nature portfolio

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

Supplementary information

Standardized annotation of translated open reading frames

In the format provided by the authors and unedited

Supplementary Information

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)

Supplementary Methods

Phase I ORF retrieval and mapping

We selected seven ORF datasets from different human studies that represent key projects for genome-wide Ribo-seq ORF identification in the last five years (**Supplementary Table 1**). Literature sources were selected based on the comprehensiveness of the dataset, specific focus on large-scale ORF detection, and transparency in reporting multiple categories of ORFs in the datafiles. Thus, additional published human Ribo-seq datasets that do not focus on ORF detection have not been analysed for **Phase I**. Also, we only collate ORFs from studies that used their own experimental and computational workflows for ORF detection, and computational studies targeting Ribo-seq datasets already used by others for detection were excluded to avoid redundancy.

We retrieved ORF exonic coordinates - when available - and ORF sequences, collecting a total set of 39,788 translated ORFs corresponding to 29,373 unique protein sequences. We only selected Ribo-seq ORFs found in long non-coding RNAs (lncRNAs), alternative protein-coding frames and/or UTRs from protein-coding genes. Ribo-seq can also describe translations within pseudogenes[1–3], i.e. loci believed to be defunct protein-coding genes or derived from protein-coding genes but disabled by deleterious mutations. However, pseudogene translations - 'pseudo-ORFs' - are not considered here due to potential complexities in mapping data at loci that can have highly similar genome paralogs. Ribo-seq reads can also be mapped onto circular RNAs (circRNAs), suggesting cap-independent translation[4–7]. However, reference annotation projects do not yet incorporate circRNAs, and the experimental evidence for the translation of 'circORFs' (or 'cORFs') remains a topic of ongoing debate[8,9]. Finally, Ribo-seq can also identify alternative isoforms of annotated proteins, including novel coding exons, in-frame N-terminal extensions (e.g. as recently reported for *STARD10* and *ZNF281*[10]), and internal translation initiation sites that produce shorter proteoforms. The latter will be of particular value to annotation projects as such

isoforms are difficult to find through conservation studies.

Each of the selected studies applied different minimum length cut-offs to define their Ribo-seq ORFs and only 4 of the studies considered near-cognate ORFs (ORFs starting with non-AUG initiation codons, see **Supplementary Table 1**). Hence, in order to maximize ORF replicability across studies we discarded 8,503 Ribo-seq ORFs shorter than 16 amino acids and 10,412 Ribo-seq ORFs starting with near-cognate codons. Next, for the five ORF datasets that were built on an older Human Genome assembly (GRCh37/hg19, **Supplementary Table 1**), we converted ORF coordinates to GRCh38/hg38 using UCSC Liftover[11]. We remapped all translated ORF sequences to Ensembl Release v.101 transcriptome (August 2020, equivalent to GENCODE v35), generating a set of 8,805 unique ORFs, after excluding 1,767 ORFs that could not be fully mapped to any transcript in this release. We note that the latter set includes 80 replicated Ribo-seq ORFs, i.e., ORFs detected in more than one study that could not be matched with a GENCODE V35 transcript model (**Supplementary Table 7**); GENCODE are currently examining these Ribo-seq ORFs and the potential transcript structures they would map to for potential inclusion. For 130 ORFs, the sequence could be mapped to more than one transcribed genomic region and the exact unique region was identified by combining the exonic coordinate data and/or the associated gene names annotated in each study. We next excluded ORFs overlapping pseudogenes (n = 423) or in-frame complete coding sequences (CDS) from protein-coding or nonsense-mediated decay transcripts (n = 560) in the current transcriptome version, since the ORF datasets used older transcriptome releases and new protein-coding sequences and pseudogenes are newly annotated in GENCODE V35 (**Supplementary Table 8**). We noticed that our dataset includes a subset of 49 Ribo-seq ORFs that overlap in-frame incomplete CDSs - without annotated start and/or stop codons (**Supplementary Table 9**) - and a subset of 98 Ribo-seq ORFs that can be assigned to annotated protein entries in Uniprot[12] (**Supplementary Tables 2 and 3**). Finally, in order to get a non-redundant list of translated ORFs for the Phase I, we adapted the clustering method of UniRef90 (UniProt[12]) and collapsed overlapping ORFs with alternative start or stop codons

in groups where multiple instances of ORF isoforms shared the same start and/or stop codon and ≥ 90% of the linear amino acid sequence, considering the longest ORF as representative. If an ORF isoform exhibited significant similarity to two or more non-collapsed ORFs, the isoforms were multiply assigned to all possible cases. This resulted in a final **Phase I** consensus set of 7,264 collapsed Ribo-seq ORFs, where only 549 of the ORFs had more than one ORF isoform – a total of 558 unique ORF sequence isoforms.

The number of ORFs per each of the included 7 studies that passed our filtering and transcript assignment criteria varied between 846 and 3,062. This substantial difference in detected ORFs is not necessarily a reflection of data quality or depth, but primarily the result of approach-specific filtering presets, or the number of replicate identifications required within each study. Updates to the **Phase I** catalog are possible as more Ribo-seq datasets become available and gene annotations continue to expand. Also, we recognise that the relative paucity of Ribo-seq data across human tissues and cell lines prohibits a biologically comprehensive list of human Ribo-seq ORFs at this time.

