
Supplementary information

Expanding the human proteome with microproteins and peptideins

In the format provided by the
authors and unedited

Supplementary Information

Expanding the human proteome with microproteins and peptideins

Eric W. Deutsch^{1,*}, Leron W. Kok^{2,3,*}, Jonathan M. Mudge^{4,*}, Cristian F. Valls^{5,6,*}, Irwin Jungreis^{7,8,*}, Jorge Ruiz-Orera⁹, Zhi Sun¹, Ulrike Kusebauch¹, Ivo Fierro-Monti^{4,10}, Jennifer G. Abelin⁸, M. Mar Alba^{11,12}, Julie L. Aspden¹³, Sreejan Bandyopadhyay¹⁴, Kaushik Banerjee^{5,6}, Pavel V. Baranov¹⁵, Ariel A. Bazzini^{16,17}, Francis Bourassa¹⁸, Elspeth A. Bruford¹⁹, Lorenzo Calviello²⁰, Steven A. Carr⁸, Anne-Ruxandra Carvunis^{21,22,23}, Sonia Chothani^{24,25}, Jim Clauwaert⁵, Kellie Dean²⁶, Pouya Faridi^{27,28}, Adam Frankish⁴, Amy Goodale⁸, Thomas Green⁸, Norbert Hubner^{9,29,30,31}, Nicholas T. Ingolia³², Manolis Kellis^{7,8}, Michele Magrane⁴, Maria Jesus Martin⁴, Thomas F. Martinez^{33,34,35}, Gerben Menschaert³⁶, Uwe Ohler^{37,38}, Sandra Orchard⁴, Alisa Potter^{2,3,39}, Owen J.L. Rackham⁴⁰, Matthew G. Rees⁸, David E. Root⁸, Jennifer A. Roth⁸, Xavier Roucou⁴¹, Fernando J. Sialana¹⁴, Sarah A. Slavoff^{42,43,44}, Michał I. Świrski⁴⁵, Jack A.S. Tierney⁴, Félix-Antoine Trifiro¹⁸, Eivind Valen⁴⁶, Valeriia Vasylieva¹⁸, Aaron Wacholder^{21,22,23}, Shengbo Wang⁴, Li Wang⁸, Jonathan S. Weissman^{47,48,49,50}, Wei Wu^{51,52}, Zhi Xie (谢志)⁵³, Jyoti S. Choudhary¹⁴, Michal Bassani-Sternberg^{54,55,56}, Juan Antonio Vizcaíno⁴, Nicola Ternette^{57,58}, Marie A. Brunet^{18,59,60}, Robert L. Moritz^{1,\$}, John R. Prensner^{5,6,\$}, Sebastiaan van Heesch^{2,3,\$}

¹Institute for Systems Biology (ISB), Seattle, WA, 98109, USA

²Princess Máxima Center for Pediatric Oncology, Utrecht, 3584 CS, The Netherlands

³Onco Institute, Utrecht, The Netherlands

⁴European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, CB10 1SD, UK

⁵Department of Pediatrics, Division of Pediatric Hematology/Oncology, University of Michigan Medical School, Ann Arbor, MI, 48109, USA

⁶Department of Biological Chemistry, University of Michigan Medical School, Ann Arbor, MI, 48109, USA

⁷Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA

⁸Broad Institute of MIT and Harvard, Cambridge, MA, 02142, USA

⁹Cardiovascular and Metabolic Sciences, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, 13125, Germany

¹⁰Biozentrum, University of Basel, Basel, 4056, Switzerland

¹¹Hospital del Mar Research Institute, Barcelona, Spain

¹²Catalan Institute for Research and Advanced Studies (ICREA), Barcelona, Spain

¹³School of Molecular and Cellular Biology, Faculty of Biological Sciences, University of Leeds, Leeds, LS2 9JT, UK

¹⁴Functional Proteomics Group, Institute of Cancer Research, Chester Beatty Labs, London, SW3 6JB, UK

¹⁵School of Biochemistry and Cell Biology, University College Cork, Cork, Ireland

¹⁶Stowers Institute for Medical Research, Kansas City, MO, 64110, USA

¹⁷Department of Molecular and Integrative Physiology, University of Kansas Medical Center, Kansas City, KS, 66160, USA

¹⁸Pediatrics Department, University of Sherbrooke, Sherbrooke, Québec, Canada

¹⁹HUGO Gene Nomenclature Committee (HGNC), Department of Haematology, University of Cambridge School of Clinical Medicine, Cambridge, UK

- ²⁰Human Technopole, Milan, 20157, Italy
- ²¹Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, 15213, USA
- ²²Pittsburgh Center for Evolutionary Biology and Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, PA, 15213, USA
- ²³Present address: Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland
- ²⁴Centre for Computational Biology and Program in Cardiovascular and Metabolic Disorders, Duke-NUS (National University of Singapore) Medical School, Singapore, 169857, Singapore
- ²⁵Genome Institute of Singapore (GIS), Agency for Science, Technology and Research (A*STAR), 60 Biopolis Street, Singapore, 138672, Singapore
- ²⁶School of Biochemistry and Cell Biology, University College Cork, Cork, Ireland
- ²⁷Centre for Cancer Research, Hudson Institute of Medical Research, Clayton, VIC, Australia
- ²⁸Monash Proteomics & Metabolomics Platform, Department of Medicine, School of Clinical Sciences, Monash University, Clayton, VIC, Australia
- ²⁹Charité-Universitätsmedizin Berlin, Berlin, 10117, Germany
- ³⁰Helmholtz-Institute for Translational AngioCardioScience (HI-TAC) of the Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC) at Heidelberg University, Heidelberg, 69117, Germany
- ³¹DZHK (German Center for Cardiovascular Research), Partner Site Berlin, Berlin, 13347, Germany
- ³²Department of Molecular and Cell Biology, Center for Computational Biology, University of California, Berkeley, Berkeley, CA, 94720-3202, USA
- ³³Department of Pharmaceutical Sciences, University of California, Irvine, Irvine, CA, 92617, USA
- ³⁴Department of Biological Chemistry, University of California, Irvine, Irvine, CA, 92617, USA
- ³⁵Chao Family Comprehensive Cancer Center, University of California, Irvine, Irvine, CA, 92617, USA
- ³⁶Biobix, Lab of Bioinformatics and Computational Genomics, Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Ghent, Belgium
- ³⁷Department of Biology, Humboldt University Berlin, Berlin, 10117, Germany
- ³⁸Berlin Institute of Medical Systems Biology (BIMSB), Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin, 10115, Germany
- ³⁹Biomolecular Mass Spectrometry and Proteomics, Bijvoet Centre for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands
- ⁴⁰School of Biological Sciences, University of Southampton, Southampton, UK
- ⁴¹Department of Biochemistry and Functional Genomics, Université de Sherbrooke, Sherbrooke, Québec, Canada
- ⁴²Department of Chemistry, Yale University, New Haven, CT, 06520, USA
- ⁴³Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, 06520, USA
- ⁴⁴Institute for Biomolecular Design and Discovery, Yale University, West Haven, CT, 06516, USA
- ⁴⁵Institute of Genetics and Biotechnology, Faculty of Biology, University of Warsaw, Warsaw, Poland
- ⁴⁶Department of Biosciences, University of Oslo, Oslo, Norway
- ⁴⁷Whitehead Institute for Biomedical Research, Cambridge, MA, 02142, USA
- ⁴⁸Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, 02142, USA
- ⁴⁹Howard Hughes Medical Institute, Whitehead Institute, Cambridge, MA, 02142, USA
- ⁵⁰David H. Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA
- ⁵¹Singapore Immunology Network (SIgN), Agency for Science, Technology and Research (A*STAR), Singapore
- ⁵²Department of Pharmacy & Pharmaceutical sciences, National University of Singapore (NUS), Singapore
- ⁵³State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China
- ⁵⁴Ludwig Institute for Cancer Research, University of Lausanne, Lausanne, 1005, Switzerland
- ⁵⁵Department of Oncology, Centre hospitalier universitaire vaudois (CHUV), Lausanne, 1005, Switzerland
- ⁵⁶Agora Cancer Research Centre, Lausanne, 1011, Switzerland
- ⁵⁷School of Life Sciences, Division Cell Signalling and Immunology, University of Dundee, Dundee, DD1 5EH, UK

⁵⁸Centre for Immuno-Oncology, Nuffield Department of Medicine, University of Oxford, Oxford, OX37DQ, UK

⁵⁹Centre de Recherche du Centre hospitalier universitaire de Sherbrooke (CRCHUS), Sherbrooke, Québec, Canada

⁶⁰Cancer Research Institute, University of Sherbrooke (IRCUS), Sherbrooke, Québec, Canada

Table of contents

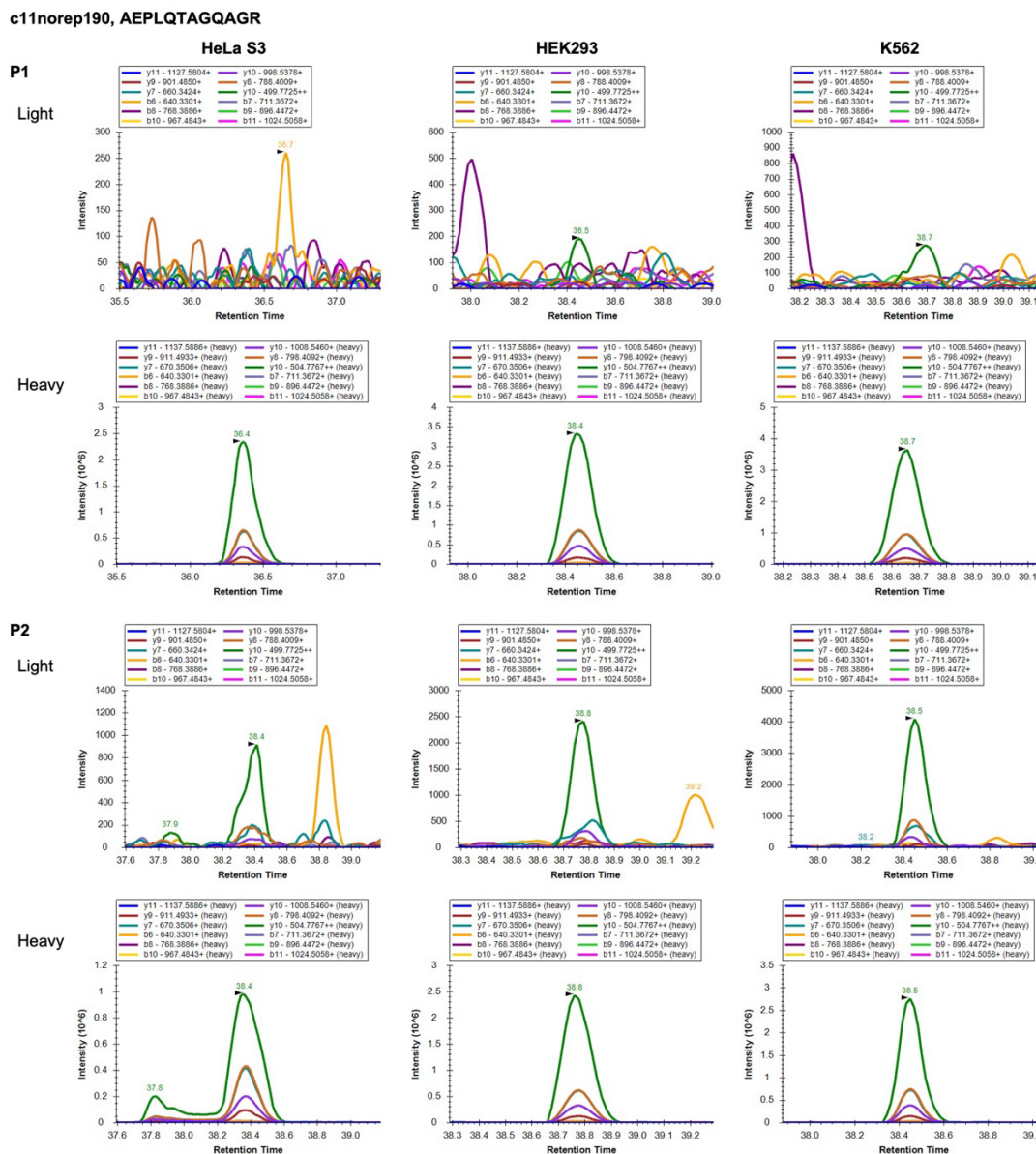
Supplementary information figures

1. Peptide verification of c11norep190 through targeted proteomics (p4)
2. Peptide verification through targeted proteomics utilizing different mass spectrometry instruments (p5)

Supplementary information results

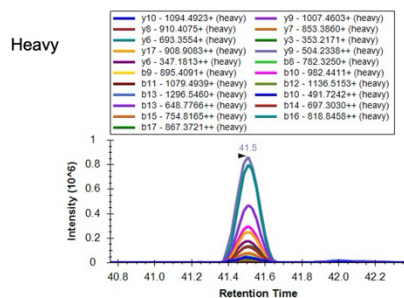
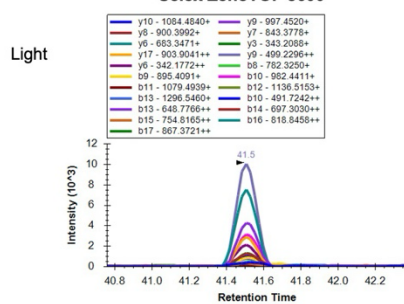
1. Discussion of ncORF detections in HLA and non-HLA data subjected to manual analysis (p6)
2. Machine learning-based prediction of ncORF detectability (p78)
3. Detection of ncORFs in human ubiquitinated proteomics datasets (p91)
4. ORBL: motivation, methodology, validation, and limitations (p93)
5. Structural predictions of ncORF-derived proteins (p105)
6. *De novo* gene annotation (p107)

Supplementary information figures

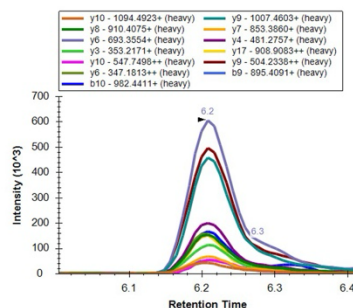
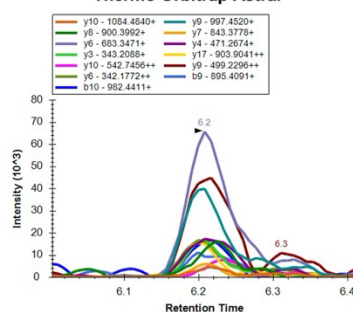


Supplementary Information Figure 1. Peptide verification of c11norep190 through targeted proteomics. Visualization of ion transitions from the endogenous (top and third row) and synthetic heavy labelled (second and bottom row) peptide AEPLQTAGQAGR (c11norep190, Tier 2A) in cell lysates from HeLa S3, HEK 293 and K562 (left to right). The endogenous peptide was detected in all cell lysates from sample processing protocol P2 (bottom two rows), but not in cell lysates from sample processing protocol P1 (top two rows). Data acquired on a Sciex ZenoTOF 8600 instrument. This peptide was predicted and not previously seen in the PeptideAtlas non-HLA build.

a c11riboseqor4, ATPGHTGCLSPGCPDQPAR
Sciex ZenoTOF 8600

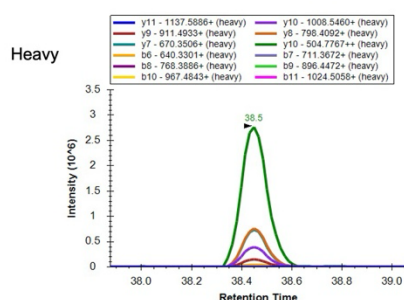
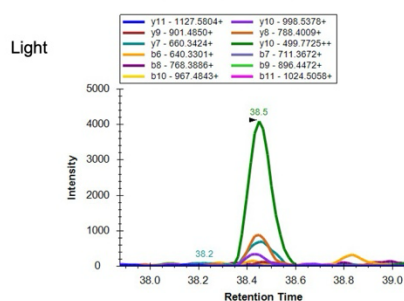


Thermo Orbitrap Astral

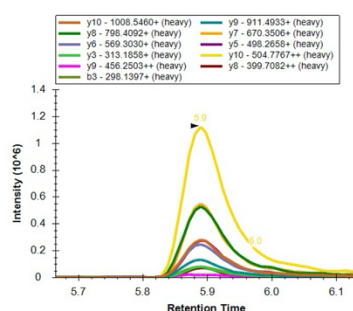
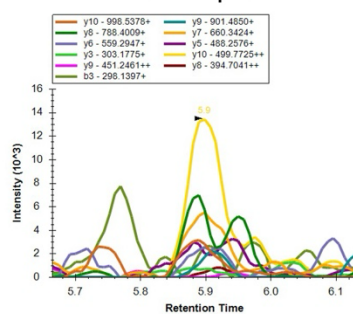


b c11norep190, AEPLQTAGQAGR

Sciex ZenoTOF 8600



Thermo Orbitrap Astral



Supplementary Information Figure 2. Peptide verification through targeted proteomics utilizing different mass spectrometry instruments. Visualization of ion transitions from the endogenous (top rows) and synthetic heavy labelled (bottom rows) peptides ATPGHTGCLSPGCPDQPAR (**A**, c11riboseqor4, Tier 1A) and AEPLQTAGQAGR (**B**, c11norep190, Tier 2A) in K562 cell lysates using sample processing protocol 2 (ACN precipitated). The endogenous peptides were detected with both instruments, the Sciex ZenoTOF 8600 (left) and the Thermo Orbitrap Astral (right).

Supplementary information results

(1) Discussion of ncORF detections in HLA and non-HLA data subjected to manual analysis

This document provides extended discussion on ncORFs initially discovered by Ribo-seq (henceforth ncORFs) and subsequently found to have support in non-HLA or HLA datasets. Specifically, it provides further information on ncORFs that have been subjected to manual analysis, which involves the examination of the supporting peptide data and the deployment of a gene annotation workflow, both being performed by expert curators. All ncORFs in Tiers 1A and 2A as part of their initial classification have been subjected to manual analysis, i.e., all ncORFs with at least one non-HLA peptide in apparent support following the proteomics survey. Similarly, all ncORFs in Tier 2A with at least 5 HLA peptides in support as per the initial proteomics survey have been subjected to manual analysis, and are presented in the same way.

Manual analysis workflow

Proteomics analysis

The first stage of the manual analysis workflow involved an examination of the supporting peptides for ncORFs, especially in terms of their spectral quality. This was performed by PeptideAtlas, in line with standard operating procedures.

Supporting peptides were examined for all 183 ncORFs that presented as Tier 1A (37) or Tier 2A (146) following the initial proteomics survey, and for all 79 ncORFs in Tier 1B supported by at least 5 HLA peptides. Note that 2 and 4 ncORFs with 5+ HLA peptides in support were categorised in Tier 1A and Tier 2A respectively, i.e., these ncORFs are also supported by non-HLA peptides. The rationale for requiring 5 or more HLA peptides is discussed below. **Supplementary Table 2** provides the complete list of non-HLA peptides matched to ncORFs. Each row represents an individual peptide, some of which have multiple peptide-spectrum matches (PSMs). For each peptide we list the peptide identifier, sequence, number of PSMs, and the ncORF entry to which it maps as taken from Mudge *et al*, as well as the ORF's PeptideAtlas category. We also provide the Universal Spectrum Identifier (USI)⁹⁵ of the best PSM for the peptide. All spectra are available in the PeptideAtlas interface for additional public scrutiny. Furthermore, USIs provide a mechanism with which the original spectrum can be viewed or downloaded for any participating ProteomeXchange resource that holds the spectrum, and annotated with the requested sequence. The easiest location to resolve USIs is at the ProteomeXchange ProteomeCentral webpage <https://proteomecentral.proteomexchange.org/usii/>. **Supplementary Table 6** provides the equivalent information for the HLA peptides.

