFs and certain classes of canonical proteins are infrequently detected in MS data, which are summarized elsewhere[28]. One consideration for HUPO is that an MS-based 'canonical' protein assignment requires multiple PSMs, ideally based on non-overlapping tryptic peptides. Although we recognise the value of these guidelines, very small proteins may be 'less discoverable' by MS, especially due to a paucity of identifiable tryptic fragments[28]. Notably, nearly 1,500 protein-coding genes annotated by GENCODE, UniProt and HGNC do not presently have MS support recognised by HUPO[24]. Moving forward, we are committed to examining all potential protein-coding Ribo-seq ORF cases with full manual gene annotation processes, and this workflow will be expanded to include manual analysis of the peptide spectra by PeptideAtlas.

Although the value of MS in identifying translated proteins is indisputable, we believe a broader 'gold standard' for evidence should employ additional methodologies, such as epitope tagging combined with western blot imaging or endogenous antibody work; HUPO already incorporates such data in collaboration with the Human Protein Atlas[24]. Consideration will also be given to

emerging proteomics technologies, such as targeted proteomics workflows and immunopeptidomics, whereas progress is being made in medium-throughput functional screening assays. For example, recent large-scale studies have translated hundreds of Ribo-seq ORFs in mammalian cells through exogenous expression, finding that nearly 50% may stably produce proteins, despite little evidence of evolutionary constraint[2,6,27].

In addition to their evaluation as proteins or regulatory units, the reference annotation of Ribo-seq ORFs necessitates the creation of integrated workflows to interpret overlapping variants, and notwithstanding great community interest in this field, standardised approaches are not yet available. We emphasise that variant interpretation pipelines designed to classify CDS mutations may be unsuitable for Ribo-seq ORFs (**Table 1**), and that a minority of overlapping variants fall within sequences displaying amino acid-level constraint. Neville *et al.*[13] found that their nORFs.org catalog contains 48 Human Gene Mutation Database or ClinVar variants already considered pathogenic or likely pathogenic, despite the fact that they do not disrupt annotated CDSs. Although these variants may affect non-canonical ORFs, it will be important to define their mechanisms of action with experimental studies, as alternative explanations for pathogenicity are supported in certain cases, such as the creation of cryptic splice sites. After excluding variants in Ribo-seq ORFs that overlap annotated CDSs, a total of 1,142 single nucleotide variants present in the ClinVar database[29] were located within our aggregated set of Phase I Ribo-Seq ORFs (**Supplementary Methods**). Fewer than 2% of these variants have been classified as pathogenic or likely pathogenic, but this is likely to be an underestimate because the absence of pathogenesis is commonly inferred due to the absence of overlap with known coding features, and because ClinVar variant coverage is heavily skewed towards annotated CDSs.

Furthermore, there is major interest in the usage of Ribo-seq for the study of human disease. In particular, it is being widely used to explore the dynamics of translation in cancer cells with aberrant proteins as diagnostic markers or targets for immunotherapy[25,26,30]. At present, reference annotation projects do not attempt to distinguish aberrant translation from those events that contribute to ‘normal’ physiology. It will be important to deduce the fraction of Ribo-seq ORFs that encode proteins that exist in normal cellular conditions. Conversely, we envisage the value

of classifying potentially aberrant translations within Phase II through a distinct annotation framework.

Our intention is for the Ribo-seq Phase I catalog to be seen as a pragmatic interim solution to a long-term problem. We believe that reference annotation databases can advance both scientific and clinical research through the propagation and standardization of Ribo-seq ORF datasets, even — and perhaps especially — while the phenotypic impact of these features remains uncertain. As biological knowledge improves, this will support the development of more accurate annotations and variant interpretations, with the potential to yield substantial insights across all aspects of human biology. In this spirit, we hope the results of Phase I of this project will be useful and beneficial to the community and invite interested labs to join our future Phase II efforts.