Identifying replicated Ribo-seq ORFs for Phase I

To date, most Ribo-seq ORFs have been detected in a limited number of common immortalized cell lines (e.g. HEK293, HeLa, K562). The identification of the same Ribo-seq ORF in independent datasets produced by different research groups offers one approach to nominate high-confidence ORFs. This has significant technical considerations, since these studies employ various protocol variations and computational pipelines. For **Phase I**, we have used the Ribo-seq ORF calls as reported in the original manuscript, and consider reproducibility of Ribo-seq ORFs between datasets to be indicator of robustness of the Ribo-seq signal, i.e., a low chance that a given ORF reflects spurious variations in data processing methods. Hence, we selected a subset of 3,085 replicated Ribo-seq ORFs that are found to be translated in more than one study. A Ribo-seq ORF was considered as translated in a specific dataset if the main ORF sequence or any of the collapsed ORF variants were found

in the ORF list generated by that study. Lastly, we recognize that many robust Ribo-seq ORFs may be identified in only one dataset at this time, particularly as the datasets that we have used in this **Phase I** effort have diverse cell types represented -- such as human heart -- which may have numerous lineage-specific ORFs that would not be identified in other datasets. Therefore, while ORF replicability suggests high-confidence, Ribo-seq ORFs identified in only one dataset should not be viewed with unwarranted skepticism.

ORF classification and transcript assignment

Ribo-seq ORFs were classified into 6 different biotypes defined by the host transcript biotype and the relative position of the ORF compared to annotated canonical protein-coding sequences (see **Table 2**). However, gene annotations usually contain several overlapping isoforms and 65.44% of the Ribo-seq ORFs could not be unambiguously mapped to a unique host isoform. For these cases, we assigned the transcript with the highest APPRIS score[13] as the most likely isoform that translates each ORF. If more than one transcript shared a similar APPRIS score, we further evaluated the Ensembl transcript support evidence (TSL) score. For 1,513 ORFs (20.82%), the sequence could still be mapped to more than a single transcript sequence with equal support evidence. However, for these cases the selection of different isoforms did not affect the assigned ORF biotype or the exonic coordinates, after which we randomly selected the transcript with the lowest Ensembl transcript id as host. All possible Ensembl transcript and gene IDs compatible with each ORF sequence, as well as the IDs of the selected host transcripts, are described in **Supplementary Tables 2 and 3**.

Multiple-species alignment and PhyloCSF

We downloaded previously generated multiple-genome alignments for 120 mammals[14] and we extracted aligned regions for each of the human Ribo-seq ORFs, including stop codons. Only species where the ORF region could be fully aligned were included in each alignment. As a result, 99.5% and 97.23% of the ORFs were aligned to at least one primate species and a non-primate mammalian species, respectively. Next, we assessed the conservation of

replicated Ribo-seq ORFs by comparing the patterns of codon evolution in different mammalian species using PhyloCSF[15] with default parameters. For comparison, we additionally built multiple alignments and ran PhyloCSF for a set of 531 annotated CDS sequences shorter than 100 amino acids, taking the longest CDS per gene (sCDSs).

Hydrophobicity and amino acid composition

We estimated average hydrophobicity indices using the Kyte-Doolittle scale for each putative translated amino acid sequence encoded by a Ribo-seq ORF or annotated CDS (aCDS). Ribo-seq ORFs displayed lower hydrophobicity than annotated proteins (mean index of -0.326 for Ribo-seq ORFs vs. -0.364 for aCDSs, **Supplementary Figure 3h**). We next calculated the proportion per type of amino acid in Ribo-seq ORFs and aCDSs. Compared to known proteins, Ribo-seq ORFs contain fewer negatively charged amino acids (D and E) and are enriched for the positively charged amino acid R (**Supplementary Figure 3i**). We also counted how many Ribo-seq ORFs contain at least 2 cysteines, since those amino acids could potentially form disulfide bonds to stabilize protein folds. Of the 3,085 replicated ORFs, 35% of the Ribo-seq ORFs contain at least one pair of cysteines (**Supplementary Figure 3j**).

Analysis of proteomics datasets

We searched for additional evidence of Ribo-seq ORF protein production by collecting 16 published datasets that identified peptides mapping to non-annotated protein-coding regions using different targeted and global mass-spectrometry (MS) approaches (e.g. LC-MS/MS, peptidomics, HLA immunopeptidomics, and selected reaction monitoring (SRM), **Supplementary Table 10**). Peptides were retrieved from the corresponding Supplementary Materials and were remapped to the full set of Ribo-seq ORF sequences. Peptide spectrum matches that uniquely mapped to a Ribo-seq ORF and did not map to any annotated protein-coding sequence were retained as potential MS evidence. We emphasise that these publications have performed MS according to different methodologies, parameters and stringencies, and that we have not - beyond remapping - attempted to reanalyse raw MS data

or standardize the data search parameters. Confidence in these MS reportings should therefore be interpreted in the context of the results as presented in the source publications.

Ongoing searching of MS datasets for Ribo-seq ORFs

We have also taken the first steps towards a standardised and systematic survey for MS evidence in large spectral datasets, emphasizing that the development of this pipeline is a work in progress. In order to confirm *in vivo* translation of Ribo-seq ORFs into proteins, many MS datasets must be searched with these sequences present in the analysis search space. PeptideAtlas[16,17] reprocesses publicly released MS datasets from ProteomeXchange[18] repositories with a large reference sequence search database, including sequences for which potential translation evidence is sought. A set of such sequences from sORFs.org[19] was added in 2017, but little evidence of translation was found at that time.

It is important to consider that, when searching a large number of datasets with speculative sequences, there will be hits to them even if none are correct. Careful error control and manual spectrum match inspection is crucial in order to exclude the possibility that the peptide-spectrum match (PSM) is false; mismatching can be driven by canonical protein sequences with unknown post-translational modifications or uncatalogued DNA variants that each can affect mass-to-charge ratios. The Human Proteome Project[20,21] (HPP) has developed a set of guidelines[22] for claiming novel detection evidence, including strict false discovery rate (FDR) control, the requirement for multiple pieces of evidence, and careful scrutiny of the spectra for such alternative explanations. These credible spectra can be assigned a unique spectrum identifier[23] to provide a reference and audit trail of the identification.