Supplementary Table 3 reconfigures the information on non-HLA peptides by listing each ncORF with matched peptides as a separate row. Certain information fields are duplicated from **Supplementary Table 2**. In addition, **Supplementary Table 3** includes the official symbol of the gene in which the ncORF is mapped and the GRCh38 coordinates of the translated sequence. We also include brief information on our manual reappraisal of the Ribo-seq evidence, which we consider to be an essential phase of the annotation process. Similarly, **Supplementary Table 7** reconfigures the information on HLA peptides in **Supplementary Table 6** based on the ncORF.

Certain peptides are listed as mapping to additional sequences in the proteomics search space, which is substantially broader than our collection of 7,264 Ribo-seq ORFs. In particular, we find mappings to annotations provided by UniProtKB or RefSeq. Having examined these in detail, we find that in almost all cases the alternative annotations are derived from the same locus as our ncORF, i.e., the peptide is (in effect) mapped to the same genomic sequence. Thus, these are not multimappings, i.e., cases where a peptide can genuinely be mapped to more than one genomic location, as would raise doubts into its provenance. Those few exceptions are the putative pseudogene cases discussed below, although we also note that several ncORFs incorporate transposon sequence, and it should be considered that additional sequences similar to these may exist outside our search space. Commentary on all relevant alternative mappings is provided in **Supplementary Tables 3** and **7** and also detailed below.

Following manual analysis, certain peptides were found to be substandard with respect to PeptideAtlas criteria, and so were no longer considered to be usable as supporting evidence. As a result, certain ncORFs had their tier classification modified. For Tier 1A: 7 were moved to Tier 2A; 2 were moved to Tier 1B (i.e., having multiple HLA peptides in support); 1 was moved to Tier 2B (i.e., having one HLA peptide in support); 7 were moved to Tier 4 (i.e., there is now considered to be no supporting peptide data of any kind). For Tier 2A: 11 were moved to Tier 1B; 11 were moved to Tier 2B; 88 were moved to Tier 4. For ncORFs supported by 5+ HLA peptides, in no cases did quality concerns lead to an ncORF now having under 5 supporting HLA peptides.

Next, the Ribo-seq evidence was manually reappraised for all ncORFs under consideration thus far. If the Ribo-seq data signal was not considered adequate to support the initial Ribo-seq ORF call in Mudge *et al* (which did not involve manual analysis of the raw data), then ncORFs with non-HLA and / or HLA support were recategorised as Tier 3. Thus, 2 ncORFs in Tier 1A were moved to Tier 3, 7 in Tier 2A were moved to Tier 3, and 2 with 5+ HLA peptides were moved to Tier 3.

Thus, following peptide analysis, 18 ncORFs were classified as Tier 1A, 39 as Tier 2A, and 75 of which were classified as Tier 1B. These 132 ncORFs were taken on to the next stage of manual analysis.

Manual gene annotation

All 143 ncORFs were subjected to manual gene annotation by the GENCODE project. GENCODE is a mature project with well established guidelines, for which expert annotators can consider any available datasets to support the annotation of transcript models and transcript functionality. Here, two aspects to this work are of most importance. Firstly, we were interested in understanding the underlying structure of the transcript to which a given ncORF had been mapped, especially to check for its accuracy and to see if any could be inferred as to how its structure may relate to the physical translation of ncORF. This process is illustrated in detail for the first ncORF example below, c12norep105, and for relevant insights are discussed for certain additional examples.

Secondly, the gene annotation featured an evolutionary component. Historically, GENCODE will annotate an ORF as protein-coding if it presents strong evidence that the translation is ancient in evolutionary terms and, more importantly, has clearly evolved under constraint at the protein-level as measured by PhyloCSF. This is because we believe that a confident observation along these lines provides evidence of protein function. Here, we anticipate that those ncORFs presenting obvious evolutionary signatures of this type have already been identified by our previous work, especially Mudge *et al.* Thus, for our present purposes, the main driver in annotation decision making (discussed below) is the quality of the peptide evidence. However, an exception is c2riboseqorf47, a uORF in germ cell-less 1, spermatogenesis associated (*GMCL1*). This ncORF is supported by 2 HLA peptides, so not initially in scope for the manual analysis process detailed here. However, this ncORF presented as of immediate interest during the subsequent experimental work, having a strong phenotype in the conditional cell knock-out assay. On further analysis, it was noted that the ncORF has strong conservation in placental mammal genomes, and - moreover - it presents with a positive PhyloCSF score. On this basis, and considering the peptide and experimental evidence, it was decided to annotate this ncORF as a novel protein-coding gene (ENSG00000310604).

The ncORF commentaries below often make reference to 'ORF conservation'. This concept considers the existence of the human ncORF in other species, with respect to the structure of the mammalian phylogenetic tree. For this work, we have developed a novel evolutionary algorithm called ORBL, as described in the main text and in detail within section 4. All 7,264 ncORFs considered from the outset of this study have calculated ORBL scores, as presented in **Supplementary Table 11**. As discussed in section 4, ORBL differs from PhyloCSF, which we also deploy here, in that it does not examine a mode of evolution based on constraint at the amino acid level. Instead, it considers the presence of the initiation codon and termination codon in other genomes, as well as the maintenance of the reading frame, i.e., looking for indels leading to frame shifts or premature termination codons.

In this document, we provide additional commentary on the evolutionary picture for certain ncORFs. This is based on the manual appraisal of multi-species genome alignments, and for every ncORF we provide a link for a representative alignment produced with the CodAlignView tool, developed by I. Jungreis, M. Lin and M. Kellis at MIT and the Broad Institute. CodAlignView uses existing multispecies genome alignments, and can be set to view alignments for 100 vertebrate or 470 mammalian genomes generated at UCSC using Multiz, or 241-genome mammalian alignments generated using Cactus⁹⁶ as part of the Zoonomia project⁹⁷. The link for

each gene is set to the view we find to be most informative in terms of species incorporated, although this can be adjusted to any alignment made available for GRCh38. These commentaries are intended to add further illumination for those interested in these specific ncORFs, complementing the ORBL scores.

Certain ncORFs are described as '*de novo* emergences'. This is an established term, although one that can carry different shades of meaning. Here, *de novo* is used to reflect our view that the ncORF has evolved from a genomic sequence that is ancestrally non-coding, and thus presumably non-translated at that point in evolutionary history. Thus, a ncORF that is described as a *de novo* emergence in humans is one that does not have a conserved ORF counterpart in other species, although we do see that there is overall multigenome sequence alignment within a clade defined with respect to an ancient common ancestor.

Protein and peptidein annotation

Prior to considering protein and peptideins, we highlight that our gene annotation efforts placed a small number of ncORFs into miscellaneous categories. First, we found that the peptide evidence for six ncORFs was consistent with the validation of protein isoforms that were not annotated when the initial Ribo-seq work for Mudge *et al* was performed (in *CCNI*, *PHLDA2*, *A1BG*, *SON*, *NUP62CL* and *TNFAIP2*). Theoretically, the existence of the ncORF and the overlapping protein isoform may not be mutually exclusive. However, in each case - as discussed below - a novel GENCODE transcript model has been added to which these peptides will be mappable in a future release, and we do not consider the ncORFs as candidates for annotation as novel independent proteins or peptideins. Second, we found a case where the ncORF corresponded to the location of pseudogenic sequence that had not been annotated by GENCODE when the Ribo-seq survey was performed. Pseudogenes present complexities for our work at present, and overlapping annotations are not regarded as trustworthy for protein or peptidein annotation. Next, we found that c9riboseqorf46 represents the transposase ORF of a Mariner transposon. While the identification of transposons ORFs is interesting, in this case the peptide evidence was not found to distinguish between the called ncORF and other candidates that could be identified upon further analysis. Finally, ncORF c6norep158 was found to be miscalled in Mudge *et al* due to a genome assembly problem on GRCh38, whereby a genomic deletion significantly disrupts the structure of known protein-coding gene *CASP8AP2*. The original Ribo-seq ORF call results from the way this locus was suboptimally represented in a previous GENCODE release, and the peptides actually map to canonical *CASP8AP2* protein.

Three ncORFs presenting proteomics data suitable to validate protein annotation had already been annotated as such following Mudge *et al*, on the basis of their evolutionary signature; c14riboseqorf117 in *EIF5*, c1riboseqorf55 in *PTP4A2*, and c3riboseqorf98 in *CGGBP1*. The discovery of proteomics data in the present study thus supports these decisions.

Following our work, three ncORFs have been annotated as novel protein-coding genes based on non-HLA peptide support. These are c12norep105 in *CYP27B1* (a complex case described in

detail below), c11riboseqorf4 in *PIDD1* and c21norep46 in *ERVH48-1*. The latter case was identified in parallel by GENCODE work examining prospective functionalised transposons, and published data was noted that supports protein function⁹⁸.

In addition, as noted above, the c2riboseqorf47 uORF in *GMCL1* was also subsequently annotated as protein-coding due to a deeper consideration of its evolutionary signature alongside other experimental support.

As noted in a parallel manuscript, translations products that are confidently detected but do not meet the criteria for protein annotation may be alternatively annotated as peptideins. For this work, the route into peptidein annotation is in effect via support from non-HLA peptides from cancer samples or immortalised cell lines, or via support from HLA peptides regardless of sample provenance. Thus, we now consider 121 ncORFs to express peptideins; 43 based on non-HLA data, 75 based on HLA data, and 3 with support from both types.

However, we note that the annotation of peptideins remains a process under development. HUPO-HPP have focused on non-HLA datasets since the inception of the project, such that mature guidelines are available for data interpretation. In contrast, HLA datasets result from a newer technology, and guiding principles for the deployment of the method at the experimental and - especially - interpretive level have not yet been codified by HUPO-HPP or in the wider community. Furthermore, HLA peptides are qualitatively different from non-HLA peptides; especially, they are on average substantially smaller. Here, we consider our requirement for 5 peptides to be a conservative threshold, and emphasise that it was used specifically for this work; it does not represent a final decision by HUPO-HPP on the level of HLA peptide support for peptidein validation. Moreover, we emphasize that the manual analysis of ncORFs for potential annotation as peptideins (or proteins) is an ongoing process beyond this publication; it is effectively now part of the general remit of our annotation projects. Thus, it is highly likely that additional ncORFs will be annotated as peptideins as this work continues, and also as more datasets become available to support reanalysis. Finally, the annotation of peptideins within the database schemas of the GENCODE, Ensembl, UniProtKB, RefSeq, RefSeq and HGNC is also work that remains under development. Thus, these official annotations are initially presented with this work as a standalone .gtf file.

Discussion on ncORF cases

The first commentary on c12nore105 is expanded to illustrate aspects of our manual annotation workflow that are also applicable to other ncORFs.

Tier 1A ncORFs

C12norep105: novel protein

This ncORF is very well detected in the PeptideAtlas non-HLA build, with 7 distinct peptides. Of these, 6 have a PSM that was deemed “excellent” upon manual inspection. Only 1 peptide (AVDHGDAPLAAPPCAWALGPPLPR) was only deemed “good”. It is quite likely correct, but missing some coverage at the N terminus. This ncORF is highlighted in the main text and fully discussed there.

A key part of our efforts to elucidate ncORFs is to interpret their translation in the context of gene annotation, noting that the provenance of any translation event is substantially governed by the nature of the initial transcription event that creates the relevant RNA. This process is especially important in cases such as c12norep105, where gene annotation requires a mechanistic explanation as to how translation manages to occur in an alternative reading frame to the canonical CDS. Initially, we saw two main routes for coupling the transcription and translation of this ORF. First, converging transcriptomics data support the presence of an alternative transcript start site (TSS) in intron 2 of *CYP27B1*, which would produce a transcript within which the ATG used in c12norep105 is a plausible cognate initiation codon (existing transcript model ENST00000546567 would include the ORF in this context). In fact, there is an inframe upstream initiation codon on this model, found at chr12:57,765,524-57,765,526, that could be used to extend the ORF in the 5' direction. Second, usage of an alternative splice acceptor site in exon 3 (as seen on model ENST00000713545, and well supported by RNA-seq data, not shown) introduces a frameshift in the *CYP27B1* CDS, moving the reading frame into that which is supported by the peptides.

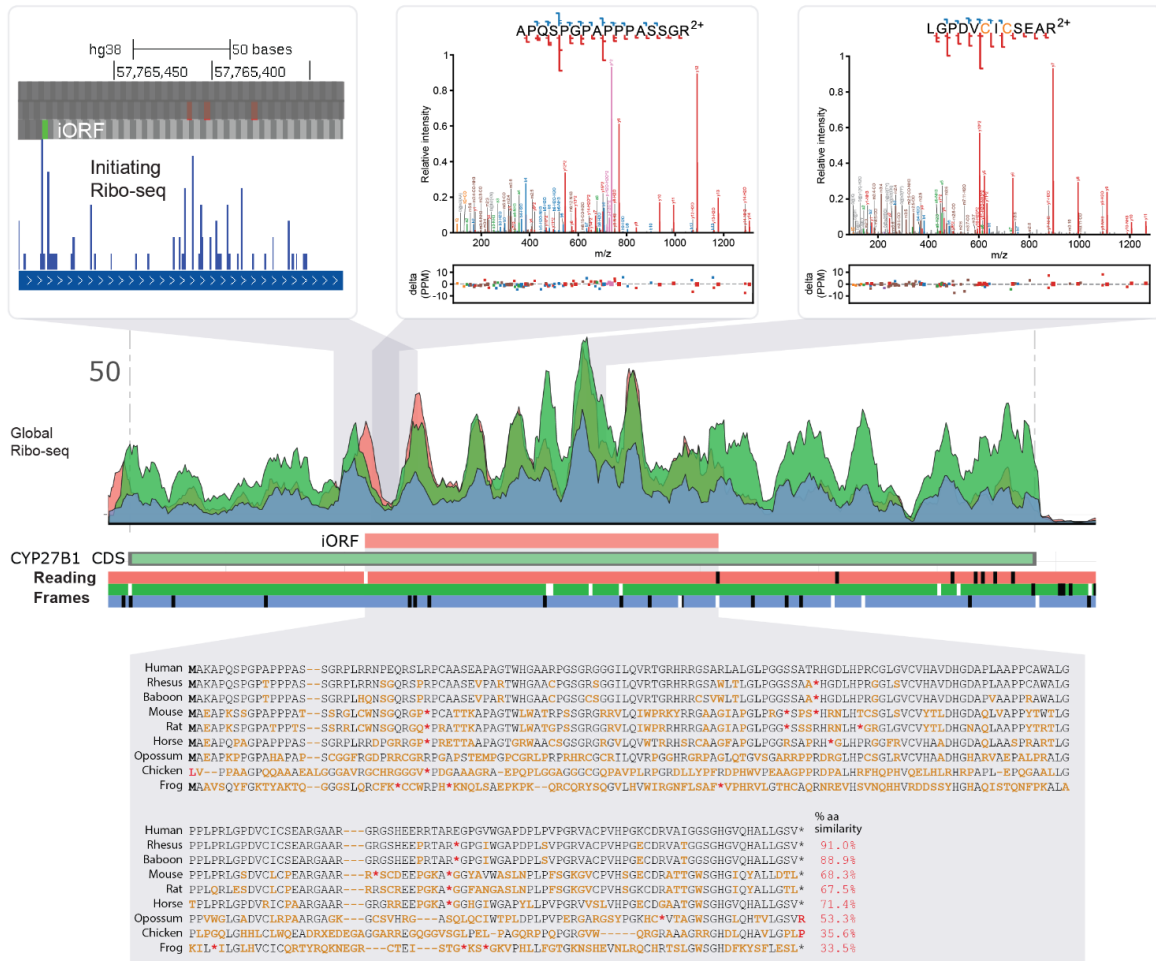
However, when manually examining the source Ribo-seq data in more detail - looking at read coverage - we see evidence that c12norep105, as called in our original study, is 5' truncated. Based on the MANE Select model ENST00000228606, the ORF remains open in the 5' direction, extending to an inframe ATG found upstream of the canonical *CYP27B1* initiation codon, and we see good Ribo-seq support for this extension. If 5' extended in this way, the c12norep105 intORF would be recontextualised as a ouORF, with a substantially larger translation of 347aa. It could in fact be that these three possibilities are not mutually exclusive, and that c12norep105 exists as a distinct ORF alongside the 5' extended form, linked to alternative transcription events. However, in contrast to the initiation codon originally called for c12norep105, this upstream ATG has excellent support in Ribo-seq experiments treated with homoharringtonine (HHT), i.e it is experimentally supported as a true initiation codon.

Regardless of its transcriptional pathway, the functionality of this translation event — i.e., its contribution to normal cellular physiology, if any — remains hard to interpret. The conservation of the ORF - whether c12norep105 or the 5' extended form - is limited to apes. As such, PhyloCSF inevitably produces a negative score, which means that this metric does not support a protein-level model of sequence evolution. Crucially, however, we observe that peptide PAp11480608 is derived from normal skin samples in a published study⁹⁹, indicating that this translation product is not aberrant or cancer specific. Meanwhile, peptide PAp11480608 was obtained from Ket-CT, which is an immortalised cell line derived from keratinocytes (the primary cell type in the outer skin). *CYP27B1* is known to be expressed in skin; it is the enzyme responsible for synthesizing the biologically active form of vitamin D. It is therefore interesting that the TSS found in intron two, with the potential to support the translation called as c12norep105, has its highest expression in keratinocytes; it must be noted though that there is also evidence of keratinocyte expression from the MANE Select TSS of *CYP27B1*.

In our view, annotation of additional coding sequence with the *CYP27B1* locus is justified by the peptide data. However, this is complicated by the outstanding questions as to the mechanism by which alternative translation occurs. GENCODE have annotated c12norep105 as protein-coding on model ENST00000546567, while the 347aa long form of the ORF is annotated as protein-coding on model ENST00000718428 (the latter is not yet present in the GENCODE public release at the time of publication). We have grouped these ORFs as part of *CYP27B1*, i.e., not as an independent gene, due to the possibility that the peptides could be validating an alternative protein isoform rather than a non-overlapping product. This is suggested by model ENST00000713545, which is also annotated as protein-coding.

The 347aa long form of the ncORF, using primate-only genomes in the Cactus alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr12%3A57%2C764%2C519-57%2C764%2C550%2Bchr12%3A57%2C764%2C754-57%2C764%2C926%2Bchr12%3A57%2C765%2C011-57%2C765%2C211%2Bchr12%3A57%2C765%2C297-57%2C765%2C499%2Bchr12%3A57%2C766%2C007-57%2C766%2C197%2Bchr12%3A57%2C766%2C847-57%2C767%2C090&strand=-&prologue=6&epilogue=6&alnset=hg38_241mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on



This figure includes the short form of the ORF, i.e., c12norep105, and indicates the lack of ORF conservation beyond apes. The top left panel indicates low-scoring though potentially discernable support for the initiation codon of this ORF in Ribo-seq experiments on HHT-treated cells. On the top right, spectra are shown for two high quality peptides.

c11riboseqorf4: novel protein

c11riboseqorf4 is an overlapping uORF in *PIDD1*, which shares substantial sequence with a second ncORF in the catalog, c11norep6. The two ORFs use distinct splice acceptor sites with respect to the exon 2-3 splice junction of the *PIDD1* MANE Select model (ENST00000347755). This does not result in a frameshift, and so c11norep6 has an additional portion of interior sequence compared to the former. Ultimately, each peptide maps to both ORFs and so cannot distinguish between them in terms of evidence for translation. The evolutionary argument is not straightforward. The ATG is perfectly conserved in mammals, and almost entirely in reptile / avian

genomes as well. The exon 2 portion of the ORF codes through in most mammals, with the exception of all rodent genomes, which have a premature STOP. The exon 3 portion is strikingly less well conserved as an ORF due to early STOPS in various lineages; the ORF is conserved as per its human length only in higher primates.