Endorsement

HUPO/HPP Executive Committee members affirming support for this work are Rudolf Aebersold, Cecilia Lindskog Bergström, Yu-Ju Chen, Fernando Corrales, Lydie Lane, Siqi Liu, Edward Nice, Gilbert Omenn, Christopher Overall, Young-Ki Paik, Charles Pineau, Michael Roehrl, and Susan Weintraub. This work is further endorsed by Piero Carninci from Human Technopole and RIKEN.

Acknowledgments

AF, JMM, FCR and PF are supported by the Wellcome Trust [Grant number 108749/Z/15/Z], the National Human Genome Research Institute of the National Institutes of Health under award number 2U41HG007234 and the European Molecular Biology Laboratory. For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission. 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. MG and YTY are supported by National Human Genome Research Institute of the National Institutes of Health under award number 2U41HG007234. IJ and MK are supported by National Human Genome Research Institute of the National Institutes of Health under award number 2U41HG007234 and award number R01 HG004037. UniProt is supported by the National Human Genome Research Institute (NHGRI) of the National Institutes of Health under Award Number [U24HG007822], European Molecular Biology Laboratory core funds and the Swiss Federal Government through the State Secretariat for Education, Research and Innovation SERI. JRP is supported by the Harvard K-12 in Central Nervous System tumors (5K12 CA 90354-18), the Alex's Lemonade Stand Foundation Young Investigator Award (#21-23983), and the Musella Foundation for Brain Tumor Research. TFM is supported by the National Institutes of Health under award number F32GM123685. MMA acknowledges funding from the Spanish Government grant PGC2018-094091-B-I00 (MCI/AEI/FEDER,EU) and AGAUR grant 2017SGR01020. JC is supported by the National Institutes of Health Pathway to Independence Award (R00 GM134154) and the Cancer Prevent and Research Institute of Texas (RR200095). JSW is supported by HHMI. AAB is supported by the Stowers Institute for Medical Research and the US National Institutes of Health (R01 GM136849). ARC is supported by funds provided by the Searle Scholars program, the Sloan

Research Fellowship in Computational and Evolutionary Molecular Biology, and the National Institute of General Medical Sciences of the National Institutes of Health award number DP2GM137422. PVB wishes to acknowledge the support from the Investigator in Science Award [Grant number 210692/Z/18/Z] by SFI-HRB-Wellcome Trust Biomedical Research Partnership and from Russian Science Foundation [Grant number 20-14-00121]. NH is recipient of an ERC advanced grant under the European Union's Horizon 2020 research and innovation programme (grant agreement n° AdG788970). NH is supported by a grant from the Leducq Foundation (11 CVD-01). The work of the HGNC is funded by the Wellcome Trust (208349/Z/17/Z) and the National Human Genome Research Institute of the National Institutes of Health (under Award Number U24HG003345). XR and MAB are funded by the Canadian Institutes of Health Research PJT-175322. XR is funded as Canada Research Chair in functional proteomics and discovery of novel proteins. NTI is supported by the U.S. National Institutes of Health under award number R01 GM130996. RLM and EWD are supported by U.S. National Institutes of Health Grants R01GM087221, R24GM127667, U19AG023122, 1S10OD026936-01 and from U.S. National Science Foundation Grant DBI-1933311. JA is supported by Biotechnology and Biological Sciences Research Council UK (BB/S007407/1). MAB is supported by a Junior 1 fellowship from the Fonds de Recherche du Québec – Santé (FRQS).

Author contributions

JMM, JR-O, JRP, SvH conceptualized the work and supervised the international collaboration. JR-O, JMG, MM, MJM, FRC, EB, EWD, RLM, JMM performed data curation. All authors contributed to standardization of the data analysis approach. All authors contributed to discussions on Phase I and II of this effort and continue to provide scientific oversight. AF, PF, MJM, M, YTY, JRP, TFM, MMA, JC, JSW, AAB, AR, PVB, NH, XR, MAB, NTI, EB, EWD, RLM provided funding. JMM, JR-O, JRP, SvH wrote the original manuscript draft. All authors reviewed the manuscript and provided edits. All authors approved the final manuscript.