The proposed consensus set of 7,264 Ribo-seq ORFs has been added to the current PeptideAtlas search space, and thus potential translation evidence will be available in future builds of PeptideAtlas after a substantial number of datasets have been reprocessed with these sequences. Some preliminary examination of results was readily available to us since

the sequences of 333 Ribo-seq ORFs were found to be contained in sequences that were already in the PeptideAtlas search space via other sources such as UniProtKB/TrEMBL and sORFs.org[19]. **Supplementary Table 11** contains 13 Ribo-seq ORFs found to have MS support that passes PeptideAtlas' criteria for inclusion in its human all-sample build[20], although additional scrutiny is needed to exclude false positives and the evidence does not meet HPP guidelines.

Out of these 13 hits, 10 Ribo-seq ORFs were supported by single peptides, In six cases the peptides were uniquely mapping, i.e. peptides that could not be also mapped to a primary UniProtKB/Swiss-Prot entry. In four cases, the peptide does map to other proteins, but is listed in the table for completeness. Single peptide evidence is not at present supported as canonical protein sequences by HUPO guidelines, which require more than one non-overlapping PSM in support. GENCODE have also previously utilised the requirement for two supporting PSMs in a survey to identify missing proteins by shotgun proteomics[24]. As such, these Ribo-seq ORFs have not been annotated as proteins by GENCODE at this point, and they will also not be coding sequences in UniProtKB or HGNC. However, we still consider that it may be useful to track and report these findings. For example, in the future, single mapping peptides could form part of a wider body of evidence to support a given translation as a protein, while knowledge of a potential supporting peptide can also inform the design of targeted proteomics methodologies or monoclonal antibody-based detection. We also recognise that other groups and projects do not use the same criteria for MS interpretation as HUPO, and may judge these PSMs along different lines. It is remarkable that, of the six ORFs with single uniquely-mapping peptide evidence, all are detected in HLA peptidome datasets using putative HLA peptides in the search space. It would seem that either there is a propensity for these ORFs to be detected in HLA peptidome samples, or a few HLA peptides map to these ORFs by chance.

These considerations are illustrated by c12riboseqorf48, one of the 10 Ribo-seq ORFs with a single high-quality PSM (**Supplementary Figure 4a**). This Ribo-seq ORF is a uoORF that

overlaps the annotated CDS of limb development membrane protein 1 like (*LMBR1L*). The Ribo-seq ORF is 45aa in size, of which only the first 4 codons do not overlap with the *LMBR1L* protein in an alternative reading frame. The Ribo-seq ORF translation does not show appreciable PhyloCSF support, which would have been indicative of protein-coding constraint with the potential to lead to protein-coding annotation, although we do not currently have clear expectations as to how PhyloCSF should perform in such 'dual-coding' scenarios. The ORF displays strong conservation at the DNA level, however, with the ATG and termination codon being found in almost all placental mammal genomes and without disruption to the frame; conservation potentially extends to bird / reptile genomes. Nonetheless, as discussed in the main article, it is known that uORFs with regulatory functions can also be highly conserved, and so we do not consider this pattern of sequence evolution to validate the Ribo-seq ORF as protein-coding *prima facie*. Experimental evidence will be required in this case. The single supporting peptide is 9aa and semi-tryptic, spans the single and dual coding frame portions of the Ribo-seq ORF, and manual analysis has found the spectra to be of high quality. All three PSMs come from an HLA peptidome dataset[25] and the peptide was identified since it was a predicted HLA peptide included in the search space for this dataset. PeptideAtlas analysis of the Ribo-seq ORF sequence predicts that there are only two tryptic peptides of suitable length with a reasonable chance of discovery, i.e. according to a theoretical trypsin digest. The detected peptide (with 1 Arg and 2 Lys) does not overlap with these suitable tryptic peptides, but instead is in a region with many Arg and Lys that would typically yield peptides that are too short for confident detection with mass spectrometry (although missed cleavages can sometimes overcome this). This peptide could become subject to future synthetic peptide designs for targeted detection. Considering all of these aspects, we believe there is a possibility that this Ribo-seq ORF is being translated into a conserved protein, but protein-coding annotation is currently being held back in lieu of the appearance of additional evidence. The ORF remains under discussion.

Supplementary Table 11 also includes three Ribo-seq ORFs supported by multiple PSMs,

although in two cases the supporting peptides were initially called as multimapping. These two potential protein sequences were detected on account of UniProtKB / TrEMBL sequences A0A024RCX2 and B4DDC5 already being in the PeptideAtlas search space, and these were found to overlap with Ribo-seq ORFs c6norep92 and c10norep59, respectively. In both cases, we found that the multi-mapping peptides - which presented generally high-quality spectra upon manual analysis - in fact represented mapping to redundant sequences in the search space, and we did not find alternative explanations for the PSMs in either case. These peptides could thus be reappraised as single-mapping.

Ribo-seq ORF c6norep92 was called by our analysis as a dORF, found within the 3' UTR of canonical protein-coding gene *PRRT1* (**Supplementary Figure 4b**). This Ribo-seq ORF was found to overlap the C-terminus of the A0A024RCX2 upon manual analysis of the MS data, an apparent isoform of *PRRT1* recognised by UniProtKB but not GENCODE (or RefSeq). However, this protein entry was found to be unusual in not beginning with a canonical ATG initiation codon. Detailed manual gene annotation of the locus was required to resolve this situation. We found that the gene translates into two alternative reading frames, and the C-terminus of A0A024RCX2 that is missing from GENCODE is a legitimate, conserved coding sequence with a RefSeq counterpart in mouse (NM_001368729.1), supported also by the MS data initially identified here. Its absence from the GENCODE annotation meant that our pipeline called the Ribo-seq ORF as a separate entity, and we can now see that in reality this is not a true dORF rather an ORF fragment of a larger protein. We also found no evidence that the non-ATG initiation used by the UniProtKB entry is a genuine biological feature, and that the 'true' isoform corresponding to the alternative C-terminus instead utilises the deeply conserved ATG of the GENCODE model ENST00000211413. The alternative reading frame is instead accessed by the usage of an alternative splice acceptor site at the second coding exon. While this unusual scenario did not lead to the validation of a Ribo-seq ORF as an *independent* protein, it instead illustrates a highly important point about our workflow: moving forward, manual gene annotation will be vital in order to describe and understand Ribo-seq

ORFs with confidence. In particular, we emphasise that our Ribo-seq ORFs are built onto a gene annotation that is not finalised, and that our project allows for transcript and ORF annotations to be improved in a reciprocal manner. Novel protein annotations that can be derived from the Ribo-seq list will not be added to GENCODE without full manual analysis according to the existing annotation guidelines for this project.