The alternative ATG of c11riboseqorf4 / c11norep6 is 31bp upstream of the canonical CDS. Essentially all transcripts produced by the gene contain both ATGs, and so it is not obvious how to distinguish between the two translational events on a transcriptional basis. The gene itself has a general expression profile.

There are 11 uniquely-mapping tryptic peptides in PeptideAtlas, with hundreds of observations. It is not listed as non-core canonical due to its similarity to c11norep6 and another predicted ncORF. But none of the 11 peptides map to UniProtKB or similar databases. Most of the PSMs are excellent and there is little doubt in this detection. Although there are substantial detections in cancer samples, there is also ample evidence from non-cancer samples. For the peptide SGLQGSPVGDGCNNGGGAR at position 2 in the protein (methionine at position 1 is never seen), all 15 PSMs are acetylated on the n terminus, as seen for example in this excellent PSM [mzspec:PXD006633:HEK293T_N_term_F03:scan:7800:\[Acetyl\]-SGLQGSPVGDGC\[Carbamidomethyl\]N\[Deamidated\]GGGAR/2](#).

The ncORF in the 120-mammal Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr11%3A803%2C398-803%2C587%2Bchr11%3A804%2C094-804%2C419&strand=-&prologue=6&epilogue=6&alnset=hg38_120mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c21norep46: novel protein

This former lncRNA ORF is now annotated as protein-coding gene *ERVH48-1* by GENCODE. This annotation decision was made in parallel to the work presented here, during a survey of UniProt / TrEMBL entries lacking corresponding GENCODE translations, and as part of a wider survey into retroelement sequences. The protein is understood to have placental function (there is transcript expression data supporting this profile), and has been considered a syncytin family member by Roberts et al.⁹⁸. The peptide evidence therefore provides good additional support for the protein-coding annotation decision, especially as six of the peptides are derived from placenta experiments. The insertion happened at the base of the old world monkey / ape clade, with the ORF being intact in its human form in apes and gibbon. Interestingly, it is observed that the 160aa protein is not obviously an *env* ORF like other syncytins, and does not in fact appear to be *env* derived. The two exon gene is indeed embedded in a HERV, but the ORF looks more likely to have evolved *de novo* from sequence that was not contributed by the transposon insertion, while

the splice junction of the 2-exon model (which has excellent transcriptomics support) also looks like an evolutionary innovation.

The ncORF using primate-only genomes in the 120-mammal Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr21%3A42%2C918%2C524-42%2C919%2C006&strand=-&prologue=6&epilogue=6&alnset=hg38_120mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c6norep158

This case is not biologically informative. The peptides simply validate the canonical protein of *CASP8AP2*, which could not be properly annotated by GENCODE due to the presence of a substantial error on the GRCh38 reference genome (an entire canonical coding exon is missing). The original Ribo-seq call onto a processed transcript model was not incorrect per se, rather the pipeline did not recognise that the GENCODE annotation was compromised.

c4riboseqorf52

c4riboseqorf52 was called as a uoORF in the 5' UTR of *CCNI*, found within MANE Select model ENST00000237654. However, gene annotation suggests an alternative provenance for the experimental evidence. The ORF has two exons. The first is clearly an ancestral coding exon, and indeed GENCODE now annotate it as such (ENST00000507788) but only as a model that then skips MANE coding exon 1 and splices in downstream and continues in the CDS frame. In other words, the first exon of the ORF can be used to make an alternative isoform of *CCNI*. The second exon of the ORF is therefore what happens when the first coding exon of *CCNI* is not skipped, and it translates dual frame with respect to coding exon 1 before reaching a termination codon. Importantly, the peptides all map to the first exon of the ORF, which means that the simpler explanation is that they support the existence of the alternative *CCNI* isoform rather than the ORF as a protein product. Thus, c4riboseqorf52 has not been annotated as a protein-coding gene.

The ncORF in the Cactus alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr4%3A77%2C066%2C298-77%2C066%2C405%2Bchr4%3A77%2C075%2C472-77%2C075%2C543&strand=-&prologue=6&epilogue=6&alnset=hg38_241mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c11norep11 / c11norep12: novel peptideins

c11norep11 is a lncRNA ORF on *IGF2-AS*, sharing substantial sequence overlap with c11norep12. The two ORFs differ in the usage of alternative splice donor sites leading into the second exon, which does not induce a frameshift; c11norep12 thus has a 37aa insertion. Both versions of the ORF are only fully conserved in ape and gibbon genomes, and look like a *de novo* emergence. UniProtKB already recognises c11norep12 as Q6U949. This protein was curated from Vu et al.¹⁰⁰, which predicts the ORF but does not provide experimental evidence for protein existence.

The peptides which support both translations - or rather cannot distinguish between them - are observed only in cancer samples or cell lines. Interestingly, the overexpression of the *IGF2-AS* transcript has been observed in certain tumours^{100,101}. Peptide PAp00139202 spans the unique splice junction of c11norep12, and is observed in blood plasma.

The ncORF using primate-only genomes in the Cactus alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr11%3A2%2C140%2C644-2%2C140%2C947%2Bchr11%3A2%2C146%2C356-2%2C146%2C447&prologue=6&epilogue=6&alnset=hg38_241mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c11riboseqorf108: peptidein

c11riboseqorf108 was called as a lncRNA ORF within *LIPT2-AS1*. On further analysis, it corresponds to an ORF contributed by a Tigger-family transposon. The insertion occurred at the base of the monkey / ape clade, and the transposase-like ORF is intact in these genomes. This could be evidence for a 'domestication' event, whereby a functional transposon ORF gains a new role in the host genome. However, all of the peptides found in support are from cancer samples or cell lines. For this reason, GENCODE have not yet decided to annotate this ORF as protein-coding. The CHES gene annotation project already considers this locus a protein-coding gene with the same ORF, CHS.9489.1.

The ncORF using primate-only genomes in the Cactus alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr11%3A74%2C497%2C035-74%2C498%2C084&prologue=6&epilogue=6&alnset=hg38_241mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c14norep4: novel peptidein

c14norep4 is a uORF of *ZNF219*. It maps to the MANE Select model; there are alternative first exons in the locus to which the ORF does not map. The ATG is found in all mammal genomes, although the length of the ORF varies substantially between species. There are 4 mapping peptides in PeptideAtlas, but the 7 amino-acid long peptide AAAAAAR is immediately discounted since it maps to many genes and so is not useful. This leaves 3 distinct peptides spread out over most of the length. Peptide AGPPPAAHNAGQGR has many matching ions but was deemed a false positive upon manual inspection. This leaves 2 peptides with excellent spectra that provide the HPP-level evidence. However, nearly all evidence comes from CPTAC cancer samples.

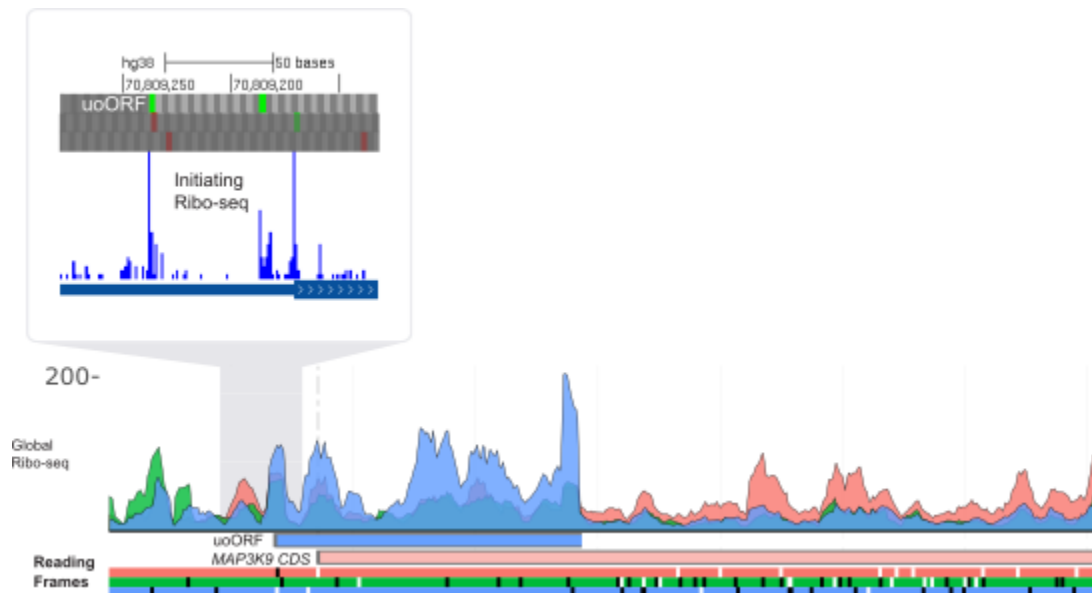
The ncORF in the 120-mammal Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr14%3A21%2C093%2C648-21%2C093%2C674%2Bchr14%3A21%2C098%2C312-21%2C098%2C527&strand=-&prologue=6&epilogue=6&alnset=hg38_120mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c14norep77: novel peptidein

c14norep77 is an upstream overlapping ORF (uoORF) of *MAP3K9*, splicing across the first two exons of MANE Select model ENST00000554752. The ATG is found in all placental mammals, and the ORF is largely conserved as well; certain lineages have earlier STOP codons. As such, it looks promising as a functional element, i.e., according to the evolutionary argument. Ribo-seq data readily resolves the c14norep77 and *MAP3K9* CDS translations, as illustrated below, including initiating Ribo-seq support for both ATGs. The 67bp portion of the ORF prior to the dual frame region is slightly PhyloCSF positive, after which the signal becomes harder to interpret. It has 5 supporting peptides, but 3 of the 5 are judged as not very good quality, primarily in the low-complexity poly-glycine region of the protein. For example, while the peptide GGGGGGGGGGGGGGGPR has some spectra that have multiple matching ions, it must be assumed that the true identification is elusive and not this sequence. These peptides are discounted in this analysis, but remain visible in the PeptideAtlas interface for inspection by the reader. This leaves 2 peptides with excellent spectra such as https://proteomecentral.proteomexchange.org/usi/?usi=mzspec:PXD020389:20171031_QE5_nLC3_AKP_UBIsite_SY5Y_CBLsKD_L-H_E1_17F_09:scan:9958:AAEPAGGHLPQQLR/2 that provide the required HPP-level evidence. Thus, GENCODE view c14norep77 as a potential

protein, but it has not yet been annotated as such because the two informative peptides are only seen in cancer cell line experiments.



The ncORF in the 120-mammal Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr14%3A70%2C801%2C068-70%2C801%2C080%2Bchr14%3A70%2C808%2C766-70%2C809%2C238&strand=-&prologue=6&epilogue=6&alnset=hg38_120mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c16riboseqorf59: novel peptidein

c16riboseqorf59 is a dORF in the 3' UTR of *RNF40*, found in MANE Select model ENST00000324685.11. The ORF is a *de novo* emergence, found intact in chimp and gorilla only. There is an inframe upstream ATG at chr16:30,775,406-30,775,408 that lacks clear Ribo-seq support, while an inframe downstream ATG at chr16:30,775,583-30,775,585 appears to have substantially higher HHT support than for the ATG originally called. Usage of either of these alternative codons for initiation does not change the evolutionary picture.

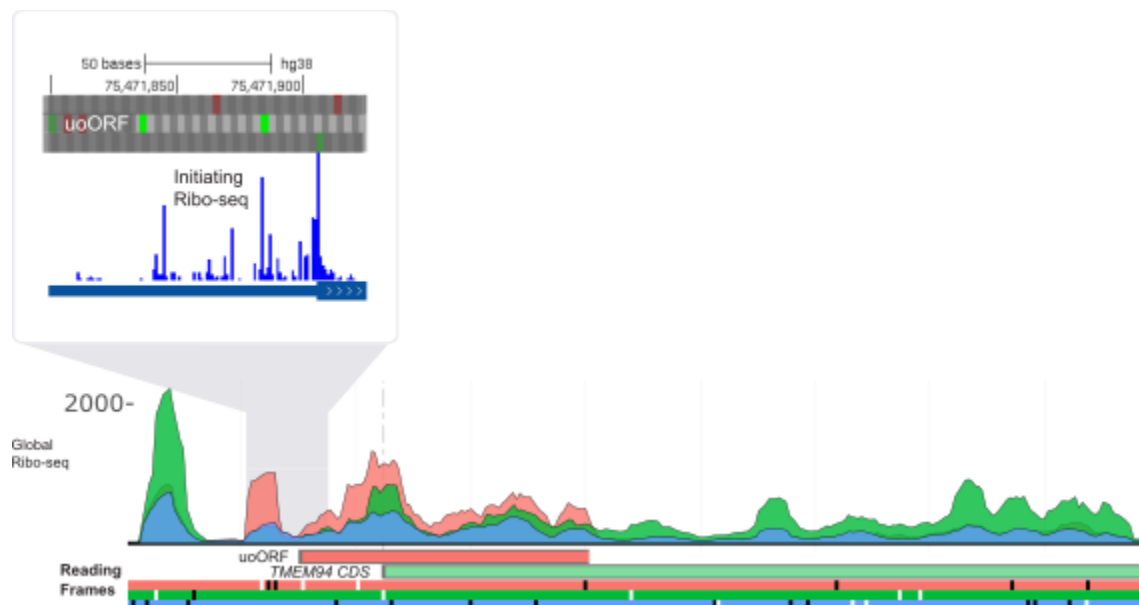
PAP14808288 was observed in normal digestive tissue, specifically as part of an N-terminus enrichment assay designed to obtain initiation sites. However, this is linked to a third downstream ATG at chr16:30,775,781-30,775,783, which does not present with clear HHT data.

The ncORF using primate-only genomes in the Cactus alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr16%3A30%2C775%2C469-30%2C776%2C053&prologue=6&epilogue=6&alnset=hg38_241mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c17riboseqorf149: novel peptidein

c17riboseqorf149 is an uoORF in the 5' UTR of *TMEM94*, found on MANE Select transcript ENST00000314256. Ribo-seq data readily resolves the c17riboseqorf149 and *TMEM94* CDS translations, as shown below, including initiating Ribo-seq support for both ATGs. The ORF is intact in the vast majority of mammal genomes, with some localised movement of the termination codon in certain lineages. The excellent spectrum is from peptide PAP11219913, which is a cancer detection. Peptide PAP11219913 is observed in multiple normal tissues. However, as this peptide is not classed as excellent, GENCODE have not annotated this ORF as protein-coding.



The ncORF using the Cactus alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr17%3A75%2C471%2C836-75%2C471%2C929%2Bchr17%3A75%2C485%2C428-75%2C485%2C547%2Bchr17%3A75%2C485%2C871-75%2C485%2C905&prologue=6&epilogue=6&alnset=hg38_241mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c19norep29: novel peptidein

c19norep29 is a uoORF in the 5' UTR of *LONP1*, found in MANE Select ENST00000360614. The ATG is found in all therian mammals, and the ORF is largely conserved with the exception of certain lineages which use alternative termination codons. All of the peptide evidence is derived from cancer samples.

The ncORF in the Cactus alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr19%3A5%2C719%2C805-5%2C720%2C143&strand=-&prologue=6&epilogue=6&alnset=hg38_241mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c19riboseqorf4: novel peptidein

c19riboseqorf4 is a uORF of the gene *STK11* found in the MANE Select transcript model ENST00000326873, and relatively large at 209aa. The 5' UTR region is ancestrally non-coding, making the ORF a *de novo* emergence. Most placental mammals have the ATG (mouse does not), but ORF conservation outside of higher primates is not consistent while PhyloCSF does not support a protein-level mode of evolution. GENCODE would not therefore annotate this ORF as protein-coding according to an evolutionary argument alone. The discovery of high quality peptide evidence is very interesting. This ncORF has 4 supporting peptides, all of which have best-PSMs deemed "excellent". For example: https://proteomecentral.proteomexchange.org/usi/?usi=mzspec:PXD020389:20171031_QE5_nLC3_AKP_UBIsite_SY5Y_CBLsKD_L-H_E1_17F_08:scan:11025:GGAFAFGTGGGTGASPETPPAK/2. It is detected in three separate datasets, PXD20389, PXD028647, and MSV000078509, all of which are derived from cancer cell line samples. In fact, each of the non-HLA peptides is derived from similar experiments, thus raising the possibility that it is an aberrant product as opposed to a functional protein molecule.

The ncORF in the 120-mammal Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr19%3A1%2C205%2C901-1%2C206%2C530&prologue=6&epilogue=6&alnset=hg38_120mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c1riboseqorf94: novel peptidein

c1riboseqorf94 is a uORF of *CDKN2C*. The gene has an unusually large 5' UTR exon containing four translated ORFs in the Ribo-seq catalog, of which this is the longest at 180aa. Note that the ORF is not found on the MANE Select transcript ENST00000371761, which has a shortened 5' UTR. The ATG is conserved in mammals, though beyond apes there is inconsistency in the termination codon used. The ORF has its own uORF - c1norep114 - which overlaps on the STOP / ATG. There are 4 uniquely-mapping tryptic peptides in PeptideAtlas. Three have multiple PSMs, some of which are excellent. One peptide has only a single PSM that is severely contaminated by other precursors in the selection window, but may well be correct. All 61 PSMs come from CPTAC cancer samples or cancer cell lines.

The ncORF in the 120-mammal Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr1%3A50%2C968%2C788-50%2C969%2C330&prologue=6&epilogue=6&alnset=hg38_120mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c5norep4: novel peptidein

c5norep4 is a dORF in the 3' UTR of *SLC9A3*. The ATG is conserved in apes and gibbon, but the ORF is human specific due to 1bp indel. The 3' UTR region is present in other mammals but not as an ORF, making this a *de novo* emergence. There are 4 uniquely-mapping tryptic peptides in PeptideAtlas, two of which have a high number of PSMs (43 and 17), although the other two are only single PSM peptides. The peptide with 43 PSMs has highly-convincing complete-coverage spectra associated with it. The other peptides are longer and do not have quite complete coverage but are quite likely correct. All PSMs come from 8 different cancer datasets from CPTAC enriched for phosphopeptides. It may be that this protein is normally of very low abundance, but brought into the detectable range via phosphopeptide enrichment.

The ncORF using primate-only genomes in the Cactus alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr5%3A472%2C072-472%2C413&strand=-&prologue=6&epilogue=6&alnset=hg38_241mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c9norep168: novel peptidein

c9norep168 is called as a lncRNA ORF within *ARRDC1-AS*, and it has a substantial same frame overlap with c9riboseqorf120. The two ORFs have the same 3' end, although c9riboseqorf120 is a single exon sequence whereas c9norep168 contains a distinct first exon as part of a two exon model. Q9H2J1 is a UniProt entry corresponding to c9riboseqorf120. This entry, apparently based on Zou et al.¹⁰², does not provide any functional annotation for the protein. The lncRNA is limited to primates, and the ORF is intact only in apes and gibbon. The peptide evidence, which does not distinguish between the two ncORFs, is derived from cancer samples. Hence GENCODE has not annotated either ncORF as protein-coding. Interestingly, Zou et al.¹⁰² provide evidence for the expression of the lncRNA in cancer samples.

The ncORF in the primate-only Cactus alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr9%3A137%2C615%2C669-137%2C616%2C132%2Bchr9%3A137%2C617%2C388-137%2C617%2C463&strand=-&prologue=6&epilogue=6&alnset=hg38_241mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c8riboseqorf15: novel peptidein

c8riboseqorf15 is a uORF of *CNOT7*, found on the MANE Select transcript ENST00000361272.9. The ORF is perfectly conserved in all mammal genomes, with the exception of platypus. The non-HLA peptides found in non-cancer samples or cell lines largely consist of the proline repeat sequence, and are not trustworthy to support reference annotation. However, there is also support from HLA peptides.