Competing interests

PVB is a co-founder of RiboMaps Ltd that provides Ribo-seq analysis as a commercial service and this includes identification of translated ORFs. ARC is a member of the scientific advisory

board for Flagship Labs 69, Inc. PF is a member of the scientific advisory boards of Fabric Genomics, Inc., and Eagle Genomics, Ltd. The other authors declare no competing interests.

1. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223 (2009).
2. van Heesch, S. *et al.* The Translational Landscape of the Human Heart. *Cell* **178**, 242–260.e29 (2019).
3. Ji, Z., Song, R., Regev, A. & Struhl, K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *Elife* **4**, e08890 (2015).
4. Calviello, L. *et al.* Detecting actively translated open reading frames in ribosome profiling data. *Nat. Methods* **13**, 165–170 (2016).
5. Martinez, T. F. *et al.* Accurate annotation of human protein-coding small open reading frames. *Nat. Chem. Biol.* (2019) doi:10.1038/s41589-019-0425-0.
6. Chen, J. *et al.* Pervasive functional translation of noncanonical human open reading frames. *Science* **367**, 1140–1146 (2020).
7. Gaertner, B. *et al.* A human ESC-based screen identifies a role for the translated lncRNA LINC00261 in pancreatic endocrine differentiation. *Elife* **9**, (2020).
8. Raj, A. *et al.* Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *Elife* **5**, (2016).
9. Olexiouk, V., Van Criekinge, W. & Menschaert, G. An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Research* vol. 46 D497–D502 (2018).

10. Michel, A. M., Kiniry, S. J., O'Connor, P. B. F., Mullan, J. P. & Baranov, P. V. GWIPS-viz: 2018 update. *Nucleic Acids Res.* **46**, D823–D830 (2018).
11. Kiniry, S. J., O'Connor, P. B. F., Michel, A. M. & Baranov, P. V. Trips-Viz: a transcriptome browser for exploring Ribo-Seq data. *Nucleic Acids Research* vol. 47 D847–D852 (2019).
12. Brunet, M. A. *et al.* OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucleic Acids Res.* **47**, D403–D410 (2019).
13. Neville, M. D. C. *et al.* A platform for curated products from novel Open Reading Frames (nORFs) prompts reinterpretation of disease variants. *Genome Res.* (2020).
14. McGillivray, P. *et al.* A comprehensive catalog of predicted functional upstream open reading frames in humans. *Nucleic Acids Res.* **46**, 3326–3338 (2018).
15. Galindo, M. I., Pueyo, J. I., Fouix, S., Bishop, S. A. & Couso, J. P. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol.* **5**, e106 (2007).
16. Vattem, K. M. & Wek, R. C. Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 11269–11274 (2004).
17. Lee, S. *et al.* Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U. S. A.* **109**, E2424–32 (2012).
18. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**, 789–802 (2011).
19. Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275–82 (2011).