Ribo-seq ORF c10norep59 was called as a lncRNA ORF on GENCODE gene LINC00839 (ENSG00000185904), and was found to correspond precisely to UniProtKB Q8NAU0 (**Supplementary Figure 4c**). This protein entry was created in 2002 during the earliest phases of genome annotation, and would not have been annotated by GENCODE as it has poor evolutionary conservation and previously lacked experimental support for its existence. Nonetheless, this ORF is found to have extensive MS support by the PeptideAtlas pipeline, and on this basis we judge that this protein is likely to exist *in vivo*. A consideration, however, is that all of the MS evidence comes from an experiment in U2OS cells, which was originally derived from an osteosarcoma sample[26]. This dataset has been enriched for SUMOylation, and it is possible that this low-abundance protein is only detectable with the aid of SUMOylation enrichment. As a result, this Ribo-seq ORF is currently being held back from protein-coding annotation in GENCODE.

Finally, we note that while these two translations have a relative abundance of distinct PSMs compared to the others in the supplementary file, they are also substantially longer (c6norep92 is 144aa; c10norep59 is 188aa) and feature more trypsin sites for potential digestion.

All sequence and spectral evidence are publicly accessible at the PeptideAtlas web site. All ORF entries discussed here are prefixed with “CONTRIB_GENCODE_” (UniProtKB identifiers do not have this prefix) at the PeptideAtlas web site, and thus a direct link to viewing the results of c10norep59 is found at the URL https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/GetProtein?atlas_build_id=502&appl

y_action=QUERY&protein_name=CONTRIB_GENCODE_c10norep59

References

1. Sisu, C. *et al.* Transcriptional activity and strain-specific history of mouse pseudogenes. *Nat. Commun.* **11**, 3695 (2020).
2. Pei, B. *et al.* The GENCODE pseudogene resource. *Genome Biol.* **13**, R51 (2012).
3. Sisu, C. *et al.* Comparative analysis of pseudogenes across three phyla. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 13361–13366 (2014).
4. Pamudurti, N. R. *et al.* Translation of CircRNAs. *Mol. Cell* **66**, 9–21.e7 (2017).
5. van Heesch, S. *et al.* The Translational Landscape of the Human Heart. *Cell* **178**, 242–260.e29 (2019).
6. Legnini, I. *et al.* Circ-ZNF609 Is a Circular RNA that Can Be Translated and Functions in Myogenesis. *Mol. Cell* **66**, 22–37.e9 (2017).
7. Yang, Y. *et al.* Extensive translation of circular RNAs driven by N6-methyladenosine. *Cell Res.* **27**, 626–641 (2017).
8. Ho-Xuan, H. *et al.* Comprehensive analysis of translation from overexpressed circular RNAs reveals pervasive translation from linear transcripts. *Nucleic Acids Res.* **48**, 10368–10382 (2020).
9. Hansen, T. B. Signal and noise in circRNA translation. *Cold Spring Harbor Laboratory* 2020.12.10.418848 (2020) doi:10.1101/2020.12.10.418848.
10. Na, C. H. *et al.* Discovery of noncanonical translation initiation sites through mass spectrometric analysis of protein N termini. *Genome Research* vol. 28 25–36 (2018).
11. Navarro Gonzalez, J. *et al.* The UCSC Genome Browser database: 2021 update. *Nucleic Acids Research* vol. 49 D1046–D1057 (2021).
12. The UniProt Consortium. UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
13. Rodriguez, J. M. *et al.* APPRIS 2017: principal isoforms for multiple gene sets. *Nucleic*

Acids Res. **46**, D213–D217 (2018).

14. Hecker, N. & Hiller, M. A genome alignment of 120 mammals highlights ultraconserved element variability and placenta-associated enhancers. *Gigascience* **9**, (2020).
15. 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).
16. Desiere, F. *et al.* Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.* **6**, R9 (2005).
17. van Wijk, K. J. *et al.* The Arabidopsis PeptideAtlas: Harnessing worldwide proteomics data to create a comprehensive community proteomics resource. *Plant Cell* **33**, 3421–3453 (2021).
18. Deutsch, E. W. *et al.* The ProteomeXchange consortium in 2020: enabling 'big data' approaches in proteomics. *Nucleic Acids Research* (2019) doi:10.1093/nar/gkz984.
19. 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).
20. Omenn, G. S. *et al.* Progress Identifying and Analyzing the Human Proteome: 2021 Metrics from the HUPO Human Proteome Project. *J. Proteome Res.* **20**, 5227–5240 (2021).
21. Adhikari, S. *et al.* A high-stringency blueprint of the human proteome. *Nat. Commun.* **11**, 5301 (2020).
22. Deutsch, E. W. *et al.* Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 3.0. *Journal of Proteome Research* vol. 18 4108–4116 (2019).
23. Deutsch, E. W. *et al.* Universal Spectrum Identifier for mass spectra. *Nat. Methods* **18**, 768–770 (2021).
24. Wright, J. C. *et al.* Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow. *Nature Communications* vol. 7 (2016).
25. Sarkizova, S. *et al.* A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat. Biotechnol.* **38**, 199–209 (2020).

26. Hendriks, I. A. *et al.* Site-specific mapping of the human SUMO proteome reveals co-modification with phosphorylation. *Nat. Struct. Mol. Biol.* **24**, 325–336 (2017).