The ncORF in the 120-mammal Multiz alignment:

<https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr8%3A17%2C245%2C236-17%2C245%2C247%2Bchr8%3A17%2C246%2C675-17%2C246%2C782&strand=->

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr6%3A26%2C376%2C596-26%2C377%2C276&prologue=6&epilogue=6&alnset=hg38_120mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c6norep53: pseudogenic

c6norep53 is a dORF in the 3' UTR of *BTN3A2*. Upon deeper analysis, it became apparent that the ORF is a 'pseudoexon' sequence. In other words the C-terminus of the standard butyrophilin family protein is lost in this primate-specific copy, due to a premature termination codon in exon 7 compared with other full-length paralogs. So the dORF is a remnant of sequence that is coding in other copies, and the ncORF was able to be called during the creation of the catalog because this sequence has not been annotated by GENCODE (ncORFs overlapping known pseudogenes were filtered out). With this knowledge, it becomes clear that the peptides do not distinguish this ORF from O00478, the UniProt entry for *BTN3A3*.

The ncORF using primate-only genomes in the Cactus alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr6%3A26%2C376%2C596-26%2C377%2C276&prologue=6&epilogue=6&alnset=hg38_241mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

Tier 2A

c10norep59: peptidein

c10norep59 is a lncRNA ORF in *LINC00839*. c10norep60 is a variant form, which uses an internal initiation codon and an intronic termination codon. Q8NAU0 is a TrEMBL entry matching c10norep59, while B4DDC5 is an alternative TrEMBL isoform from the same locus, distinct from both c10norep59 and c10norep60. The single candidate peptide thus matches to all four translations in the same location. Gene annotation finds that c10norep59 is substantially better supported by Ribo-seq data, and that the alternative termination codon of c10norep60 is potentially a dubious call. Hence only the former has been annotated. This peptide is observed in a cancer sample.

The ncORF using primate-only genomes in the Cactus alignment:

<https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr10%3A42%2C475%2C573-42%2C475%2C880%2Bchr10%3A42%2C477%2C261->

42%2C477%2C442%2Bchr10%3A42%2C486%2C979-
42%2C487%2C055&prologue=6&epilogue=6&alnset=hg38_241mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c10riboseqorf22

c10riboseqorf22 is a lncRNA ORF on *WAC-AS1*. This ORF used to have a corresponding Trembl entry D3DRW4, which was recently removed from the database. There are three other ncORFs in the gene, none of which overlap. The lncRNA itself is only present in higher primates, with the ORF intact only in ape and gibbon. The ORF is entirely embedded within a LINE element. The single excellent peptide is derived from a cancer sample.

The ncORF using primate-only genomes in the Cactus alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr10%3A28%2C523%2C873-28%2C524%2C190&strand=-&prologue=6&epilogue=6&alnset=hg38_241mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c11norep103

C11norep103 is a uORF in *MAP3K11*; it was previously called as a ncORF on a processed transcript, as the model used for initial interpretation was annotated as a retained intron model (ENST00000524856.1). The ORF is conserved in higher primates, and represents a *de novo* emergence. The supporting peptide is observed only in cancer cell lines.

The ncORF using primate-only genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr11%3A65%2C614%2C947-65%2C615%2C156&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c11norep16

c11norep16 is an doORF in *PHLDA2*, i.e., it initiates in an alternative frame to the CDS then translates downstream of the canonical termination codon and into the 3' UTR. The initiation codon of the ncORF overlaps with the *PHLDA2* termination codon as [ATGA], although - despite strong read coverage for the gene as a whole - no support for this initiation codon in HHT-treated samples is observed. In fact, following manual gene annotation an alternative, simpler explanation for the Ribo-seq signal and peptide support is apparent: usage of a cryptic splice donor site in exon 1 at chr11:2929166-2929168 causes a frameshift leading into exon 2, and switches the *PHLDA2* reading frame into that which is supported by the peptide evidence. Thus, it is likely that the Ribo-seq signal was originally miscalled as a doORF, when in reality it is derived from an alternative isoform of *PHLDA2*. This new coding sequence has been annotated by GENCODE as ENST00000718435.

c11norep190: novel peptidein

c11norep190 is a dORF in *H2AX*. The ATG is present in higher primate genomes, with the ORF conserved only in apes and gibbon. The supporting peptide has been identified in the frontal cortex; the gene itself is expressed in all tissues, based on transcriptomics data.

The ncORF using primate-only genomes in the Cactus alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr11%3A119%2C094%2C706-119%2C094%2C885&strand=-&prologue=6&epilogue=6&alnset=hg38_241mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c11norep89: novel peptidein

c11norep89 is a lncRNA ORF in PPP1R14B-AS1. The ATG and ORF are conserved in the clade defined by apes, old world monkeys and new world monkeys. The supporting peptide is observed only in cancer samples.

The ncORF using primate-only genomes in the Cactus alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr11%3A64%2C247%2C221-64%2C247%2C259%2Bchr11%3A64%2C247%2C953-64%2C248%2C063&prologue=6&epilogue=6&alnset=hg38_241mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c12norep197: novel peptidein

c12norep197 is a uoORF in *MLXIP*, found in MANE Select model ENST00000319080. The ATG is found in all mammals (noting that some genome alignments are missing in Multiz / Cactus), and the ORF is almost entirely intact throughout the order with the exception of a few lineages. The supporting peptide is derived from a cancer sample.

The ncORF in the Cactus alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr12%3A122%2C078%2C831-122%2C079%2C253&prologue=6&epilogue=6&alnset=hg38_241mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c13riboseqorf32: novel peptidein

c13riboseqorf32 is a uORF in *CAB39L*. The ATG and ORF are almost entirely conserved in primates genomes; conservation in other mammal genomes is harder to interpret, with the ORF disrupted in several lineages. The single supporting peptide is non-tryptic, and only observed in cancer cell lines.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr13%3A49%2C433%2C382-49%2C433%2C393%2Bchr13%3A49%2C434%2C086-49%2C434%2C151&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c14riboseqorf98: novel peptidein

c14riboseqorf98 is an uoORF in *FOXN3*. The ATG and ORF are well conserved in mammals, with some localised movement of the termination codon. The single peptide is only observed in cancer samples.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr14%3A89%2C412%2C451-89%2C412%2C490%2Bchr14%3A89%2C416%2C871-89%2C417%2C070&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c15riboseqorf103: novel peptidein

c15riboseqorf103 is a uORF in *CHSY1*. The ATG and ORF are conserved in mammalian genomes, with few exceptions. The single peptide has an observation in placenta; the gene is very highly expressed in this tissue based on RNAseq and CAGE data. The ORF is 124aa, so if protein-coding it would not be classed as a microprotein.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr15%3A101%2C251%2C543-101%2C251%2C914&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c17norep1: novel peptidein

c17norep1 is a lncRNA ORF in *RPH3AL-AS1*, conserved only in the human genome. The single peptide is observed only in cancer samples.

The ncORF using primate-only genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr17%3A331%2C275-331%2C598&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c18riboseqorf10: novel peptidein

c18riboseqorf10 is a lncRNA ORF in *LINC00526*, conserved only in apes. The single peptide is observed only in cancer samples. UniProt annotate this precise ORF as Q96FQ7, although it is classified by them as PE5 ('product of a dubious gene prediction') so is not taken forward by HUPO-HPP.

The ncORF using primate-only genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr18%3A5%2C237%2C364-5%2C237%2C651&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c19norep18: novel peptidein

c19norep18 is an intORF within the CDS of *NFIC*, found in MANE Select model ENST00000443272. The ATG of the ncORF was called immediately downstream of a splice acceptor site, and without HHT support for initiation, which suggests it may not have an accurately interpreted 5' end. However, we do not observe a 5' extension to the ORF that is convincingly supported by Ribo-seq or evolutionary analysis. The ATG of the ncORF is found in all mammals, although the termination codon shows substantial variability and the ORF is fully conserved only in higher primates. The supporting peptide is derived from a cancer sample.

The ncORF in the Cactus alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr19%3A3%2C381%2C713-3%2C381%2C964&prologue=6&epilogue=6&alnset=hg38_241mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c19riboseqorf160: novel peptidein

c19riboseqorf160 is a uoORF in the 5' UTR of *ZNF524*, found in MANE Select model ENST00000301073.4. The ATG is found in all mammals, although the termination codon shows positional variation and the ORF is only fully conserved in higher primates. The supporting peptide is derived from a cancer sample.

The ncORF in the Cactus alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr19%3A55%2C600%2C398-55%2C600%2C408%2Bchr19%3A55%2C602%2C075-55%2C602%2C378&prologue=6&epilogue=6&alnset=hg38_241mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c19riboseqorf180: novel CDS isoform

c19riboseqorf180 is a dORF in *A1BG*. The ATG and ORF are conserved in higher primates. The single peptide is observed in blood plasma, which is interesting because *A1BG* is a glycoprotein secreted into plasma. The ORF as called by Ribo-seq is potentially translated from a two exon model based on a strong TSS found at approximately chr19:58347673, i.e., as seen on current GENCODE model ENST00000598345. Alternatively, the usage of an alternative splice donor site in the penultimate exon of the *A1BG* mRNA at chr19:58347427 would bring the *A1BG* CDS into the same frame as the dORF, and in this case the peptide could be explained as supporting a new protein isoform. This alternative splice junction has substantial transcriptomics support and so can be annotated as ENST00000850950, but is nonetheless a minor transcript in comparison to MANE Select model ENST00000263100. These two explanations may not be mutually exclusive. Ultimately, the locus is intriguing, though hard to comprehensively appraise without further data.

The ncORF using primate-only genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr19%3A58%2C346%2C882-58%2C347%2C022&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c19riboseqorf183: novel peptidein

c19riboseqorf183 is a lncRNA ORF in ENSG00000232098 / *ZNF584-DT*. The ATG and ORF are conserved in apes. The single peptide is observed in non-cancer cell lines.

The ncORF using primate-only genomes in the 470 mammals Multiz alignment:

<https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr19%3A58%2C406%2C715-58%2C406%2C770%2Bchr19%3A58%2C408%2C127->

[58%2C408%2C343&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on](https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr20%3A16%2C526%2C147-16%2C526%2C205%2Bchr20%3A16%2C528%2C371-16%2C528%2C440%2Bchr20%3A16%2C573%2C229-16%2C573%2C282&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on)

c20riboseqorf22: novel peptidein

c20riboseqorf22 ouORF in *KIF16B*. The ATG and ORF are conserved in mammals with some localised movement of the termination codon. The single peptide is observed in cancer cell lines.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr20%3A16%2C526%2C147-16%2C526%2C205%2Bchr20%3A16%2C528%2C371-16%2C528%2C440%2Bchr20%3A16%2C573%2C229-16%2C573%2C282&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c20riboseqorf61

c20riboseqorf61 is a lncRNA ORF in *MHENCR*. The ATG and ORF are human specific. The single peptide is observed only in cancer samples and cell lines.

The ncORF using primate-only genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr20%3A63%2C627%2C323-63%2C627%2C392%2Bchr20%3A63%2C628%2C190-63%2C628%2C362&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c21norep26

c21norep26 was called as a dORF in the 3' UTR of *SON*. However, manual gene annotation makes it clear that the reading frame should instead be understood as the C-terminus of an alternative protein isoform of *SON*, which results from an alternative splicing reaction. This

alternative isoform was recognised and annotated by GENCODE as ENST00000704334. The supporting peptide is observed in numerous normal tissues.

c2norep161: novel peptidein

c2norep161 is a lncRNA ORF in *LCT-AS1*. The ATG and ORF are specific to apes. The single peptide is observed only in cancer cell lines.

The ncORF using primate-only genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr2%3A135%2C820%2C546-135%2C820%2C639%2Bchr2%3A135%2C821%2C675-135%2C821%2C871&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c2riboseqorf145: novel peptidein

c2riboseqorf145 is a uORF in *PLCL1*. The ATG and ORF are conserved in mammal genomes, with very few exceptions. The ORF is highly alanine-rich, in a manner that is linked to length variation in different species. The single peptide is observed in cancer cell lines only.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr2%3A197%2C804%2C922-197%2C804%2C987&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c2riboseqorf167: novel peptidein

c2riboseqorf167 is an uORF in the 5' UTR of *MARCHF4*, found in the MANE Select model ENST00000273067. The ATG initiation codon of the ORF overlaps with the STOP codon of c2norep257 found immediately upstream. The c2riboseqorf167 ORF is fully intact in almost all therian mammal genomes. The supporting peptide is derived from a cancer sample.

The ncORF in the Cactus alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr2%3A216%2C371%2C336-216%2C371%2C599&strand=-&prologue=6&epilogue=6&alnset=hg38_241mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c2riboseqorf90

c2riboseqorf90 is a dORF in the 3' UTR of *PTPN18*. The ORF is conserved in apes and gibbon, and the genomic alignment of the region in other mammals indicates that it is a *de novo* emergence. There are 3 uniquely-mapping tryptic peptides in PeptideAtlas. However, one is nested within another and thus does not count for HPP guidelines, so only 2 HPP-compliant peptides remain. One has 5 PSMs, some of which are extremely compelling with complete coverage, but the other only has a single PSM. This PSM has good coverage but suffers from contamination from another precursor in the selection window, and is thus not ideal. Most PSMs come from cancer samples. Only a single PSM comes from a deep pituitary gland sample.

The ncORF using primate-only genomes in the Cactus alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr2%3A130%2C373%2C716-130%2C373%2C880&prologue=6&epilogue=6&alnset=hg38_241mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c3riboseqorf16: novel peptidein

c3riboseqorf16 is a lncRNA ORF in *FGD5-AS1*, located entirely with a TcMar-Mariner family DNA transposon; see also comments for c9riboseqorf46. It overlaps with c3norep21 and c3norep22 in alternative reading frames. The transposon is inserted in higher primates, and the ORF is only conserved in apes. The single unique-mapping peptide is observed in non-cancer cell lines. Other peptides cannot distinguish this ORF from e.g. *SETMAR* protein.

The ncORF using primate-only genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr3%3A14%2C945%2C419-14%2C945%2C598&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c3riboseqorf166: novel peptidein

c3riboseqorf166 is a uoORF in *FNDC3B*, with the start codon immediately upstream of the CDS. The ATG is conserved in vertebrate genomes, with the position of the termination codon showing variability. The peptide is noted as multimapping by PeptideAtlas because it also maps to smORFs within the locus published by Cui *et al.* It is observed in cancer samples only.

The ncORF in the 100 tetrapod alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr3%3A172%2C112%2C476-172%2C112%2C590%2Bchr3%3A172%2C133%2C471-172%2C133%2C490&prologue=6&epilogue=6&alnset=hg38_100_tetrapod&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c4riboseqorf34: novel peptidein

c4riboseqorf34 is a uORF in *OCIAD1*. The ATG and ORF are conserved only in chimp and gorilla genomes. The supporting peptide includes observations in normal tissues. The peptide maps also to RefSeq models XP_024309873.1, XP_024309877.1 within the locus, both of which contain an alternative N-terminus that represents the first portion of the Ribo-seq ORF before a splice junction leads the translation into the same frame as the canonical CDS. The peptide maps to this shared sequence and so it is plausible that it validates this putative isoform as opposed to the ORF. These possibilities are not mutually exclusive; nonetheless, the Ribo-seq data provides substantially clearer support for the uORF as an intact translated sequence, i.e., as originally called, and not for the putative isoform. As such, the ncORF has been annotated as a peptidein.

The ncORF using primate-only genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr4%3A48%2C831%2C102-48%2C831%2C206&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c5riboseqorf2: novel peptidein

c5riboseqorf2 was called as a lncRNA ORF in *SLC9A3-OT1*, but it is also a dORF in *SLC9A3*; the two genes overlap on the same strand. The ATG and ORF are specific to the human genome, being a *de novo* emergence. The ORF is identical to the unreviewed TrEMBL entry Q71RB1. One supporting peptide has been observed in normal tissues.

The ncORF using primate-only genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr5%3A471%2C793-471%2C823%2Bchr5%3A472%2C679-472%2C935&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c6norep103: novel peptidein

c6norep103 is an uoORF in *DAXX*, sharing its first of two coding exons with c6riboseqorf53 and c6riboseqorf54, which are uORFs. This first coding exon - to which the peptide maps - can potentially be used to produce an alternative DAXX protein isoform via a splice site shift, as seen in model ENST00000706094. Inspection of the Ribo-seq data indicates that the Ribo-seq ORF exists as called, as do ncORFs c6riboseqorf53 and c6riboseqorf54. The putative isoform, in contrast, has weak RNAseq support, suggesting that the ncORF forms are the dominant expression product. Hence, c6norep103 has been annotated as a peptidein. While the ATG and ORF are observed in most mammal genomes, several lineages are missing the former and others have alternative termination codons. The supporting peptide is observed in cancer samples.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr6%3A33%2C321%2C345-33%2C321%2C567%2Bchr6%3A33%2C322%2C862-33%2C322%2C914&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c6norep174: novel peptidein

c6norep174 is a dORF in *HDAC2*. The ATG is observed in most primate species, although the termination codon is unique to human. The supporting peptide is observed only in cancer cell lines.

The ncORF using primate-only genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr6%3A113%2C939%2C147-113%2C939%2C239&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c6riboseqorf17: novel peptidein

c6riboseqorf17 is an uoORF in *MRS2*. The ATG and ORF are conserved in mammal genomes, although with some localised termination codon movement. The supporting peptide is observed only in cancer cell lines.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr6%3A24%2C403%2C001-24%2C403%2C236%2Bchr6%3A24%2C405%2C168-24%2C405%2C171&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSite=s=human&hideInserts=on&hideJumps=on

c7norep19: novel peptidein

c7norep19 is a dORF in *ACTB*. The evolutionary history is complex. The region as a whole is conserved in vertebrates, strikingly so across several sequences. In terms of phylogenetics, the ORF seems most likely to be a mammalian *de novo* emergence that has been lost in several lineages, although length variation across a simple repeat region complicates the picture. The single peptide is observed in cancer cell lines.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr7%3A5%2C527%2C342-5%2C527%2C683&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c8norep15: novel peptidein

c8norep15 is a uoORF in *LZTS1*. The ATG is conserved in mammals with the exception of rodents, and also in reptiles and avian species. The ORF is conserved only in higher primates due to indels and substantial variation in the position of the termination codon. The peptide is observed only in cancer samples.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr8%3A20%2C255%2C055-20%2C255%2C279&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c8riboseqorf118: novel peptidein

c8riboseqorf118 is a uORF in *MAF1*. The ATG and ORF are conserved in mammalian genomes. The single peptide is observed in urine, while *MAF1* has general expression.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr8%3A144%2C104%2C625-144%2C104%2C747&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c9riboseqorf46: no annotation

c9riboseqorf46 is a lncRNA ORF in *UBQLN1-AS1*. The entire ORF is contributed by a TcMar-Mariner transposon, giving a transposase-like putative protein. The insertion happened prior to the radiation of higher primates, although the ORF is only intact in apes and gibbon. As for c11riboseqorf8, this could be a 'domestication' event, whereby a functional transposon ORF gains a new role in the host genome. This ORF corresponds to B7Z6N6, an unreviewed TrEMBL entry.

However, on further inspection, the peptides also map to Q53H47, the UniProt entry for known protein-coding gene *SETMAR*. Thus, GENCODE will not annotate this ORF as protein-coding until locus-specific peptide evidence becomes available.