20. Levy, A. How evolution builds genes from scratch. *Nature* **574**, 314–316 (2019).
21. Ruiz-Orera, J. & Albà, M. M. Translation of Small Open Reading Frames: Roles in Regulation and Evolutionary Innovation. *Trends Genet.* **35**, 186–198 (2019).
22. Whiffin, N. *et al.* Characterising the loss-of-function impact of 5'untranslated region variants in 15,708 individuals. *Nat. Commun.* **11**, 1–12 (2020).
23. Lim, Y. *et al.* Multiplexed functional genomic analysis of 5' untranslated region mutations across the spectrum of prostate cancer. *Nature Communications* vol. 12 (2021).
24. Adhikari, S. *et al.* A high-stringency blueprint of the human proteome. *Nat. Commun.* **11**, 5301 (2020).
25. Ouspenskaia, T. *et al.* Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. *Nat. Biotechnol.* (2021) doi:10.1038/s41587-021-01021-3.
26. Laumont, C. M. *et al.* Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Transl. Med.* **10**, (2018).
27. Prensner, J. R. *et al.* Noncanonical open reading frames encode functional proteins essential for cancer cell survival. *Nat. Biotechnol.* **39**, 697–704 (2021).
28. Omenn, G. S., Lane, L., Lundberg, E. K., Overall, C. M. & Deutsch, E. W. Progress on the HUPO Draft Human Proteome: 2017 Metrics of the Human Proteome Project. *J. Proteome Res.* **16**, 4281–4287 (2017).
29. Landrum, M. J. *et al.* ClinVar: improvements to accessing data. *Nucleic Acids Res.* **48**, D835–D844 (2020).
30. Chong, C. *et al.* Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat. Commun.* **11**, 1293 (2020).

411 **Table 1: Approaches to interpret Ribo-seq ORFs**

412

Possible cellular interpretation of Ribo-seq ORF translation	**Comments**
A Ribo-seq ORF encodes a 'missing' conserved protein	Ribo-seq ORFs may be recognised as canonical – in accordance with existing protein annotations – on the basis that the sequence of the protein they encode shows clear evidence of being maintained by evolutionary selection over a significant period of evolutionary time.
A Ribo-seq ORF encodes a taxonomically restricted protein.	Ribo-seq ORFs may encode proteins whose sequence and molecular activities are specific to one species or lineage. Evidence for selection or conservation across distant species or lineages is lacking for these ORFs, either because the protein sequence has diverged beyond recognition from its orthologues, or because the protein evolved recently from previously noncoding material and homologues do not exist in other species or lineages.
A Ribo-seq ORF regulates protein or RNA abundance.	Ribosome engagement of regulatory ORFs does not result in a protein product under selection but regulates the abundance of a canonical protein or RNA. This paradigm is well established for uORFs and uoORFs, as noted in **Table 2,** though it is applicable to other transcript scenarios. Regulatory ORFs

	may compete for ribosomes with their downstream canonical ORFs or produce nascent peptides that stall ribosomes, leading to the controlled 'dampening' of protein expression. Alternative modes of action, such as the induction of RNA decay pathways, the processing of small RNA precursors or the adjustment of RNA stability, have also been inferred.
A Ribo-seq ORF is the result of random translation.	The translation of some Ribo-seq ORFs may simply be 'noise'. Since translation has a high bioenergetic cost, a protein that results from random translation is likely to be translated at lower levels than a canonical CDS and evolve neutrally; it may also be unstable in comparison, and be potentially rapidly degraded. Nonetheless, it is theoretically possible that certain proteins do exist as stable 'junk' proteins, or that random translation events affect the expression of the canonical protein. The detection of random Ribo-seq ORFs is less likely to be reproducible.
A Ribo-seq ORF encodes a disease-specific protein.	This protein would not be produced under normal physiological homeostasis but could be of major interest for diagnostics and therapeutics**.** Insights are especially emerging in cancer biology, where transcription and translation are known to be dysregulated. This leads to the production of 'aberrant', possibly

	rapidly-degraded proteins that are commonly antigenic and presented on the cell surface by the HLA system, offering the prospect of neoantigens. In addition, antigens resulting from disease-specific dysregulated ribosome activity - sometimes called defective ribosomal products (DRiPs) - have also been explored.

Note: a given ORF may encompass several of these possibilities, e.g., a translated ORF that is both regulatory and implicated in disease neoantigen production.