Data and code availability

All analyses in this study are performed using published and publicly available analytical tools or software packages. Published Ribo-seq ORF datasets and processed mass-spectrometry peptide datasets were retrieved from the Supplementary Material of each referenced study as described in **Supplementary Tables 1 and 10**. The code used for generating the list of Phase I ORFs is available at https://github.com/jorruior/gencode-riboseqORFs.

Supplementary Tables

Supplementary Table 1. Description of the 7 human Ribo-seq datasets used for this study

Supplementary Table 2. Table with 3,085 replicated Ribo-seq ORFs that were found in at least two Ribo-seq studies. LncRNA-ORFs were divided into two categories (see ORF biotype): lncRNA (lncRNA-ORFs in lncRNA genes) and processed_transcript (lncRNA-ORFs in protein-coding genes)

Supplementary Table 3. Table with 4,179 Ribo-seq ORFs that are from a specific Ribo-seq study. LncRNA-ORFs were divided into two categories (see ORF biotype): lncRNA (lncRNA-ORFs in lncRNA genes) and processed_transcript (lncRNA-ORFs in protein-coding genes)

Supplementary Table 4. Table with 254 Ribo-seq ORF sequences discarded due to the presence of near-cognate initiation codons. These ORFs were replicated in at least two Ribo-seq studies.

Supplementary Table 5. Table with 1,520 Ribo-seq ORF sequences discarded due to the short size (< 16 amino acids). These ORFs were replicated in at least two Ribo-seq studies.

Supplementary Table 6. This sheet lists 10 ORFs that have been annotated as protein-coding by GENCODE as part of this work, and a further 15 that were previously annotated as part of preliminary investigative work into Ribo-seq datasets combined with in-house PhyloCSF analysis. Proteins that are listed as appearing in GENCODE v38-39 are not in a public ‘genebuild’ release at the time of publication, and so gene and transcript IDs are not yet available for these models. Note that while the ‘comments’ provide brief explanations for the annotation, they do not attempt to establish provenance for the initial identification of the ORF or protein. Further support for these annotations could potentially be found in additional resources or publications. These annotation decisions were made by GENCODE according to an interpretation of the balance of probability when considering all available evidence, i.e. these ORFs are considered *most likely* to be protein-coding, in line with standard annotation

criteria. GENCODE recognise that further experimental characterization will be required to support these annotations.

Supplementary Table 7. Table with 80 Ribo-seq ORF sequences that could not be mapped to any of the current transcript annotations in Ensembl v.101. These ORFs were replicated in at least two Ribo-seq studies.

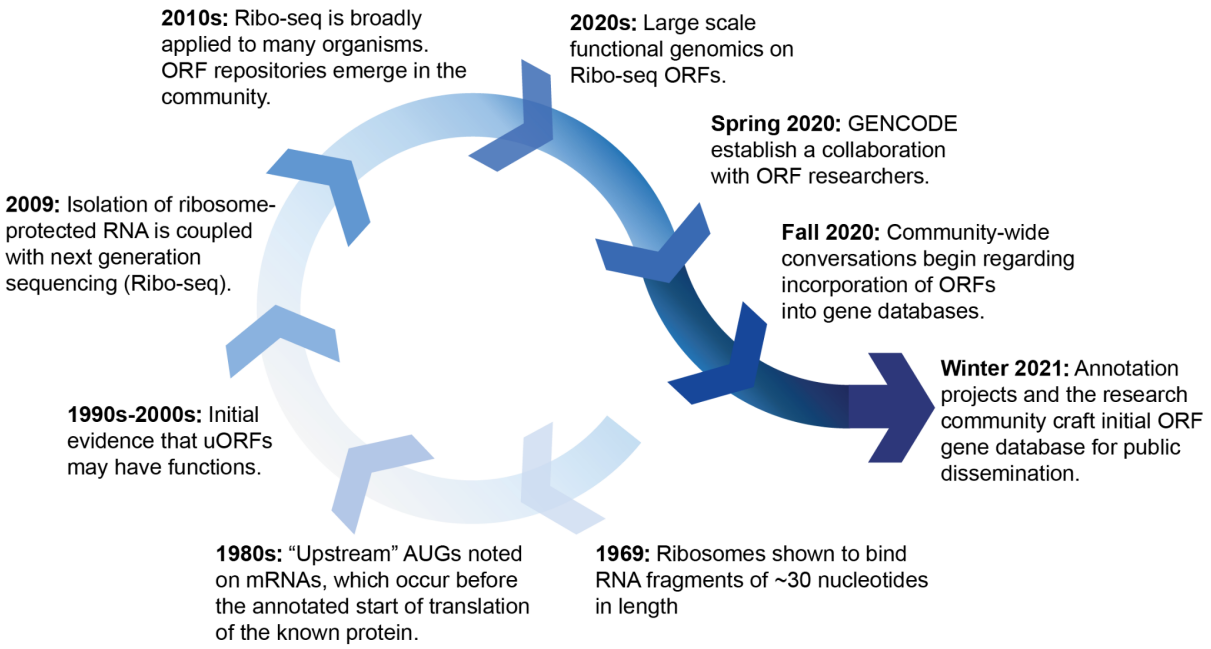
Supplementary Table 8. Table with 957 Ribo-seq ORFs currently annotated as protein-coding (CDS or NMD) or overlapping pseudogenes in Ensembl v.101. ORFs partially or totally overlapping in-frame CDS were included in this table. If two or more ORFs shared >= 90% of the amino acid sequence, only the longest one was included.

Supplementary Table 9. Table with 49 Ribo-seq ORFs that overlap in-frame annotated CDS without annotated start and/or stop codons.