The ncORF using primate-only genomes in the Cactus alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr9%3A83%2C712%2C501-83%2C712%2C995&prologue=6&epilogue=6&alnset=hg38_241mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c9riboseqorf74: novel peptidein

c9riboseqorf74 is an uoORF in *PHF19*. The ATG is conserved in mammals, with the ORF conserved only in primates. The peptide is observed only in cancer cell lines.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr9%3A120%2C874%2C047-120%2C874%2C060%2Bchr9%3A120%2C874%2C556-120%2C874%2C756%2Bchr9%3A120%2C877%2C091-120%2C877%2C142&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

cXnorep77: novel peptidein

cXnorep77 is a doORF in *NUP62CL*. However, it appears that the ncORF represents a truncation of an alternative protein isoform of *NUP62CL* that was not annotated when the set was constructed; splicing in a cassette exon at chrX:107,150,856-107,150,967 induces a frameshift in the *NUP62CL* CDS which puts the translation into the same frame as the ncORF. Given the absence of HHT data to support the internal initiation codon called for the ncORF, it seems most likely that the peptide supports instead this alternative protein isoform, which is represented by ENST00000432145. The peptide is observed only in cancer cell lines.

c12norep33: novel peptidein

c12norep33 was called as a processed transcript in *PRH1*, which in GENCODE terms is a model within a protein-coding gene that is not annotated with a translation of any kind. Further inspection indicates that the ncORF is better contextualised as a uORF, although these 5' UTR exons are shared with *PRR4* as part of a complex locus. The ATG and ORF are conserved in higher primates. The peptide has been observed in normal tissues, and the ncORF is also supported by 16 HLA peptides; one of the highest counts. F1T0A8 is an unreviewed TrEMBL entry corresponding to the same ORF.

The ncORF using primate-only genomes in the Cactus alignment:

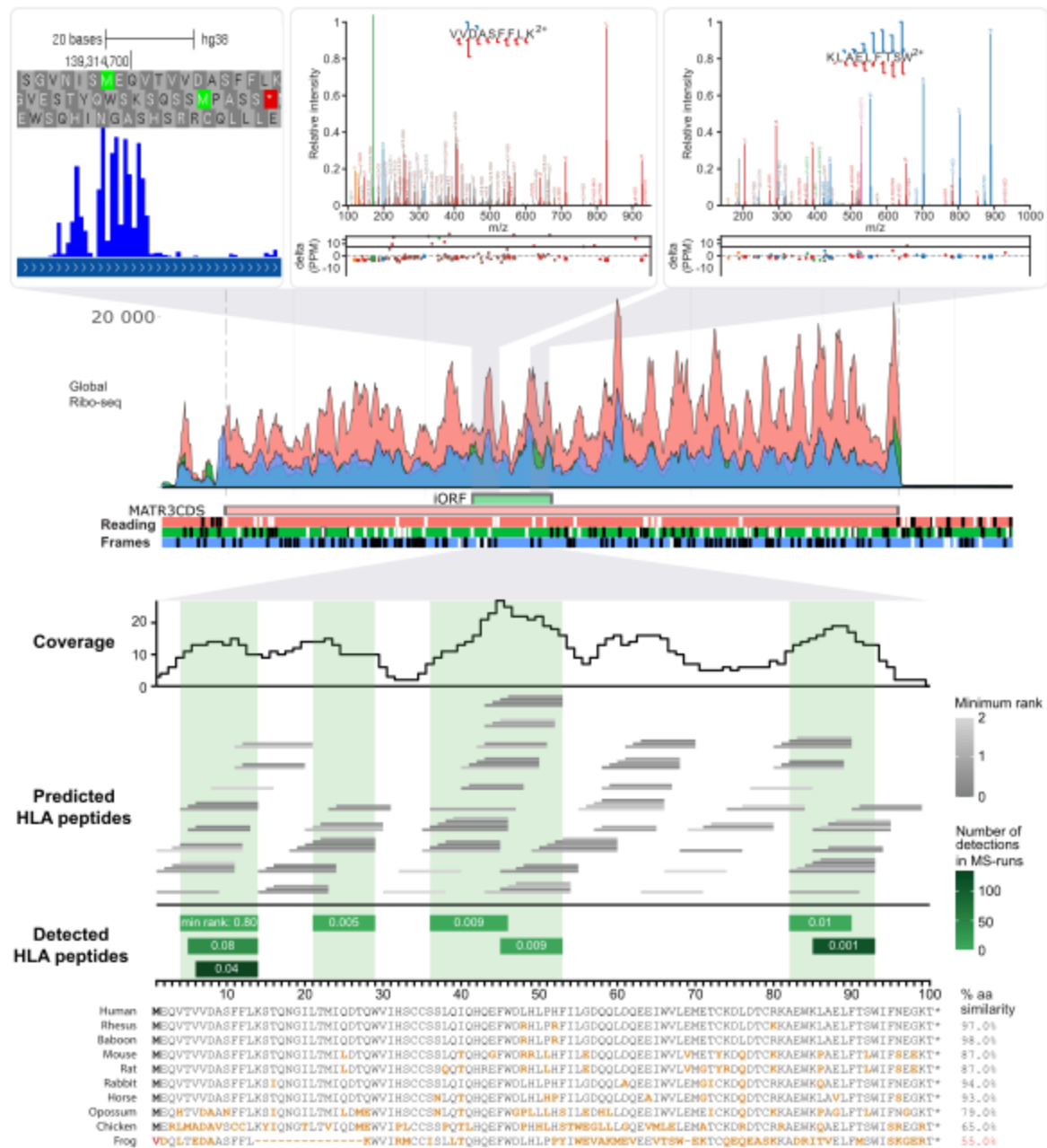
https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr12%3A11%2C121%2C140-11%2C121%2C180%2Bchr12%3A11%2C171%2C422-11%2C171%2C599&strand=-&prologue=6&epilogue=6&alnset=hg38_241mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c5norep142: novel peptidein

c5norep142 is an intORF in *MATR3*. The single non-HLA peptide is observed in cancer cell lines. It appears the translation is accessed by alternative splicing: transcriptomics analysis indicates that the exon containing the start of the *MATR3* CDS is commonly 'skipped' in mRNA, and the alternative frame intORF becomes the first plausible translation in transcripts where skipping happens (i.e., considering 5'-3' ribosome scanning). The ORF is conserved in almost all mammal genomes.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr5%3A139%2C314%2C694-139%2C314%2C736%2Bchr5%3A139%2C315%2C697-139%2C315%2C738%2Bchr5%3A139%2C316%2C076-139%2C316%2C188%2Bchr5%3A139%2C317%2C053-139%2C317%2C105%2Bchr5%3A139%2C317%2C596-139%2C317%2C647&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on



Overview of data available for c5norep142, an intORF in the *MATR3* gene. Ribo-seq data shows the initiation of translation at the methionine translation initiation codon (green), as determined by enrichment of ribosomes at the TIS. Two peptide spectral matches for HLA-I peptides VVDASFFLK and KLAELFTSW are shown having nearly complete sequence coverage (USIs are mzspect:PXD037270:Liv32_1176935F:scan:33690:VVDASFFLK/2 and mzspect:PXD011628:PBM009_msms37:scan:16281:KLAELFTSW/2, respectively). The lowest panel shows the position of all 8 peptides that were observed in the immunopeptidomics data. The color shading indicates the number of MS runs in which each peptide was observed. The middle panel shows all peptides that are predicted with NetMHCpan to be observable in the MS runs (i.e., they are predicted to bind with NetMHCpan score <2 to at least one allele in one of the

samples in which peptides were observed). The top part shows the number of predicted binding peptides in which each amino acid was located. Green shadings indicate which part of the ORF sequence was observed. Except for the region near the offset of 62, detected peptides occurred in the regions with the highest numbers of predicted binders.

Tier 1B, supported by 5+ HLA peptides

c10norep31: novel peptidein

c10norep31 is an intORF in *CDC13*. The initiation codon is 10bp downstream of the CDS ATG, and has clear Ribo-seq support in HHT datasets. It is not obvious how the gene utilises the downstream ATG instead of the CDS ATG. The ORF is conserved in primates; it seems to have originated at the based on placental mammals, with ATG loss in certain lineages.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr10%3A12%2C196%2C259-12%2C196%2C319%2Bchr10%3A12%2C198%2C705-12%2C198%2C733&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c10riboseqorf41: novel peptidein

c10riboseqorf41 is a uoORF in *TFAM*. The ORF is conserved in higher primates.

The ncORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr10%3A58%2C385%2C433-58%2C385%2C648%2Bchr10%3A58%2C386%2C220-58%2C386%2C303&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c11norep30: novel peptidein

c11norep30 is an intORF found within three coding exons of *EIF4G2*. The ATG is isolated immediately prior to a splice donor site, and considering also the lack of HHT support it could potentially be miscalculated. This ATG is almost perfectly conserved in placental mammals. However, the ORF given by using either of two alternative in-frame upstream ATGs (chr11:10,804,143-10,804,145 and chr11:10,804,167-10,804,169) is conserved far deeper into vertebrates, and this would not change the peptide mapping. Neither of these ATGs has HHT support either though. In terms of access to the translation, this has the potential to occur via skipping of the first codon exon of MANE Select model ENST00000339995, which is reasonably supported in RNAseq data. Thus, the ncORF as called has been annotated as a peptidein, although we consider its N-t structure uncertain.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr11%3A10%2C803%2C899-10%2C804%2C050%2Bchr11%3A10%2C804%2C137-10%2C804%2C139&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c11riboseqorf106: novel peptidein

c11riboseqorf106 is a uORF in *UCP2*. The initiation codon is found upstream of the TSS of MANE Select model ENSG00000175567; however, there is evidence for transcriptional extension in the 5' direction to accommodate this, and so the ORF can instead be translated on model ENST00000310473. In this context, it is notable that HHT support and Ribo-seq signal is stronger for the usage of a downstream initiation codon within the ORF (giving translation chr11:73,981,512-73,981,622), which would be the first ATG with the MANE Select model. However, there is also HHT support for the called ATG, albeit notably weaker. It therefore appears that the ORF exists in both the form as called, but more commonly in an N-t truncated form. One of the peptides is specific to the N-t extended form. The extended form has poor conservation, whereas the ORF from the downstream ATG is deeply conserved into vertebrates.

The ncORF in the 100 tetrapod alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr11%3A73%2C981%2C512-73%2C981%2C632%2Bchr11%3A73%2C982%2C721-73%2C983%2C034&strand=-&prologue=6&epilogue=6&alnset=hg38_100_tetrapod&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c11riboseqorf112: novel peptidein

c11riboseqorf112 is a uoORF in *EMSY*, spanning four exons. The ORF is almost perfectly conserved in mammalian genomes, as well as reptile and avian genomes.

The ncORF in the 100 tetrapod alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr11%3A76%2C445%2C106-76%2C445%2C128%2Bchr11%3A76%2C446%2C900-76%2C447%2C008%2Bchr11%3A76%2C451%2C858-76%2C451%2C957%2Bchr11%3A76%2C453%2C314-76%2C453%2C357&prologue=6&epilogue=6&alnset=hg38_100_tetrapod&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c12riboseqorf145: novel peptidein

c12riboseqorf145 is a uORF in *TAOK3*. The ORF is almost perfectly conserved in mammalian genomes.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr12%3A118%2C255%2C649-118%2C255%2C655%2Bchr12%3A118%2C266%2C655-118%2C266%2C698&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c12riboseqorf47: novel peptidein

c12riboseqorf47 is a uoORF in *TMEM106C*. The ATG is present in almost all mammalian genomes. The ORF is consistently open until downstream of the canonical *TMEM106C* initiation, after which there is some variability in the position of the ORF termination.

The ncORF in the 470 mammals Multiz alignment:

<https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr12%3A47%2C964%2C172->

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr12%3A49%2C566%2C198-49%2C566%2C215%2Bchr12%3A49%2C566%2C722-49%2C566%2C841%2Bchr12%3A49%2C568%2C048-49%2C568%2C050&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&codonPos=2&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c12riboseqorf52: novel peptidein

c12riboseqorf52 is a uoORF in *MCRS1*. The ORF is conserved in mammals.

The ncORF in the 100 vertebrate alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr12%3A49%2C566%2C198-49%2C566%2C215%2Bchr12%3A49%2C566%2C722-49%2C566%2C841%2Bchr12%3A49%2C568%2C048-49%2C568%2C050&strand=-&prologue=6&epilogue=6&alnset=hg38_100_tetrapod&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c14riboseqorf117: novel protein

c14riboseqorf117 is a uoORF in *EIF5*. It was already annotated as protein-coding by GENCODE (ENSG00000291313) following Mudge *et al*, due to the observation of strong evidence for protein-level constraint across the translation in vertebrate genomes.

c14riboseqorf73: novel peptidein

c14riboseqorf73 is a uoORF in *COX16*. The ORF is conserved in higher primates.

The ncORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr14%3A70%2C342%2C695-70%2C342%2C729%2Bchr14%3A70%2C359%2C519-70%2C359%2C603&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c15riboseqorf1: novel peptidein

c15riboseqorf1 is a uORF within the long, complex 5' UTR of *NIPA2*. It is found upstream of uORF c15riboseqorf2 on certain transcript models. The ORF is highly conserved in mammals, with a small number of species using a different termination codon.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr15%3A22%2C838%2C702-22%2C838%2C985&prologue=6&epilogue=6&alnset=hg38_470mammals&codonPos=2&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c15riboseqorf45: novel peptidein

c15riboseqorf45 is a uORF in *CIAO2A*. The ORF is almost perfectly conserved in mammalian genomes, and presents good evidence of conservation in reptile / avian genomes.

The ncORF in the 100 vertebrate alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr15%3A64%2C093%2C641-64%2C093%2C823&strand=-&prologue=6&epilogue=6&alnset=hg38_100_tetrapod&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c16norep103: novel peptidein

c16norep103 is an intORF in *NFATC3*. The initiation codon has support from stalled Ribo-seq-datasets, and there is an in-frame termination codon a short distance upstream in the same exon. The transcription event that would lead to expression of this ORF is not obvious, although it is potentially linked to the usage of alternative first exons that lack the initiation codon of the canonical CDS. The ORF is highly conserved in vertebrates, with disruptions in very few lineages.

The ncORF in the vertebrate alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr16%3A68%2C122%2C077-68%2C122%2C184&prologue=6&epilogue=6&alnset=hg38_100_tetrapod&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c16riboseqorf1: novel peptidein

c16riboseqorf1 is a uoORF in *SNRNP25* with respect to the MANE Select model (ENST00000293861), but it is an intORF with respect to an alternative model that contains an N-t extension to a non-ATG initiation codon (ENST00000710415). The ORF is conserved in primate genomes, with a few exceptions.

The ncORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr16%3A53%2C952-54%2C058%2Bchr16%3A55%2C459-55%2C489&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c16riboseqorf58: novel peptidein

c16riboseqorf58 is a uoORF in *SRCAP*. The ORF is conserved in vertebrates. Upon further analysis, *SRCAP* also presents Ribo-seq and evolutionary evidence for upstream initiation via a non-ATG codon, in frame with the known CDS and so giving an N-t extended isoform. The uoORF is an intORF with respect to this novel model (ENST00000973048).

The ncORF in the 100 vertebrate alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr16%3A30%2C700%2C709-30%2C700%2C878%2Bchr16%3A30%2C704%2C064-30%2C704%2C082&prologue=6&epilogue=6&alnset=hg38_100_tetrapod&wrap=50&spliceSite=s=human&hideInserts=on&hideJumps=on

c16riboseqorf91: novel peptidein

c16riboseqorf91 is one of three consecutive uORFs in the 5' UTR of *CES2*, and the most upstream. The ORF is conserved in ape genomes.

The ncORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr16%3A66%2C934%2C550-66%2C934%2C678&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c17norep146: novel peptidein

c17norep146 is uoORF in *PSMC5*. It is disrupted in several mammalian lineages.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr17%3A63%2C827%2C462-63%2C827%2C514%2Bchr17%3A63%2C828%2C138-63%2C828%2C209%2Bchr17%3A63%2C829%2C494-63%2C829%2C497&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c17norep98: novel peptidein

c17norep98 is a lncRNA ORF in *LINC00910*. It falls entirely within a HERV element. The ORF is conserved in ape and gibbon genomes.

The ncORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr17%3A43%2C371%2C394-43%2C371%2C426%2Bchr17%3A43%2C377%2C671-43%2C377%2C813%2Bchr17%3A43%2C379%2C123-43%2C379%2C262%2Bchr17%3A43%2C381%2C273-43%2C381%2C331&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c17riboseqorf58: novel peptidein

c17riboseqorf58 is an intORF of *KIAA0100*, since renamed *BLTP2*. The intORF has a very clear HHT peak. It is not obvious how this ORF is translated as opposed to the canonical CDS, as there is very little transcriptional diversity at the 5' end of the gene. The ORF is conserved in mammal genomes, with an earlier termination codon in marsupials.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr17%3A28%2C643%2C178-28%2C643%2C311%2Bchr17%3A28%2C643%2C610-28%2C643%2C659%2Bchr17%3A28%2C644%2C058-28%2C644%2C095&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c17riboseqorf78: novel peptidein

c17riboseqorf78 is a uORF in *IGFBP4*. The ORF is conserved in mammalian genomes.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr17%3A40%2C443%2C505-40%2C443%2C558&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c18norep5: novel peptidein

c18norep5 is a uoORF in *NDC80*. The ORF is a *de novo* emergence in primates, although it is disrupted in several lineages.

The ncORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr18%3A2%2C571%2C602-2%2C571%2C683%2Bchr18%3A2%2C572%2C977-2%2C573%2C062&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c18riboseqorf50: novel peptidein

c18riboseqorf50 is a dORF in *PMAIP1*. The ORF is found in higher primate genomes, with a few exceptions. It is not clear how the translation would be accessed by the ribosome. Nonetheless, Ribo-seq support is clear.

The ncORF in primate genomes in the 470 mammals Multiz alignment:

<https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr18%3A59%2C902%2C884->

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr19%3A17%2C101%2C693-17%2C102%2C163&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSite=s=human&hideInserts=on&hideJumps=on

c19riboseqorf40: novel peptidein

c19riboseqorf40 is a uoORF in *MYO9B*. The ATG appears to be ancestral in vertebrates, although it is lost in certain lineages such as avians. The position of the termination codon is variable once the initiation codon of *MYO9B* is bypassed.

The ncORF in the 100 vertebrate alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr19%3A17%2C101%2C693-17%2C102%2C163&prologue=6&epilogue=6&alnset=hg38_100_tetrapod&wrap=50&spliceSite=s=human&hideInserts=on&hideJumps=on

c19riboseqorf61: novel peptidein

c19riboseqorf61 is a lncRNA ORF in the first exon of ENSG00000267575. The ORF is a *de novo* emergence in apes.

The ncORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr19%3A27%2C793%2C510-27%2C793%2C647&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSite=s=human&hideInserts=on&hideJumps=on

c19riboseqorf66: novel peptidein

c19riboseqorf66 is a uORF in *URI1*. The ORF is conserved in mammalian genomes.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr19%3A29%2C942%2C370-29%2C942%2C564&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSite=s=human&hideInserts=on&hideJumps=on

c1norep182: novel peptidein

c1norep182 is a uORF in *IGSF3*. The evolutionary signature is complex; it seems most likely ancestral to placental mammals, but having been lost in numerous lineages.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr1%3A116%2C666%2C632-116%2C666%2C739&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c1norep256: novel peptidein

c1norep256 is an intORF in *PRRC2C*. The ORF is deeply conserved in vertebrates. Translational access to the intORF is possible via the skipping of the exon containing the *PRRC2C* initiation codon, i.e exon 2 of ENST00000647382, which has substantial support in RNAseq datasets.

The ncORF in the 100 vertebrate alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr1%3A171%2C513%2C007-171%2C513%2C111&prologue=6&epilogue=6&alnset=hg38_100_tetrapod&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c1riboseqorf177: novel peptidein

c1riboseqorf177 is a uORF in *PI4KB*. The ORF is conserved in higher primates.

The ncORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr1%3A151%2C326%2C176-151%2C326%2C343&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c1riboseqorf231: novel peptidein

c1riboseqorf231 is one of eight lncRNA ORFs in *GAS5*, which is a small nucleolar RNA host gene. The initiation codon is close to a splice acceptor site and lacks stalled ribosome support, and it could be that the translation initiates further upstream. The ORF is specific to higher primates.