Table 2. Terminology and categories of Ribo-seq ORFs

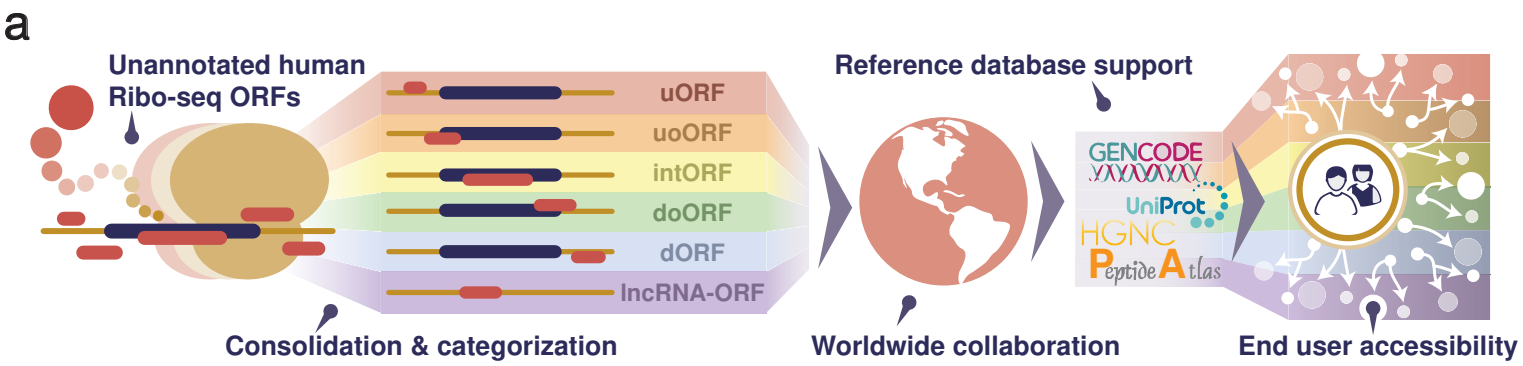
Term	Definition	Biological role
Ribo-seq ORF	Translated sequences identified by the Ribo-seq assay that have not already been annotated by reference annotation projects Also known as: non-canonical ORFs, alternative ORFs (altORFs), novel ORFs (nORFs). If <100 amino acids in size: small ORFs (smORFs), short ORFs (sORFs). Putative encoded proteins in smORFs/sORFs are also known as: microproteins, micropeptides, short ORF-encoded polypeptides (SEPs).	See below
Upstream ORFs (uORFs)	Translated sequences located within the exons of the 5' untranslated region (5' UTR) of annotated protein-coding genes.	Regulation of the translational efficiency of the downstream canonical protein. Cellular stress-related translation. May produce independently-functional proteins.
Upstream overlapping ORFs (uoORFs)	Translated sequences beginning in the 5' UTR of an	Similar to uORFs. Regulation translation of their