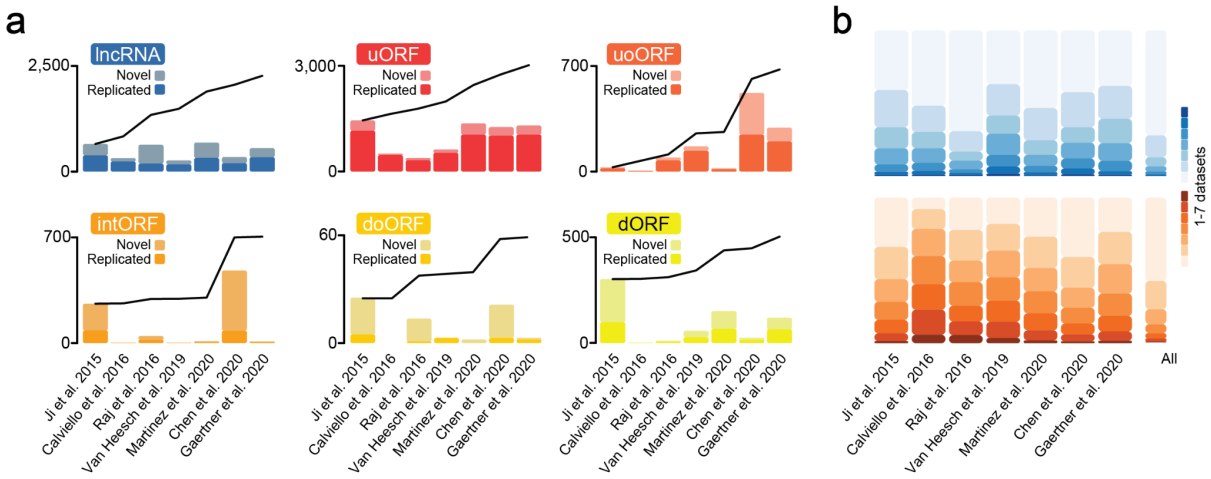
Supplementary Table 10. Description of the 16 mass-spectrometry datasets used for this study. Identified peptides were retrieved from supplementary data and re-mapped to Ribo-seq ORFs.

Supplementary Table 11. Table of Ribo-ORFs which have had their protein-coding potential manually assessed due to their presentation of peptide evidence following analysis by PeptideAtlas. These ORFs have not been annotated as protein-coding by GENOCDE, UniProt or HGNC at the present time, nor recognised as protein-coding by HUPO / HPP. All except two entries have only a single peptide. Most spectra are ‘good’ according to manual inspection by experts in PeptideAtlas, but various confounding factors call into question the detection of these ORFs. Exceptions are c6norep92 and c10norep59, which are supported by multiple peptides. The former is explicable as a previously unrecognised alternative protein isoform, while we consider that protein-coding annotation of the latter will depend on the prior development of guidelines for non-standard experimental datasets, as discussed above.