The ncORF ORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr1%3A173%2C865%2C239-173%2C865%2C282%2Bchr1%3A173%2C866%2C528-173%2C866%2C567%2Bchr1%3A173%2C866%2C761-173%2C866%2C796%2Bchr1%3A173%2C866%2C991-173%2C867%2C041&strand=-&prologue=6&epilogue=6&alInset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c1riboseqorf248: novel peptidein

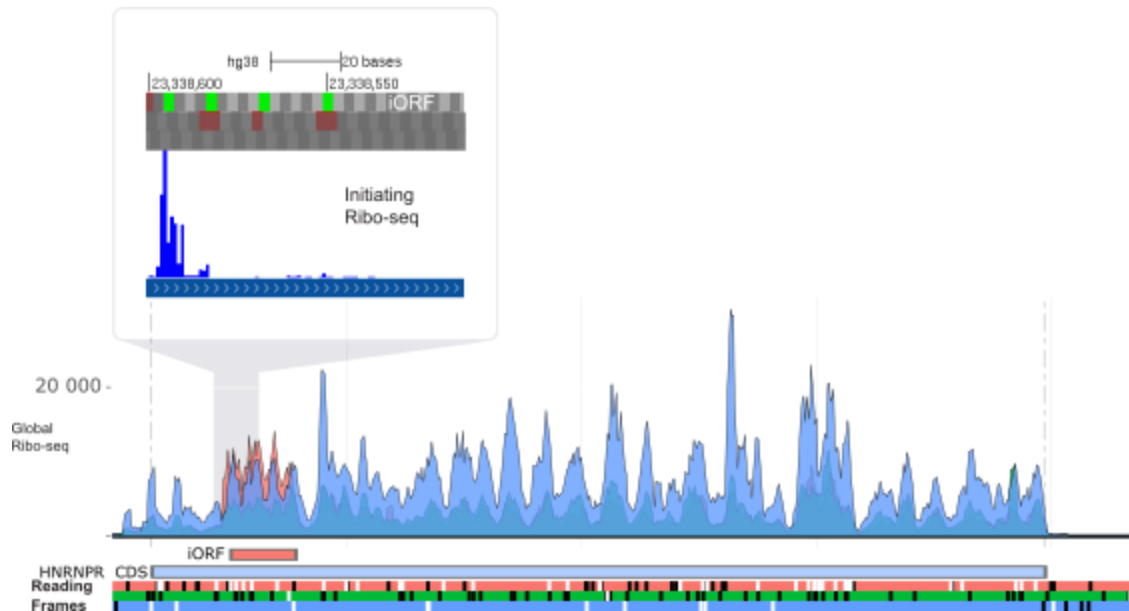
c1riboseqorf248 is a uORF in *IVNS1ABP*. The ORF is conserved in mammals.

The ncORF in the 100 vertebrate alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr1%3A185%2C311%2C281-185%2C311%2C322%2Bchr1%3A185%2C316%2C953-185%2C316%2C985&strand=-&prologue=6&epilogue=6&alInset=hg38_100_tetrapod&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c1riboseqorf39: novel peptidein

c1riboseqorf39 is an intORF within the CDS of *HNRNPR*, at least on MANE Select transcript ENST00000302271; the first coding exon of this model is commonly skipped due to alternative splicing, in which case the initiation codon of the ncORF (which has clear support from HHT data; see below) could also be considered as an uoORF with respect to models such as ENST00000476451. The ncORF is perfectly conserved in mammalian genomes, and also exhibits strong conservation in avian / reptile genomes. Of note, the initiation codon of the ncORF is disrupted by ClinVar variant NM_005826.5(*HNRNPR*):c.170A>T (p.Tyr57Phe), which is extremely rare in the gnomAD dataset (allele frequency 7.03987e-07) and linked to unspecified inborn genetic disease. However, the variant is also missense with respect to the *HNRNPR* canonical CDS.



The ncORF in the 100 vertebrate alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr1%3A23%2C337%2C831-23%2C337%2C861%2Bchr1%3A23%2C338%2C490-23%2C338%2C596&strand=-&prologue=6&epilogue=6&alnset=hg38_100_tetrapod&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c1riboseqorf55: known protein

c1riboseqorf55 is a uORF in *PTP4A2*. GENCODE previously annotated this ncORF as protein-coding based on PhyloCSF support. In fact, it is a shorter form of the translation that was picked up, chr1:31,919,563-31,919,628 (-), which is clearly conserved and constrained in vertebrates. c1riboseqorf55 is an N-t extended form, with the extension being conserved in mammals.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr1%3A31%2C919%2C563-31%2C919%2C658%2Bchr1%3A31%2C937%2C987-31%2C938%2C010&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c20norep82: novel peptidein

c20norep82 is an intORF in *ADNP*. The initiation codon is supported by stalled Ribo-seq data, and almost perfectly conserved in vertebrate genomes. There is some movement of the termination codon. Access to the translation is possible by skipping the exon which contains the initiation codon of the canonical CDS.

The ncORF in the 100 vertebrate alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr20%3A50%2C894%2C233-50%2C894%2C430&strand=-&prologue=6&epilogue=6&alnset=hg38_100_tetrapod&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c21riboseqorf14: novel peptidein

Not done c21riboseqorf14 is a uORF in *RUNX1*, not found on the MANE Select. The ORF is conserved in mammalian genomes, with the exception of a few lineages that have lost the initiation codon and some localised termination codon movement. There are three other Ribo-seq ORFs in this 5' UTR.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr21%3A34%2C887%2C800-34%2C887%2C862&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c22norep57: novel peptidein

c22norep57 is a uORF in *MFNG*. The ORF is conserved in primate genomes.

The ncORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr22%3A37%2C486%2C244-37%2C486%2C345&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c2riboseqorf155: novel peptidein

c2riboseqorf155 is a uORF in *BMPT2*. Interestingly, another uORF further downstream in the same 5' UTR was already annotated as protein-coding due to clear PhyloCSF support

(ENSG00000289490). The ncORF is younger, originating *de novo* at the base of the mammalian radiation, with the initiation codon being lost in certain lineages.

The ncORF in the 100 vertebrate alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr2%3A202%2C376%2C869-202%2C376%2C934&prologue=6&epilogue=6&alnset=hg38_100_tetrapod&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c2riboseqorf55: novel peptidein

c2riboseqorf55 is a uORF in *WBP1*. The ORF is conserved in mammalian genomes. It can be annotated as a peptidein due to proteomics evidence.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr2%3A74%2C458%2C484-74%2C458%2C594&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c3norep241: novel peptidein

c3norep241 is an intORF in *P3H2*. There is no clear HHT support, although the Ribo-seq signal is visible in the correct frame. The initiation codon is dual frame with respect to the first coding exon of the CDS, and it is not obvious how it is accessed by the ribosome. The ORF is found in all primate genomes, with a couple of exceptions.

The ncORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr3%3A189%2C995%2C313-189%2C995%2C442%2Bchr3%3A190%2C120%2C252-190%2C120%2C313&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c3riboseqorf1: novel peptidein

c3riboseqorf1 is a uORF in *THUMPD3*. The ORF is conserved in higher primate genomes.

The ncORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr3%3A9%2C363%2C424-9%2C363%2C525&prologue=6&epilogue=6&alnset=hg38_470mammals_prime&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c3riboseqorf154: novel peptidein

c3riboseqorf154 is a uORF in *MBNL1*. Its evolutionary history is fairly complex. The ATG is ancient, found at the base of vertebrates, yet it is lost in several mammal species and lineages, and the STOP has similarly imperfect conservation.

The ncORF in the 100 vertebrate alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr3%3A152%2C269%2C026-152%2C269%2C076&prologue=6&epilogue=6&alnset=hg38_100_tetrapod&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c3riboseqorf51: novel peptidein

c3riboseqorf51 is a uORF in *TCA1M*. The initiation codon is found in most mammalian genomes, although the position of the termination codon is variable.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr3%3A44%2C354%2C745-44%2C354%2C798&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c3riboseqorf98: known protein

c3riboseqorf98 is a uORF in *CGGBP1*. GENCODE already chose to annotate an N-t truncated form of this ORF as protein-coding due to its strong PhyloCSF score (ENSG00000288654). This shorter form has clear HHT data in support for initiation. Nonetheless, the extended form of the ORF as called is also supported by Ribo-seq data, and two of the six peptides map to what would be the N-t extension region. However, in contrast to the annotated protein, this extended form would be specific to higher primates.

The ncORF in primate genomes in the 470 mammals Multiz alignment:

<https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr3%3A88%2C058%2C139-88%2C058%2C223%2Bchr3%3A88%2C058%2C815-88%2C059%2C047&strand=->

&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c4riboseorf2: novel peptidein

c4riboseorf2 is a uORF in *CTBP1*, found entirely within LINE sequence. The ORF is perfectly conserved in primate and rodent genomes. The initiation codon is older, being found in most mammals.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr4%3A1%2C241%2C335-1%2C241%2C433&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c4riboseorf30: novel peptidein

c4riboseorf30 is a uORF in *SLC30A9*. The ncORF is a de novo emergence in ape genomes.

The ncORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr4%3A41%2C990%2C504-41%2C990%2C632&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c4riboseorf57: novel peptidein

c4riboseorf57 is a uORF in *SEC31A*. The ncORF is conserved in higher primate genomes.

The ncORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr4%3A82%2C900%2C167-82%2C900%2C223%2Bchr4%3A82%2C900%2C334-82%2C900%2C435&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c4riboseqorf60: novel peptidein

c4riboseqorf60 is a uORF in *PTPN13*. The ORF is conserved in primate genomes. The ATG is conserved in rodents and lagomorphs. A second uORF (c4riboseqorf61) is found a short distance downstream.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr4%3A86%2C594%2C356-86%2C594%2C451&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c4riboseqorf75: novel peptidein

c4riboseqorf75 is a uoORF in *GSTCD*. Most mammal genomes have the initiation codon, although the ORF is poorly conserved beyond apes. The ncORF has a substantial overlap with c4riboseqorf74 in an alternative frame.

The ncORF ORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr4%3A105%2C708%2C818-105%2C709%2C016%2Bchr4%3A105%2C717%2C593-105%2C717%2C738&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c4riboseqorf83: novel peptidein

c4riboseqorf83 is a lncRNA ORF in *SNHG8*, non-overlapping with c4riboseqorf84. While the sequence region is conserved in mammals the ORF is primate specific, and so would be a *de novo* emergence.

The ncORF in primate and rodent genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr4%3A118%2C278%2C820-118%2C278%2C978&prologue=6&epilogue=6&alnset=hg38_470mammals_supraprimate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c4riboseqorf84: novel peptidein

c4riboseqorf84 is a lncRNA ORF in *SNHG8*, non-overlapping with c4riboseqorf83. The initiation codon is found in higher primate genomes, although the position of the termination codon is variable.

The ncORF in primate and rodent genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr4%3A118%2C279%2C412-118%2C279%2C431%2Bchr4%3A118%2C279%2C624-118%2C279%2C714&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c5riboseqorf128: novel peptidein

c5riboseqorf128 is a uORF in *PRELID1*. The initiation codon is conserved in mammals, although the alanine homopolymer is variable in length and the location of the termination codon is also inconsistent.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr5%3A177%2C303%2C816-177%2C303%2C869&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c5riboseqorf144: novel peptidein

c5riboseqorf144 was called as a uoORF in *ZFP62* with respect to models including ENST00000512132, but it is an intORF in the MANE Select model ENST00000502412. The ORF is conserved in higher primates.

The ncORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr5%3A180%2C851%2C347-180%2C851%2C466&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c5riboseqorf22: novel peptidein

c5riboseqorf22 is a lncRNA ORF in *NNT-AS1*. The ORF is specific to ape genomes, and incorporates MER39B and MSTB transposon sequence.

The ncORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr5%3A43%2C575%2C896-43%2C575%2C935%2Bchr5%3A43%2C583%2C467-43%2C583%2C566%2Bchr5%3A43%2C602%2C778-43%2C602%2C814&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c5riboseqorf27: novel peptidein

c5riboseqorf27 is a lncRNA ORF in AC008966.1, since named as *MOCS2-DT*. Inspection of the data indicates that the ORF may be more commonly translated via an [ATT] codon several 18bp upstream, although there is also some HHT support for the [ATG] as called. The ATG is found in most primate genomes, but the ORF is human specific.

The ncORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr5%3A53%2C109%2C930-53%2C110%2C148%2Bchr5%3A53%2C112%2C240-53%2C112%2C278&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c5riboseqorf9: novel peptidein

c5riboseqorf9 is a lncRNA ORF in *MIR4458HG*. Two other ncORFs were called by Ribo-seq, and c5riboseqorf7 shares an out of frame overlap. The ORF is a de novo emergence in apes.

The ncORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr5%3A8%2C459%2C958-8%2C460%2C062&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c5riboseqorf95: novel peptidein

c5riboseqorf95 is a uORF in *FAM53C*. It could also be called as an uoORF with respect to alternative isoform ENST00000511276. The ORF is conserved in mammals.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr5%3A138%2C341%2C195-138%2C341%2C263&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c6norep97: novel peptidein

c6norep97 is a lncRNA ORF in *PSMB8-AS1*. Most of its sequence is antisense to CDS of *TAP1*. The ORF is conserved in higher primate genomes.

The ncORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr6%3A32%2C845%2C552-32%2C845%2C743&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c6riboseqorf128: novel peptidein

c6riboseqorf128 is a uORF in *BCLAF1*. The ATG is almost perfectly conserved in therian mammal genomes, although the position of the termination codon is variable.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr6%3A136%2C289%2C742-136%2C289%2C801&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c6riboseqorf99: novel peptidein

c6riboseqorf99 is a uoORF in *CCNC*. The ORF is conserved in mammals.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr6%3A99%2C562%2C928-99%2C562%2C948%2Bchr6%3A99%2C568%2C496-99%2C568%2C588&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c7norep125: novel peptidein

c7norep125 is a uORF in *SLC25A40*. The ATG is almost perfectly conserved in mammals with the exception of rodents, although the ORF is otherwise specific to apes and gibbon due to the presence of a second exon of three that is embedded within an Alu element.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr7%3A87%2C858%2C740-87%2C858%2C751%2Bchr7%3A87%2C860%2C072-87%2C860%2C169%2Bchr7%3A87%2C860%2C572-87%2C860%2C626&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c7norep236: novel peptidein

c7norep236 is a lncRNA ORF in ENSG00000216895. Exon 1 has a partial overlap with LINE sequence. The ORF is a *de novo* emergence in apes.

The ncORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr7%3A155%2C611%2C300-155%2C611%2C417%2Bchr7%3A155%2C643%2C468-155%2C643%2C568&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c7riboseqorf115: novel peptidein

c7riboseqorf115 is a uORF in *TMEM168*. The ORF has almost perfect conservation in mammal genomes, with the exception of termination codon movement in a small number of lineages.

The ORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr7%3A112%2C790%2C313-112%2C790%2C372&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c7riboseqorf131: novel peptidein

c7riboseqorf131 is a dORF in *CYREN*. The ORF is conserved only in apes, although the ATG is found in most primate genomes. Stalled Ribo-seq data supports in the initiation codon, although it is not obvious how the translation is accessed.

The ncORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr7%3A135%2C166%2C342-135%2C166%2C521&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c7riboseqorf18: novel peptidein

c7riboseqorf18 is a uORF in *SNX13*. The ORF is perfectly conserved in mammals.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr7%3A17%2C940%2C325-17%2C940%2C483&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c7riboseqorf64: novel peptidein

c7riboseqorf64 is a uoORF in *TMEM60*. The ORF is conserved in mammalian genomes, although a few lineages have a different termination codon.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr7%3A77%2C794%2C288-77%2C794%2C423%2Bchr7%3A77%2C798%2C254-77%2C798%2C348&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c8norep107: novel peptidein

c8norep107 is a lncRNA ORF in *ZFPM2-AS1*. The ORF is human specific.

The ncORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr8%3A105%2C787%2C547-105%2C787%2C596%2Bchr8%3A105%2C798%2C325-105%2C798%2C442%2Bchr8%3A105%2C927%2C651-105%2C927%2C717%2Bchr8%3A106%2C060%2C381-106%2C060%2C442&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c8riboseqorf104: novel peptidein

c8riboseqorf104 was called as a lncRNA ORF in *CASC19*, although GENCODE now consider it to be part of the same gene as *CCAT1* and *PCAT1*, thus making a much larger and more complex lncRNA locus. It contains eight ncORFs. This ORF is conserved in higher primates.

The ncORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr8%3A127%2C209%2C393-127%2C209%2C602&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c8riboseqorf12: novel peptidein

c8riboseqorf12 is a uoORF in *NEIL2*. The ATF is conserved in primate genomes with a few exceptions, although the termination codon is variable and the ORF is specific to higher primates.

The ncORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr8%3A11%2C770%2C235-11%2C770%2C335%2Bchr8%3A11%2C771%2C446-11%2C771%2C536&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c8riboseqorf67: novel peptidein

C8riboseqorf67 is a uoORF in *PEX2*, one of four that each share an overlap. The ORF is conserved in ape genomes.

The ncORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr8%3A76%2C984%2C108-76%2C984%2C195%2Bchr8%3A76%2C986%2C187-76%2C986%2C284&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c9riboseqorf24: novel peptidein

c9riboseqorf24 was called on a processed transcript in *CREB3*. It is better understood as a uoORF with respect to models including ENST00000881110; it cannot be mapped to the MANE Select model ENST00000353704, which has been set to use a downstream TSS. The ORF is conserved in ape / gibbon.

The ncORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr9%3A35%2C732%2C636-35%2C732%2C824&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c9riboseqorf81: novel peptidein

c9riboseqorf81 is a uORF in *RABEPK*. The ATG is conserved in primate genomes, and the ORF in higher primates.

The ncORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr9%3A125%2C200%2C795-125%2C200%2C860&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

cXriboseqorf30: novel peptidein

cXriboseqorf30 is a uoORF in *FAM104B*, since renamed *VCF2*. The ORF is conserved in primate genomes. It has a minor splicing difference compared to cXriboseqorf31.

The ncORF in the primate 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chrX%3A55%2C146%2C233-55%2C146%2C311%2BchrX%3A55%2C159%2C129-55%2C159%2C228%2BchrX%3A55%2C161%2C137-55%2C161%2C176&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

cXriboseqorf44: novel peptidein

cXriboseqorf44 is a lncRNA in *JPX*, with cXriboseqorf45 representing an alternative spliceform. The initiation codon has almost perfect conservation in therian mammals, but the ORF is highly dissimilar between species and only fully conserved in apes.

The ncORF in the primate 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chrX%3A73%2C944%2C354-73%2C944%2C455&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

cXriboseqorf59: novel peptidein

cXriboseqorf59 is a uORF in *MORF4L2*. The ORF is highly conserved in therian mammals, aside from the fact that the termination codon often shows highly localized variability.