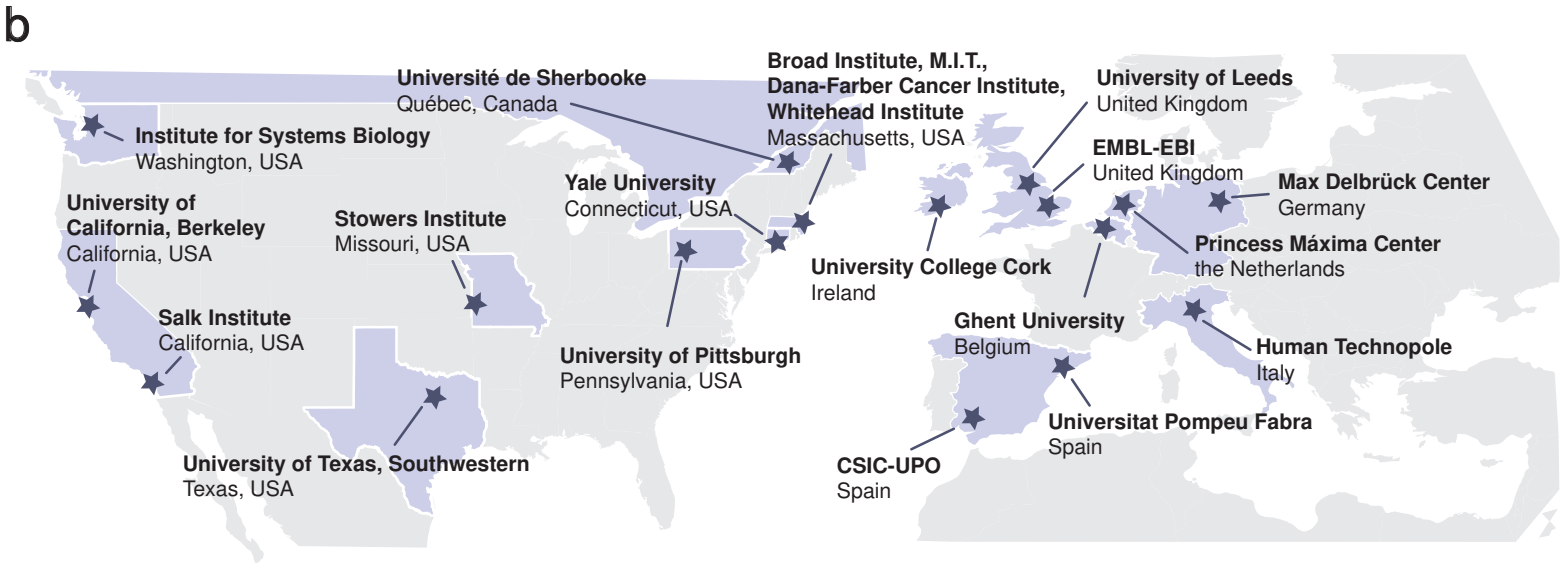
	annotated protein-coding gene and partially overlapping its coding sequence in a different reading frame.	overlapping CDS, but potentially stronger regulatory potential compared to uORFs. May produce independently-functional proteins.
Downstream ORFs (dORFs)	Translated sequences located within the 3' UTR of annotated protein-coding genes	Less commonly detected and generally poorly understood. May act as *cis* translational regulators.
Downstream overlapping ORFs (doORFs)	Translated sequences beginning in the genomic coordinates of an annotated CDS but continuing beyond the annotated CDS and terminating in the 3' UTR of the annotated protein-coding gene.	Similar to dORFs
Internal out-of-frame ORFs (intORFs)	Translated sequences located on the mRNA of an annotated protein-coding gene and completely encompassed within the canonical CDS, but translated via a different reading frame. Also known as: altCDSs, nested ORFs, dual-coding regions.	May regulate translation similar to uORFs in some cases. Detection of intORFs with Ribo-seq is possible but difficult due to the convolution of triplet periodicity signals from two reading frames; it largely depends on the length and translation level of the intORF relative to the overlapping canonical CDS.