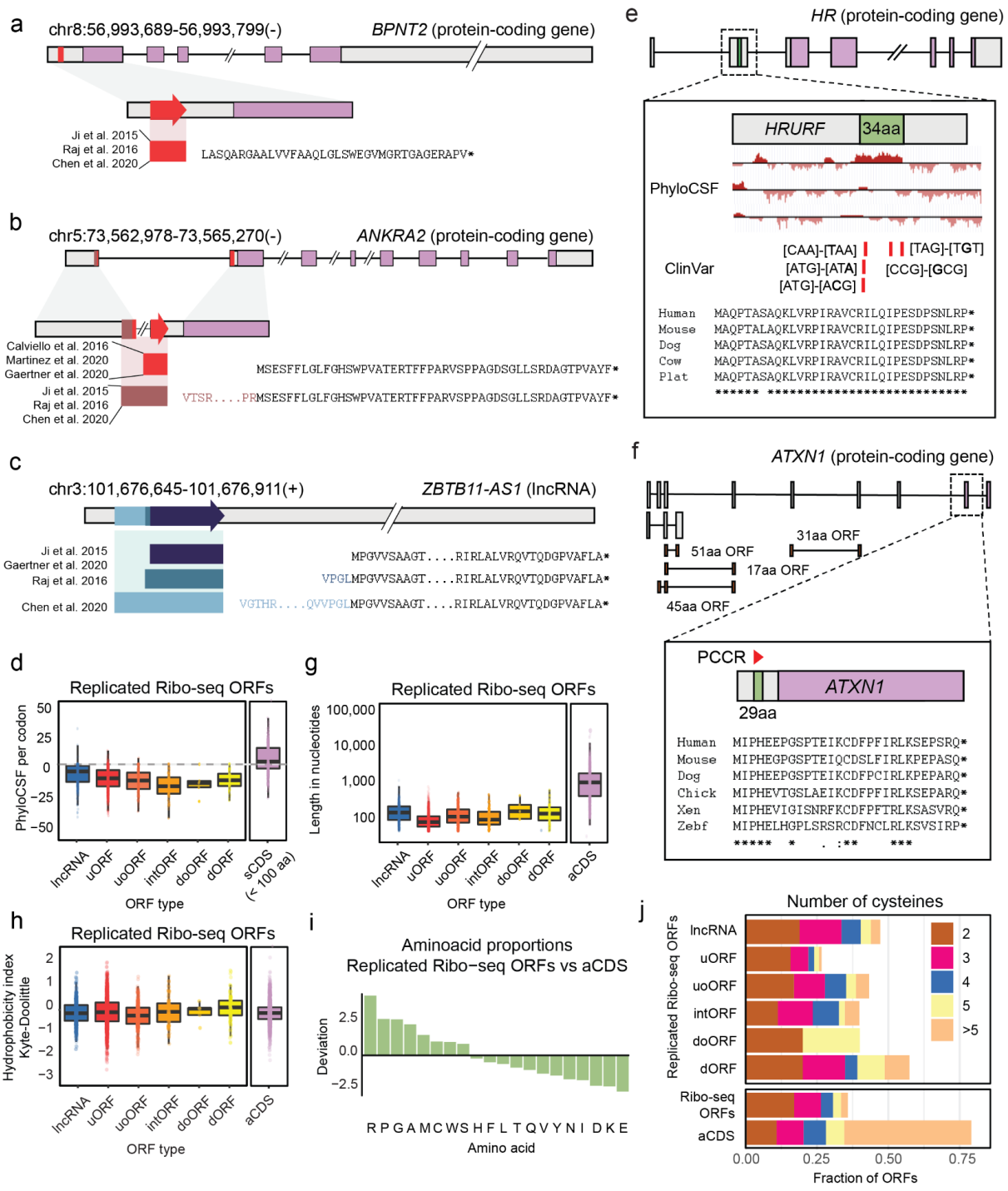
Supplementary Figures



Supplementary Figure 1: A timeline showing the formation of this community consensus resource for Ribo-seq ORFs in relation to major scientific advances in understanding these ORFs.



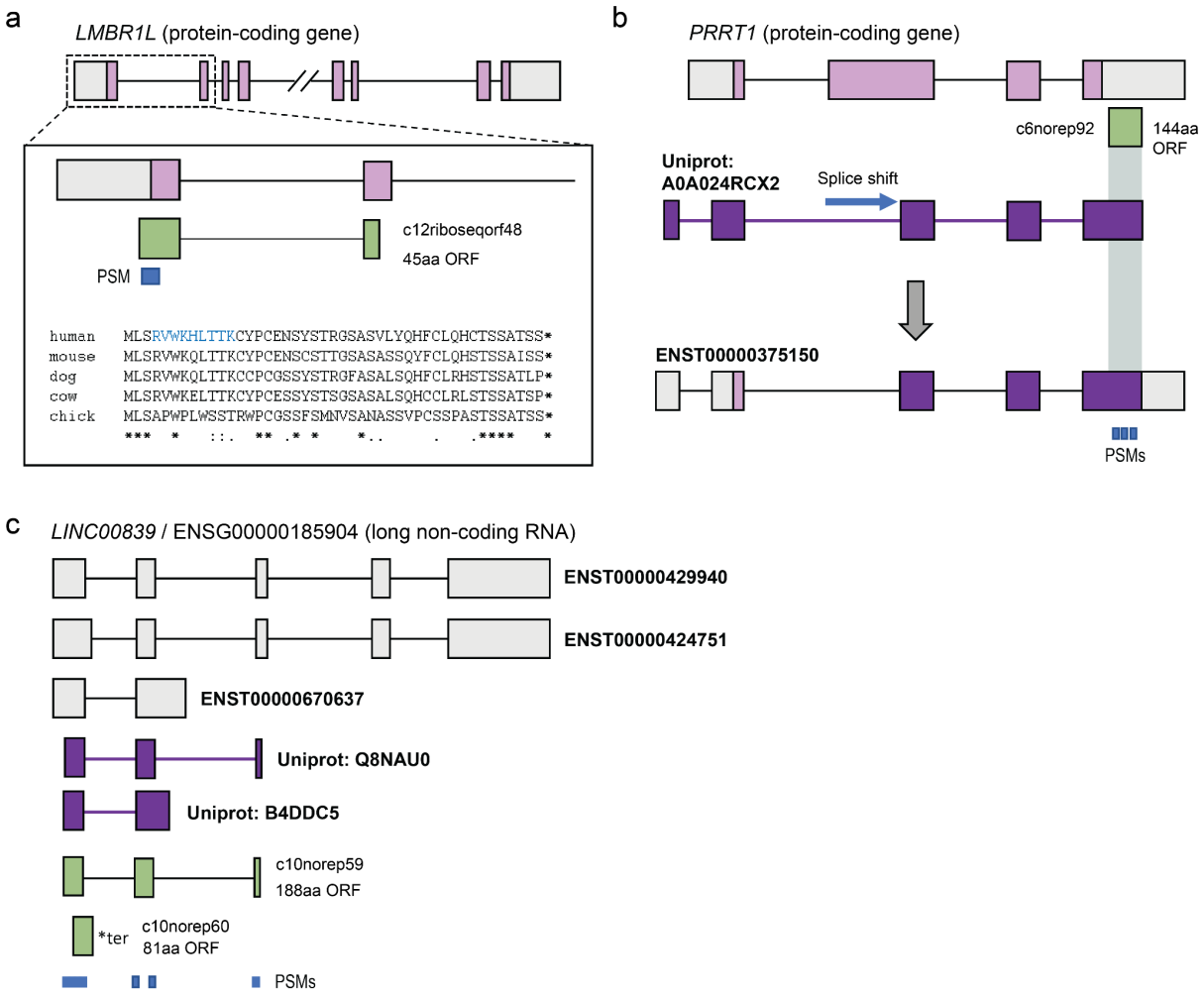
Supplementary Figure 2: (a) Bar plots with abundances of replicated and newly identified ORFs across datasets. Cumulative lines illustrate the evolution in the total number of new unique ORFs identified across studies, sorted chronologically. **(b)** Replicated Ribo-seq ORF identifications within lncRNAs (top) and mRNAs (bottom), organized by study.