The ncORF in the 470 mammals Multiz alignment:

<https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chrX%3A103%2C685%2C228-103%2C685%2C260%2BchrX%3A103%2C686%2C631-103%2C686%2C696&strand=->

&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c14norep128: novel isoform

c14norep128 is a dORF in *TNFAIP2*. The ORF was initially called in Tier 1A, but was changed to Tier 1B after the non-HLA peptides in support were found to be inadmissible as evidence. The ORF substantially overlaps with an alternative coding isoform of the gene, and while there is clear Ribo-seq signal over this region the initiation codon for c14norep128 cannot be resolved with confidence. As such, it was decided not to annotate this ORF as a peptidein at the present time.

c16riboseqorf104: novel peptidein

c16riboseqorf104 is a uORF in *PSKH1*. The ORF was initially called in Tier 1A, but was changed to Tier 1B after the non-HLA peptides in support were found to be inadmissible as evidence. Three of the HLA peptides are alanine rich, though judged to be of excellent quality. The ORF is highly conserved in therian mammal genomes, with length variation in the alanine repeat region.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr16%3A67%2C893%2C276-67%2C893%2C371%2Bchr16%3A67%2C908%2C680-67%2C908%2C685&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

ncORFs nominated for inspection as peptideins/proteins based upon CRISPR lethality screen data

c7riboseqorf76

c7riboseqorf76 is a uORF in *FZD1*. The ORF is conserved in therian mammal genomes, and the initiation codon is conserved across to coelacanth. The ORF has a 13 residue arginine homopolymer at the N-t. Arginine stretches at the N-t can be nuclear localisation signals, but this

is unusually long and so would be expected to be highly potent. However, arginine tracts are also used in ribosome stalling, which could point to alternative function if the ORF is instead a regulatory translation.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr7%3A91%2C264%2C487-91%2C264%2C588&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c22riboseqorf37

c22riboseqorf37 is a uORF in *RBFOX2*, found in an alternative first coding exon compared to the MANE Select model. Translation of the ORF has the potential to be controlled via differentiation TSS usage. The ORF is conserved in mammal genomes.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr22%3A35%2C840%2C369-35%2C840%2C443&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c4norep181

c4norep181 is a uORF in *RAPGEF2*, in the very large 5' UTR. The ATG and termination codon are largely conserved in mammalian genomes, with length variability in the glycine homopolymer at the N-t.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr4%3A159%2C103%2C385-159%2C103%2C507&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c13riboseqorf46

c13riboseqorf46 is a uORF in *RAP2A*. The ORF is disrupted in rodent and lagomorph genomes, but otherwise is conserved in mammals. Positive PhyloCSF signal in places is potentially explained by alanine-rich content.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr13%3A97%2C434%2C282-97%2C434%2C425&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c2norep171

c2norep171 is a uORF in *GALNT13*. Entirely found within the LTR of an ancient ERV, so not incorporating a component ORF of the ERV. The insertion occurred and was fixed near the base of mammals, and the ORF is conserved in mammal genomes.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr2%3A153%2C872%2C048-153%2C872%2C113&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c7norep93

c7norep93 is a uORF in *PSPH*, which is replete with ClinVar variants. Found upstream of c7norep92 in the same 5' UTR exon. The ORF seems most likely to be a *de novo* emergence in higher primates, although there is an ATG in the region in several other mammalian lineages.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr7%3A56%2C051%2C252-56%2C051%2C320&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c1riboseqorf249

c1riboseqorf249 is a uORF in *CAMSAP2*. The ORF seems to have emerged at the base of the primate / rodent / lagomorph clade, although in fact the C-t and termination codon is conserved in mammals.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr1%3A200%2C739%2C611-200%2C739%2C682&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c11riboseqorf154

c11riboseqorf154 is a uORF in *ZNF202*. The evolutionary picture is not straightforward. The ORF is found in most primate genomes, and certain other mammalian lineages as well. Overall though it is disrupted in most species.

The ncORF using primate-only genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr11%3A123%2C740%2C439-123%2C740%2C498&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c10riboseqorf97

c10riboseqorf97 is a uORF in *SFXN3*. Evolutionary evidence suggests the ORF most likely evolved at the base of mammals, before being lost in certain lineages.

The ncORF in the 470 mammals Multiz alignment:

<https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr10%3A101%2C032%2C405->

101%2C032%2C458&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c2riboseqorf47: novel protein

c2riboseqorf47 is a uORF in *GMCL1*. Now annotated as protein-coding after considering also the evolutionary picture, where PhyloCSF supports protein-coding potential in placental mammals. The protein-coding gene ID is ENSG00000310604.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr2%3A69%2C829%2C739-69%2C829%2C798&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c1riboseqorf161

c1riboseqorf161 is a uORF in *NBPF15*. This gene has many highly similar paralogs within segmental duplications on chromosome 1, at least some of which also contain a version of the uORF (the T2T genome likely contains additional and distinct copies, i.e., the locus has copy number variation). The family is understood as primate specific. This copy presents as conserved in apes and gibbon, although the underlying genome alignments are potentially not true 1:1 orthologies.

The ncORF using primate-only genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr1%3A144%2C461%2C549-144%2C461%2C617&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c14riboseqorf118: novel peptidein

c14riboseqorf118 is a uORF in *MARK3*. The ORF is conserved in mammalian genomes with few exceptions. It can be annotated as a novel peptidein due to peptide evidence.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr14%3A103%2C385%2C546-103%2C385%2C662&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c19riboseqorf187

c19riboseqorf187 is a uORF in *MZF1*. The 5' UTR is highly complex, with substantial alternative splicing. RefSeq model XM_011527264.4 is a computational prediction with a splice donor site midway through the ORF, which splices into the MANE Select first coding exon. There is good long-read and short-read support for this splice junction, so it can also be annotated by GENCODE. Ribo-seq indicates that the single exon ORF form as called does exist, but it is possible the splicing form exists as well, whereby the translation would be understood as an alternative isoform of *MZF1*. The shared initiation codon is specific to higher primates, and the ORF as called is intact in higher primate genomes. Meanwhile, the splice donor site does not display mammalian conservation.

The ncORF using primate-only genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr19%3A58%2C573%2C195-58%2C573%2C296&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c19riboseqorf128

c19riboseqorf128 is a uORF in *FKRP*. The 5' UTR is highly complex. There are ClinVar variants on the initiation codon. One has conflicting interpretations, though has been considered pathogenic for a form of MD (and it's absent in GnomAD). <https://www.ncbi.nlm.nih.gov/clinvar/RCV000626047.3/>. *FKRP* CDS mutations have been associated with MD (and other things). Also, this variant disrupts the termination codon: <https://www.ncbi.nlm.nih.gov/clinvar/variation/391158/> (no associated information though; absent in GnomAD). There are a few additional ClinVar variants that are missense within the ORF (if it is protein-coding). The initiation codon is conserved in mammals, and the ORF largely so although with some termination codon movement in certain lineages.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr19%3A46%2C746%2C069-46%2C746%2C090%2Bchr19%3A46%2C748%2C027-46%2C748%2C085&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c16norep126

c16norep126 is a uORF in *USP10*. Three *USP10*-like pseudogenes are found in the genome, each incorporating the uORF sequence though with sequence mismatches. The initiation codon appears to be ancestral to mammals, though it is lost in certain lineages. Similarly, the ORF is intact across most of the phylogeny, though with termination codon movement and indels in certain species.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr16%3A84%2C699%2C996-84%2C700%2C079&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c17riboseqorf64

c17riboseqorf64 is a uORF in *RFFL*. It is found in an alternative 5' UTR exon, not in the MANE Select model. The ORF is conserved in mammalian genomes, with few exceptions.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr17%3A35%2C089%2C180-35%2C089%2C254&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c1riboseqorf270

c1riboseqorf270 is a uORF in *VASH2*. The initiation codon is conserved in mammals and bird / reptile genomes. The ORF is largely intact also, although with substantial variation in the length of the alanine-rich region and some localised termination codon movement.

The ORF in the 100 vertebrate alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr1%3A212%2C950%2C602-212%2C950%2C724&prologue=6&epilogue=6&alnset=hg38_100_tetrapod&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c2riboseqorf95

c2riboseqorf95 is a uORF in *R3HDM1*. The ORF is almost perfectly conserved in mammalian genomes.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr2%3A135%2C531%2C530-135%2C531%2C628&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c12riboseqorf127

c12riboseqorf127 is a uORF in *PHETA1*. The ORF is intact in mammalian genomes, with few exceptions.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr12%3A111%2C368%2C926-111%2C368%2C976&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c6riboseqorf67

c6riboseqorf67 is a dORF in *TOMM6*. No evidence of stalled ribosome support for the initiation codon, while usage of the splice acceptor site at chr6:41,789,568-41,789,568 (first bp of the exon) in combination with the skip of coding exon 2 would give an alternative *TOMM6* isoform that

incorporates most of the ORF sequence. There is long-read support for this transcript, so it can be annotated by GENCODE. The ORF stretch prior to the splice acceptor does have support in Ribo-seq data, but there does seem to be an increase in support to its 5'. The evolutionary picture is hard to read. The ORF as called is conserved in higher primates, and although the ATG is seen in some other mammalian lineages, the C-t is poorly conserved.

The ORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr6%3A41%2C789%2C536-41%2C789%2C655&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c17norep95

c17norep95 is a uORF in *IFI35*. The ATG is potentially ancestral with respect to mammals though apparently lost in numerous lineages, while the termination codon is only found in certain ape genomes. The ATG is not found in the MANE Select, being a few base pairs upstream of the TSS called for the model. It is transcribed though, which suggests the potential for expression modulation via TSS switching.

The ORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr17%3A43%2C006%2C771-43%2C006%2C839&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c12riboseqorf77

c12riboseqorf77 is a uORF in *ZBTB39*. The ATG is conserved in mammalian genomes, with variation in the position of the termination codon in certain lineages.

The ncORF in the 470 mammals Multiz alignment:

<https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr12%3A57%2C004%2C929-57%2C004%2C961%2Bchr12%3A57%2C006%2C405-57%2C006%2C431&strand=->

&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInser
ts=on&hideJumps=on

c1riboseqorf169

c1riboseqorf169 is a uORF in *ANP32E*. The ORF is a *de novo* emergence, likely in apes and gibbon.

The ORF in primate genomes in the Cactus alignment:

[https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr1%3A150%
2C235%2C977-150%2C236%2C063&strand=-
&prologue=6&epilogue=6&alnset=hg38_241mammals&wrap=50&spliceSites=human&hideInser
ts=on&hideJumps=on](https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr1%3A150%2C235%2C977-150%2C236%2C063&strand=-&prologue=6&epilogue=6&alnset=hg38_241mammals&wrap=50&spliceSites=human&hideInser
ts=on&hideJumps=on)

c16riboseqorf67

c16riboseqorf67 is a uORF in *NETO2*. The ORF is conserved in mammalian genomes, with the exception of rodents and lemur species. It is found entirely within ERV LTR sequence, which is therefore an ancient insertion event.

The ncORF in the 470 mammals Multiz alignment:

[https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr16%3A47%
2C143%2C755-47%2C143%2C856&strand=-
&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInser
ts=on&hideJumps=on](https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr16%3A47%2C143%2C755-47%2C143%2C856&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInser
ts=on&hideJumps=on)

c17riboseqorf93

c17riboseqorf93 is a lncRNA ORF in ENSG00000267002. It is found entirely within ERV sequence (though apparently does not represent an ERV component ORF), a very large insertion that happened at the base of apes / gibbon. All related sequences in the human genome are less than 90% similar at the DNA level over the ORF, and none represent an equivalent ORF. The lncRNA gene is transcribed antisense from the *NBR1* promoter, and has high, general expression. Its first exon is also used as an alternative first 5' UTR exon of *BRCA1* downstream with

reasonable RNAseq support, although the ORF is not present on any transcript reads that also incorporate *BRCA1* coding exons.

The ORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr17%3A43%2C167%2C100-43%2C167%2C213&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c3riboseqorf106: novel peptidein

c3riboseqorf106 is a lncRNA ORF in *ZBTB11-AS1* / ENSG00000256628, and the ORF is fully antisense to *ZBTB11* CDS. Substantial Ribo-seq data is clearly present on the correct strand, including stalled support for the initiation codon. The lncRNA itself has high, general expression. The *ZBTB11* CDS is found in all vertebrates, which complicates evolutionary analysis of the ORF. The ATG is found in most mammalian genomes, and the ORF is typically intact. Losses do not follow an obvious phylogenetic pattern. It can be annotated as a novel peptidein due to peptide evidence.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr3%3A101%2C676%2C645-101%2C676%2C911&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c17riboseqorf52

c17riboseqorf52 is a lncRNA ORF in *CCDC144NL-AS1* / ENSG00000233098, embedded within the REP522 sequence. This is understood as a telomeric repeat, but here it is found in the pericentromeric region. Most other genomic sequences are highly divergent, although sequences at chr17:21515813-21515892 and chr7:57639349-57639410 are ~90% similar at the DNA level. Evolutionary interpretation is complicated by the fact that multispecies alignments are unlikely to be accurate. The repeat itself is understood as primate specific, while an intact ORF does not seem to be an ancestral or typical component.

The ORF in primate genomes in the 470 mammals Multiz alignment:

<https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr17%3A20%2C868%2C498->

20%2C868%2C578&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c10norep118

c10norep118 is a lncRNA ORF in *ADIRF-AS1* / ENSG00000272734. It is entirely embedded within an ERV sequence, but does not seem to be an ERV component ORF. There are multiple related sequences in the human genome, but all below 90% sequence identity at the DNA level. The insertion itself is ancient, having occurred at the base of the primate radiation.

The ORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr10%3A86%2C965%2C862-86%2C966%2C035&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals_primate&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c10riboseqorf92: novel peptidein

c10riboseqorf92 is a lncRNA ORF in *OLMALINC* / ENSG00000235823. Exon 1 of 2 overlaps with c10norep140 and c10norep141, while exon 2 is unique to the Ribo-seq ORF catalog and embedded within the ERV sequence. Exon 1 represents *de novo* emergence in higher primates, while the ERV insertion happened at the base of new world monkeys and old world monkeys. The full length ORF is only intact in ape and gibbon genomes. *OLMALINC* itself has a high, general expression. It can be annotated as a novel peptidein due to peptide evidence.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr10%3A100%2C373%2C947-100%2C374%2C012%2Bchr10%3A100%2C381%2C216-100%2C381%2C452&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c6norep15: novel peptidein

c6norep15 is a lncRNA ORF in *LYRM4-AS1* / ENSG00000272142. Exon 1 of 2 is antisense to *RPP40* CDS, and exon 2 is embedded within Alu sequence. The conservation of exon 1 is hard to assess due to the *RPP40* complexity, while exon 2 - and the ORF itself - to the old world monkey / ape clade. It can be annotated as a novel peptidein due to peptide evidence.

The ncORF in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr6%3A5%2C003%2C910-5%2C003%2C959%2Bchr6%3A5%2C021%2C381-5%2C021%2C435&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c9norep47

c9norep47 is a lncRNA ORF in *GLIDR* / ENSG00000278175. Exon 1 of 3 has a partial MIR element overlap. The exon 1-2 region is duplicated as 7 copies on chromosome 9 pericentromeric regions, at ~95% sequence identity, including as part of lncRNAs *FAM88C*, *LERFS*, *FGF7P3*, *FAM88E*, *FAM88F* and *ENSG00000291170*; in effect, these lncRNA would seem to be part of a gene family. *GLIDR* itself has high, general expression. The sequence region is primate specific, although the multispecies alignments are unlikely to be accurate so the evolution of the ORF is difficult to judge beyond that. Intact copies do seem to exist in monkey genomes.

The ORF in primate genomes in the 470 mammals Multiz alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr9%3A39%2C808%2C741-39%2C808%2C761%2Bchr9%3A39%2C809%2C486-39%2C809%2C615%2Bchr9%3A39%2C809%2C731-39%2C809%2C855&strand=-&prologue=6&epilogue=6&alnset=hg38_470mammals&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

c2riboseqorf84

c2riboseqorf84 is a lncRNA ORF in *NIFK-AS1* / ENSG00000236859. Both exons are found within the same Alu element, which seems to have inserted in higher primates. There are thus numerous related sequences in the genome, although none are full length with higher than 92% sequence similarity at the DNA level, and none appear to incorporate related ORFs. The ORF is conserved in apes and old world monkeys, with the exception of gibbon.

The ORF in primate genomes in the Cactus alignment:

https://data.broadinstitute.org/compbio1/cav.php?controlsState=show&intervals=chr2%3A121%2C650%2C327-121%2C650%2C381%2Bchr2%3A121%2C650%2C468-121%2C650%2C511&prologue=6&epilogue=6&alnset=hg38_241mammals_prime&wrap=50&spliceSites=human&hideInserts=on&hideJumps=on

(2) Machine learning-based prediction of ncORF detectability

To better understand and estimate the differences between non-coding ORFs (ncORFs) that we detect and those that we do not detect and the differences between HLA-I peptides that we detect and those that we do not detect by Mass Spectrometry (MS), we have trained two models based on the properties of the detected and undetected ncORFs and HLA-I peptides. Inferences from these model results do not necessarily represent causality, yet they estimate how several amino acid sequence-based features may influence detectability by MS. Section 1 below describes the ncORF classification model, and Section 2 describes the HLA-I peptide classification model.

1. ncORFs microproteins classification model

To discern potential distinctions between detected and undetected ncORF microproteins, we curated a dataset encompassing both categories and applied a Statistical Learning model for analysis. Our analysis focused on a cohort of 7264 ncORFs, comprising 1785 that were detected and 5479 that remained undetected. Utilising the amino acid sequences corresponding to all ncORF microproteins, we computed a comprehensive set of 35 attributes. This process involved leveraging the Bio.SeqUtils.ProtParam module from BioPython and the Amino Acid Indices version 9.2 (<https://www.genome.jp/>). Subsequently, we employed the Boruta algorithm (<https://gitlab.com/mbq/Boruta/>) to select the most relevant features from the pool of 35 attributes.

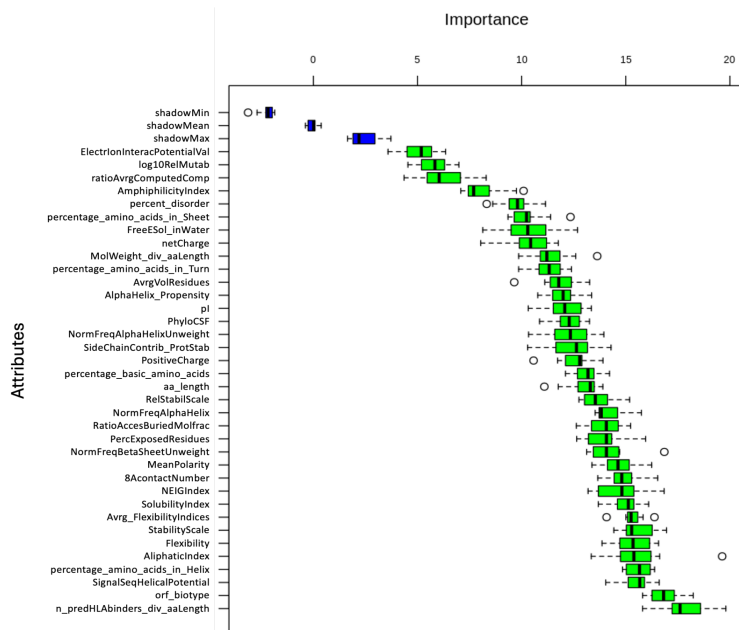


Figure 1. Boruta selected attributes. The Boruta algorithm tries to capture all the important features with respect to the outcome variable represented by ncORF peptide detectability. The Z-score, also known as the standard score, is a statistical measurement that describes a value's relationship to the mean of a group of values. The Boruta algorithm iteratively assesses the importance of each feature by comparing the Z-scores of actual features against those of

randomly permuted shadow features. Features that consistently have higher Z-scores than the shadow features are considered important, while those with lower Z-scores are considered unimportant and are removed from the model. Blue box plots correspond to minimal, average and maximum Z score of a shadow attribute. Red and green box plots represent Z scores of respectively rejected and confirmed, and yellow boxplots are considered tentative attributes.

Boruta confirmed 36 attributes (**Figure 1**, green) and no attributes were deemed unimportant. Since orf_biotype is a 'categorical feature' with 7 different types, we replaced orf_biotype with 7 features representing each category. We added the Instability Index attribute, and altogether, we used 43 attributes to test a cost-sensitive binary classification TensorFlow-Keras model.