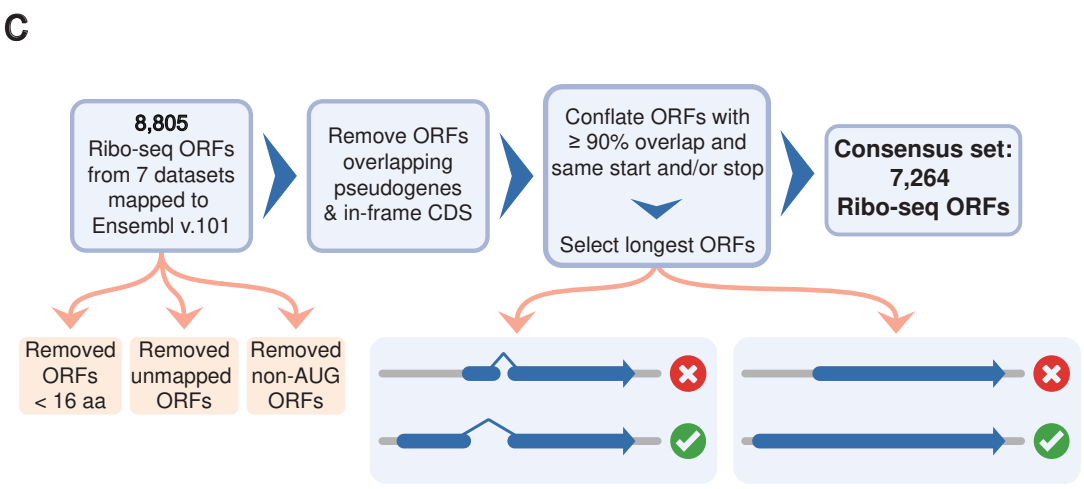
Long non-coding RNA ORFs (lncRNA-ORF)	Translated sequences located within transcripts currently annotated as long non-coding RNAs (lncRNAs), including long intervening/intergenic noncoding RNAs (lincRNAs), long non-coding RNAs that host small RNA species (encompassing microRNAs, snoRNAs, etc), antisense RNAs, and others	May produce independently-functional proteins. Typically lack strong sequence conservation.

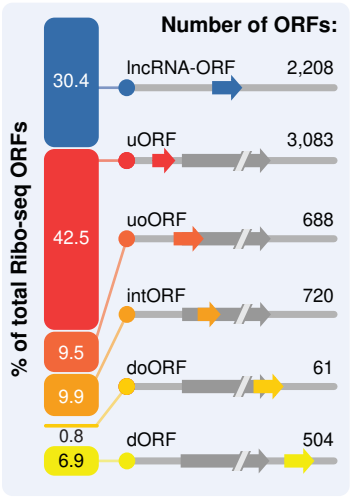
418

Figure 1. Characterization of a consensus set of Ribo-seq ORFs for annotation by GENCODE. **(a)** A schematic of the main steps and goals for this consortium effort. **(b)** A map showing the participating institutions included in this effort. **(c)** A schematic overview of employed filtering steps used to create the consensus set of ribosome profiling (Ribo-seq) ORFs. **(d)** A diagrammatic representation of all Ribo-seq ORFs according to ORF type (see **Table 2** for more information).


a
Unannotated human Ribo-seq ORFs
uORF
uoORF
intORF
doORF
dORF
lncRNA-ORF
Consolidation & categorization
Worldwide collaboration
Reference database support
GENCODE
UniProt
HGNC
PeptideAtlas
End user accessibility


b
Université de Sherbooke
Québec, Canada
Broad Institute, M.I.T.,
Dana-Farber Cancer Institute,
Whitehead Institute
Massachusetts, USA
University of Leeds
United Kingdom
Institute for Systems Biology
Washington, USA
EMBL-EBI
United Kingdom
Yale University
Connecticut, USA
Max Delbrück Center
Germany
University of
California, Berkeley
California, USA
Stowers Institute
Missouri, USA
Princess Máxima Center
the Netherlands
University College Cork
Ireland
Salk Institute
California, USA
Ghent University
Belgium
Human Technopole
Italy
University of Pittsburgh
Pennsylvania, USA
Universitat Pompeu Fabra
Spain
University of Texas, Southwestern
Texas, USA
CSIC-UPO
Spain


c
8,805
Ribo-seq ORFs
from 7 datasets
mapped to
Ensembl v.101
Remove ORFs
overlapping
pseudogenes
& in-frame CDS
Conflate ORFs with
≥ 90% overlap and
same start and/or stop
Select longest ORFs
Consensus set:
7,264
Ribo-seq ORFs
Removed
ORFs
< 16 aa
Removed
unmapped
ORFs
Removed
non-AUG
ORFs


d
Number of ORFs:
% of total Ribo-seq ORFs
30.4
lncRNA-ORF
2,208
42.5
uORF
3,083
uoORF
688
intORF
720
9.5
doORF
61
9.9
0.8
dORF
504
6.9