Supplementary Figure 3: (a) A uORF located in the *BPNT2* gene has multiple datasets supporting translation at a near-cognate initiation codon. **(b)** A uORF located in the *ANKRA2* gene has a near-cognate initiation codon but also a separate annotation utilizing a methionine initiation codon. **(c)** A Ribo-seq ORF located in the *ZBTB11-AS1* lncRNA has several proposed initiation codons utilizing either a methionine initiation codon or a near-cognate

initiation codon. **(d)** Box plots with nucleotide sequence lengths of replicated Ribo-seq ORFs compared to annotated CDSs (aCDS). ORFs are separated into each respective class. Ribo-seq ORFs are significantly shorter than annotated CDS (two-sided Wilcoxon test, p-value < 10^{-10}) **(e)** A 34 amino acid uORF (green box) identified within the 5' UTR of *HR* (UTR sequences in grey boxes; CDS in purple boxes) has been annotated as protein-coding (ENSG00000288677), now recognised as *HRURF* by HGNC. The protein-coding nature of the ORF was inferred by PhyloCSF, according to the positive signal in the top reading frame. Further support was provided by in depth comparative annotation of other vertebrate genomes, demonstrating that the protein likely evolved at the base of the therian mammal radiation; an illustrative alignment is included ('Plat' standing for platypus). *HR* has an ortholog in avians and reptiles; the equivalent sequence in these genomes lacks coding potential (not shown), indicating that *HRURF* evolved *de novo*. Five ClinVar variants fall within *HRURF*: RCV000007766.4, RCV001030440.1, RCV000007767.4, RCV000007768.4 and RCV000007769.3 in 5' order. Each is classed as 'Pathogenic', although non-coding. Following the new CDS annotation, mutations RCV000007766.4 and RCV001030440.1 are seen to disrupt the initiation codon, RCV000007767.4 and RCV000007768.4 are missense mutations, while RCV000007769.3 disrupts the termination codon. **(f)** A 29 amino acid uORF within the complex 5' UTR of *ATXN1* has been annotated as protein-coding, and will appear in a future GENCODE release. This translation has been evolving as coding sequence across the vertebrate radiation ('Chick' is chicken, 'Xen' is Xenopus, 'Zebf' is zebrafish), and the strong PhyloCSF signal (not shown) produced a PhyloCSF Candidate Coding Region (red triangle), indicative of a non-annotated CDS. The canonical transcript of *ATXN1* (ENST00000244769, top model) has six additional 5' UTR exons with three uORFs inferred from the Ribo-seq datafile (the 17 and 45 amino acid ORFs are overlapping in different reading frames), while a final ORF has been mapped to an alternatively spliced non-coding transcript (ENST00000467008, second model). While these various UTR exons are generally conserved and supported by transcript evidence in other mammal genomes, the additional ORFs are not strongly conserved and do not present signatures of purifying selection as CDS.

(g) Box plots with phyloCSF scores of replicated Ribo-seq ORFs assessing amino acid purifying selection for ORFs compared to annotated CDSs less than 100 amino acids in length (short CDS, sCDS). Only 2.4% of the replicated Ribo-seq ORFs displayed positive PhyloCSF scores, in contrast to 48% of the sCDS. **(h)** Box plots with Kyle-Doolittle hydrophobicity indices of replicated Ribo-seq ORFs compared to aCDS. ORFs are separated into each respective class. Ribo-seq ORFs are significantly less hydrophobic than annotated CDS (two-sided Wilcoxon test, p-value < 0.05) **(i)** Bar plots with the relative proportion of amino acids in replicated Ribo-seq ORFs compared to aCDSs. For each amino acid, the difference between the relative proportion in Ribo-seq ORFs and aCDSs was calculated and divided by the relative proportion in aCDSs. **(j)** Bar plots with the fraction of Ribo-seq ORFs and aCDSs that contain 2, 3, 4, 5, and more than 5 cysteines in their amino acid chains.



Supplementary Figure 4: (a) Mass spectrometry data analysis for c12riboseqorf 48. The 45aa ouORF (green) is found within the *LMBR1L* gene, with its initiation codon 4 codons upstream of the initiation codon of the canonical *LMBR1L* CDS (lilac). The Ribo-seq ORF is supported by one 9aa PSM (blue box; blue highlight in the protein alignment), which is single-mapping and fully tryptic. The translation is highly conserved at the DNA level in placental mammals, with support also in reptile / avian genomes for which sequence alignments are available (chick = chicken). Marsupial genomes have a premature termination codon in common (not shown). **(b)** Mass spectrometry data analysis for c6norep92. The 144aa dORF (green) was found in the 3' UTR of *PRRT1* (CDS in lilac; UTR in grey). The dORF was found to be encompassed by UniProt entry A0A024RCX2, at the C-terminal end. This protein entry is translated in a completely different frame (purple) to the canonical *PRRT1* CDS, and is set with a non-canonical initiation codon for which no supporting evidence could be found in this

analysis. Manual gene annotation found that the alternate reading frame is instead highly likely to be accessed by a splice acceptor site shift in coding exon 3 in combination with the canonical initiation codon of *PRRT1*. This is supported by transcriptional evidence (note that the 5' UTR of the gene has additional transcriptional complexity that is not represented here for clarity), and a new model ENST00000375150 has been added to GENCODE and is anticipated to first appear in release GENCODEv40. The Ribo-seq ORF as called by this work is thus now considered to be 5' truncated. PSMs supporting c6norep92 are shown as blue rectangles; additional PSMs supporting the N-t extended form are not shown. **(c)** Mass spectrometry data analysis for c10norep59. This 188aa Ribo-seq ORF is found with lncRNA ENSG00000185904, named by HGNC as *LINC00839*, and was found to correspond precisely to UniProt entry Q8NAU0, which can be aligned to GENCODE transcript ENST00000429940. A second Ribo-seq ORF has been described within the locus, C10norep60, found on alternative GENCODE transcript ENST00000424751. It uses an internal initiation codon in the same reading frame compared to c10norep59, and, due to a splice donor site shift in exon 1 of ENST00000424751, terminates 4bp into the intronic sequence of ENST00000429940. Translation in the locus is supported by a series of PSMs, although the alternative C-terminus of C10norep60 is not distinguished. An additional, distinct UniProt entry B4DDC5 has also been identified within the locus, mapping to alternative GENCODE transcript ENST00000670637. The alternative C-terminus of this translation is supported by a PSM, and so it may represent an alternative protein isoform. However, at present, all peptide data associated with the locus is from a cell-line. As the translations do not display evolutionary conservation, protein-coding annotation in GENCODE has not yet been decided upon pending further discussion.