1. ElectrIonInteracPotentialVal - Electron-ion interaction potential (Veljkovic et al., 1985)
2. log10RelMutab - Relative mutability (Dayhoff et al., 1978b)
3. ratioAvgComputedComp - Ratio of average and computed composition (Nakashima et al., 1990)
4. AmphiphilicityIndex - Amphiphilicity index (Mitaku et al., 2002)
5. percent_disorder - Bio.SeqUtils.ProtParam
6. netCharge - Net charge (Klein et al., 1984)
7. FreeESol_inWater - Free energy of solution in water, kcal/mole (Charton-Charton, 1982)
8. percentage_amino_acids_in_Sheet - Bio.SeqUtils.ProtParam
9. AvgVolResidues - Average volumes of residues (Pontius et al., 1996)
10. AlphaHelix_Propensity - Alpha-helix propensity derived from designed sequences (Koehl-Levitt, 1999)
11. pI - Isoelectric point (Zimmerman et al., 1968)
12. SideChainContrib_ProtStab - Side-chain contribution to protein stability (kJ/mol) (Takano-Yutani, 2001)
13. percentage_amino_acids_in_Turn - Bio.SeqUtils.ProtParam
14. NormFreqAlphaHelixUnweight - Normalised frequency of alpha-helix (Maxfield-Scheraga, 1976)
15. PhyloCSF - Phylogenetic Codon Substitution Frequencies
16. Percentage_basic_amino_acids - Bio.SeqUtils.ProtParam
17. PositiveCharge - Positive charge (Fauchere et al., 1988)
18. aa_length - length of ncORFs expressed as the corresponding number of amino acids
19. RatioAccesBuriedMolfrac - Ratio of buried and accessible molar fractions (Janin, 1979)
20. PercExposedResidues - Percentage of exposed residues (Janin et al., 1978)
21. RelStabilScale - The relative stability scale extracted from mutation experiments (Zhou-Zhou, 2004)
22. NEIGIndex - NNEIG index (Cornette et al., 1987)
23. NormFreqAlphaHelix - Normalised frequency of alpha-helix (Maxfield-Scheraga, 1976)

24. MeanPolarity- Mean polarity (Radzicka-Wolfenden, 1988)
25. Flexibility - Bio.SeqUtils.ProtParam
26. 8AcontactNumber - 8 A contact number (Nishikawa-Ooi, 1980)
27. AliphaticIndex - Bio.SeqUtils.ProtParam
28. NormFreqBetaSheetUnweight - Normalised frequency of beta-sheet, unweighted (Levitt, 1978)
29. SolubilityIndex - Bio.SeqUtils.ProtParam
30. MolWeight_div_aaLength - ratio Molecular Weight (Molecular weight Fasman, 1976) to the corresponding length in number of amino acids
31. SignalSeqHelicalPotential - Signal sequence helical potential (Argos et al., 1982)
32. Avrg_FlexibilityIndices - Average flexibility indices (Bhaskaran-Ponnuswamy, 1988)
33. StabilityScale - The stability scale from the knowledge-based atom-atom potential (Zhou-Zhou, 2004)
34. percentage_amino_acids_in_Helix - Bio.SeqUtils.ProtParam
35. n_predHLAbinders_div_aaLength - ratio of number of predicted HLA-binder peptides (this study) to the corresponding length expressed in number of amino acids
36. InstabilityIndex - Bio.SeqUtils.ProtParam
37. Orf_biotype - uoORF, upstream overlapped open reading frame
38. Orf_biotype - uORF, upstream open reading frame
39. Orf_biotype - intORF, internal open reading frame
40. Orf_biotype - lncRNA, long non-coding RNA
41. Orf_biotype - dORF, downstream open reading frame
42. Orf_biotype - processed transcript
43. Orf_biotype - doORF, downstream overlapped open reading frame

As the orf_biotype values represented categorical values and numeric values are required to generate and test the statistical model, 1 (as a value) was assigned to each ncORF with the corresponding orf_biotype category, and 0 when this did not apply. The ncORF Molecular Weight and the number of predicted HLA-binding peptide values were divided by the corresponding amino acid length.

Given that 1,785 ncORFs were detected while 5,479 remain undetected, presenting an approximate ratio of 1:3, the dataset exhibits an inherent imbalance. Therefore, we employed a balanced weight for imbalanced classification in Keras to address the imbalance, and a neural network analysis to build, train, and evaluate a TensorFlow-Keras model¹⁰³. The dataset with the selected attributes was used to implement this model using Python 3, and pandas, numpy, matplotlib, and various components of TensorFlow and Keras for building and evaluating the model. Training and testing sets were separated from the target variable, allocating 80% for training and 20% for testing, using the train_test_split function from scikit-learn.

Before fitting the model, the features were standardised using the StandardScaler, a preprocessing step to verify the features are on the same scale. The model was built as a

sequential model consisting of multiple layers: the input layer with 16 neurons, ReLU activation, and L2 regularisation. Batch normalisation and dropout layers were added after each hidden layer to prevent overfitting. The output layer consisted of a single neuron with sigmoid activation for binary classification.

The model was compiled using the Adam optimiser with a learning rate of 1e-3, binary cross-entropy loss function, and accuracy as the metric. It was trained on the training data. The training was run for 60 epochs with a batch size of 12. The model was used to predict the target variable for the test set, and the predictions were used to calculate the Receiving Operating Characteristic Area Under the Curve, ROC AUC score and make binary predictions.

Evaluation of the model:

ROC AUC Score: 0.678

**Confusion Matrix: [[620 454]
[119 260]]**

Classification Report:

	precision	recall	f1-score	support
0	0.84	0.58	0.68	1074
1	0.36	0.69	0.48	379
Accuracy			0.61	1453
macro avg	0.60	0.63	0.58	1453
weighted avg	0.72	0.61	0.63	1453

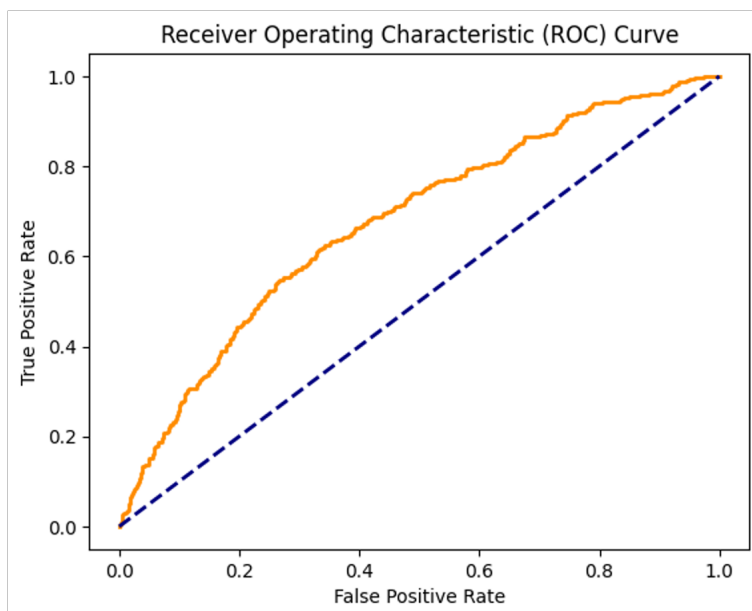


Figure 2. Receiving Operating Characteristic (ROC) Curve.

Figure 2 shows the results of the model on the test set as a ROC plot. The ROC AUC score is an indicator of how well the Tensorflow-Keras binary classification model discriminated between

detected and undetected ncORF peptides. A value of **0.68** indicates that the model exhibits a moderate capacity to distinguish between the detected and undetected open ORFs. The ROC AUC curve (orange) specifically shows the relationship between the true positive rate (sensitivity) and the false positive rate (1-specificity) for this model.

Considering the confusion matrix results, the model made 260 correct predictions for the negative class (undetected) and 620 correct predictions for the positive class. However, it also misclassified 119 negative instances as positive and 454 positive instances as negative.

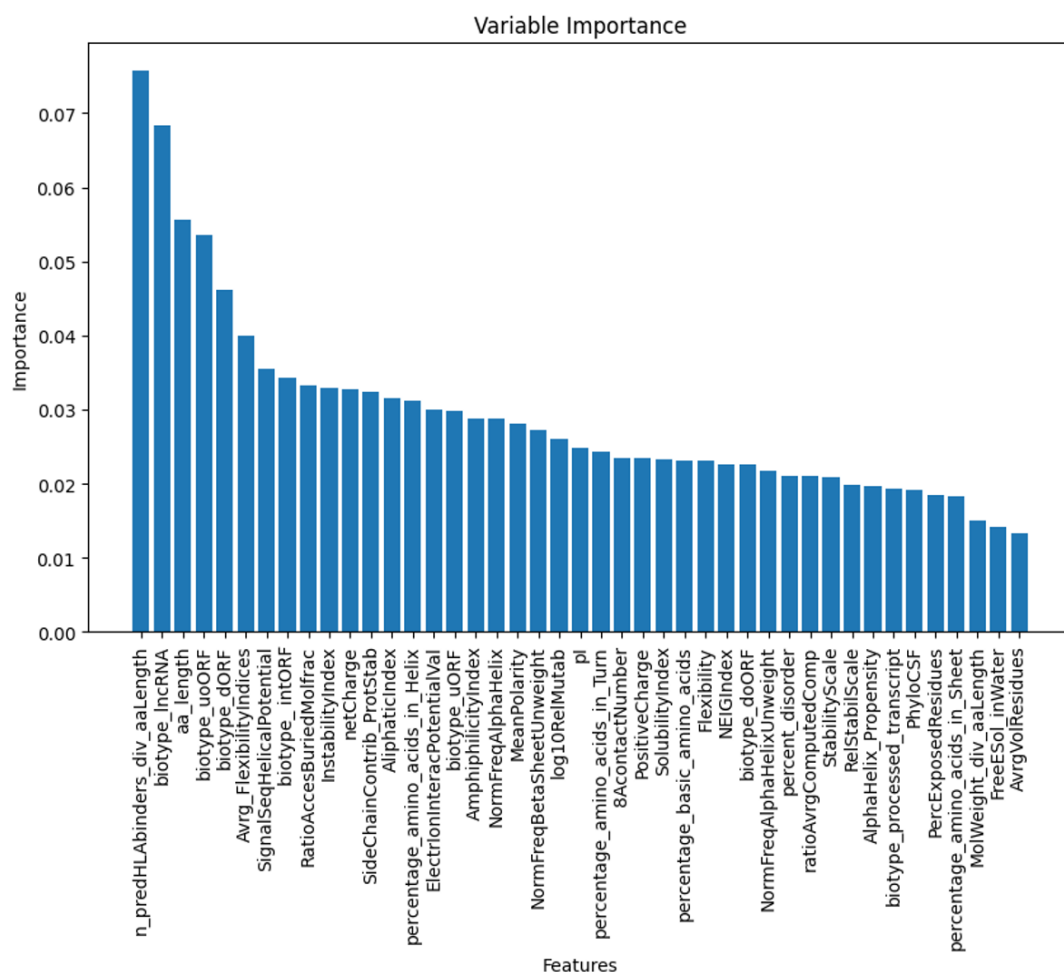


Figure 3. Variable Importance scores bar plot.

Supplementary Table 9 provides a complete list of all 7,264 ncORFs, including their identifier, sequence, the 43 features that the model used for training, and the output probabilities from the TensorFlow-Keras model.

The ratio of number of predicted HLA-binding peptides to ORF length (in amino acids) emerged as the highest importance attribute. The frequency of ORF detectability along this ratio and the proportions of detected and undetected are shown in **Figure 4**. Histogram distributions for detected and undetected ORFs appeared fairly similar, yet the model profits from these small

differences. Besides, more than 94% of peptides (2,937 out of 3,116; 94.3%) matching ncORFs were found to be presented by HLA.

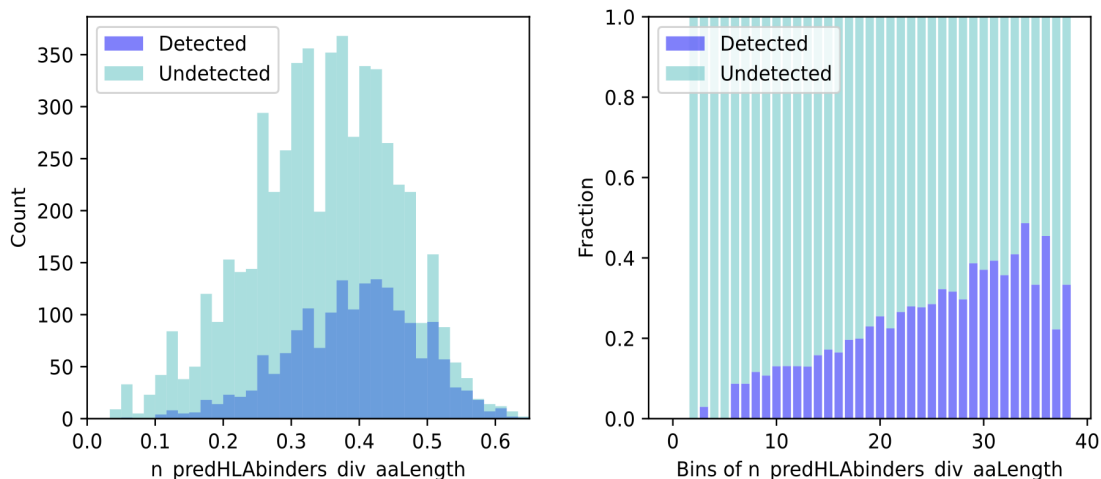


Figure 4. Ratio of the Number of predicted HLA-binding peptides to the ORF amino acid length's frequency represented as a Histogram plot (left side) and the corresponding Proportions plot (right side).

The ORF Biotype, specifically 'lncRNA', becomes the second more important variable. **Figure 5** represents the number of ORF lncRNA microproteins with all the designated ORF biotypes along the ORF detectability (left side) and the corresponding proportions on the right side.

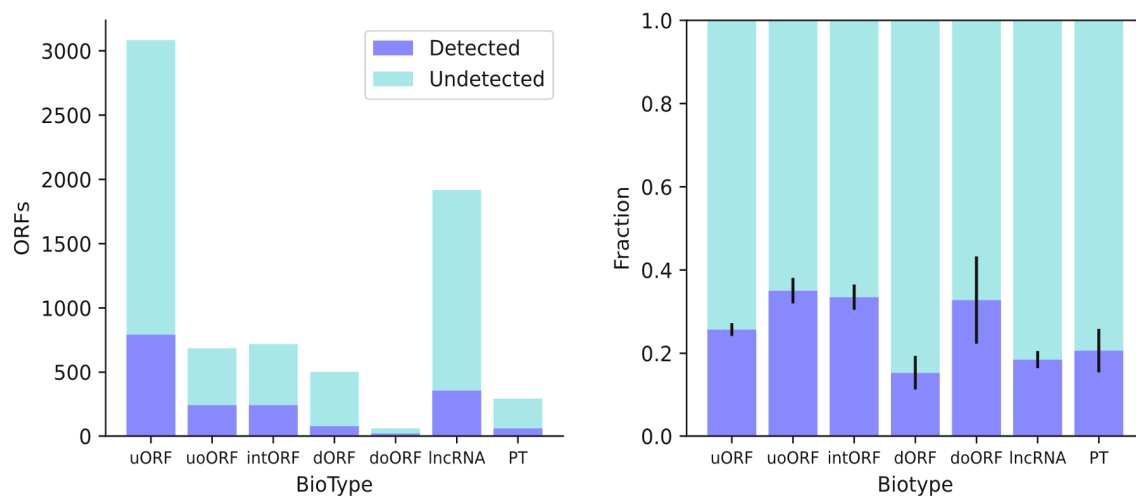


Figure 5. ORF Biotype counts versus Biotype. A bar plot (left side) representing the designated ORF biotypes along with ORF detectability (left side), and the corresponding Proportions plot (right side).

The ORF length expressed as the number of amino acids emerged as the third variable of importance. **Figure 6** represents the frequency of the detected or undetected ORFs along their amino acid length (left side) and the corresponding proportions (right side).

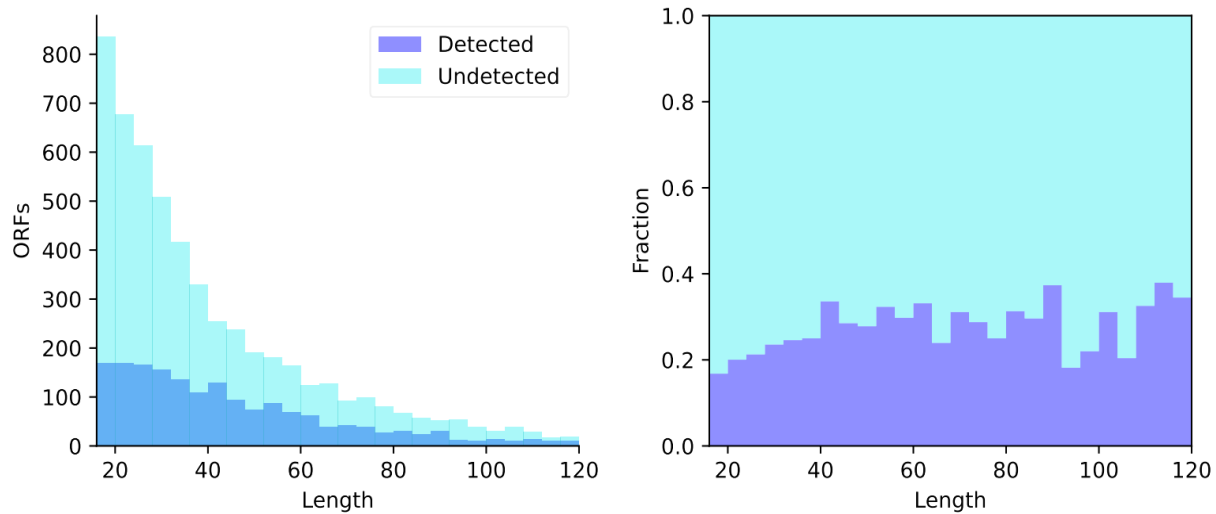


Figure 6. ORF amino acid length. The frequency of ORF detectability and the corresponding ORF length in number of amino acids are illustrated as a histogram plot (left side) together with the Proportions plot (right side).

While the model showed moderate discriminatory ability (as indicated by the ROC AUC score of 0.68), a substantial number of instances were misclassified, as revealed by the confusion matrix, thus there are likely additional factors that influence the detectability of ncORF.

The Number of Predicted HLA-binding peptides to length (amino acids) ratio appeared to be the most important attribute. Unsurprisingly, ncORFs rich in predicted HLA-binding peptides should be more easily detected.

The designated biotype of the ORF, specifically whether it originates from a lncRNA, emerged as the second most important attribute for their potential detectability. ORF microproteins derived from lncRNAs and dORFs are a little less likely to be detected, while the uORFs, uoORFs, and intORFs are moderately more likely to be detected.

Overall, the ncORF length affects detectability, as the shortest ncORFs are more difficult to detect.

2. ncORF HLA-I predicted binder peptides classification model

Considering the model results in section 1, where the number of predicted HLA-binding peptides may play an influential role in the ncORF detectability, and most MS-based proteomics approaches identify peptides, the next step was to focus on the analysis of ncORF peptide sequences. Then, to discern potential differences between detected and undetected HLA-I predicted binder peptides, we curated a dataset comprising both categories and applied a Statistical Learning model for analysis.

We focused on 9-amino acid (9aa) long detected HLA-I predicted peptides, originating from ncORFs. We created a list of 341 9aa detected peptides, each with a binding score (min_rank)

falling within the range of 0.1 to 1.7. We then paired these peptides, along with their corresponding best alleles, with undetected HLA-I predicted peptides of the same length (9aa), ensuring alignment with the closest possible binding score and allele.

From the combined set of all peptides (totaling 677), we derived a comprehensive array of 63 attributes. These attributes were computed utilising both the Bio.SeqUtils.ProtParam module from BioPython and the Amino Acid Indices version 9.2 (<https://www.genome.jp/>). Employing the Boruta algorithm (<https://gitlab.com/mbq/Boruta/>), we selected the most discriminative features from the pool of 63 attributes.

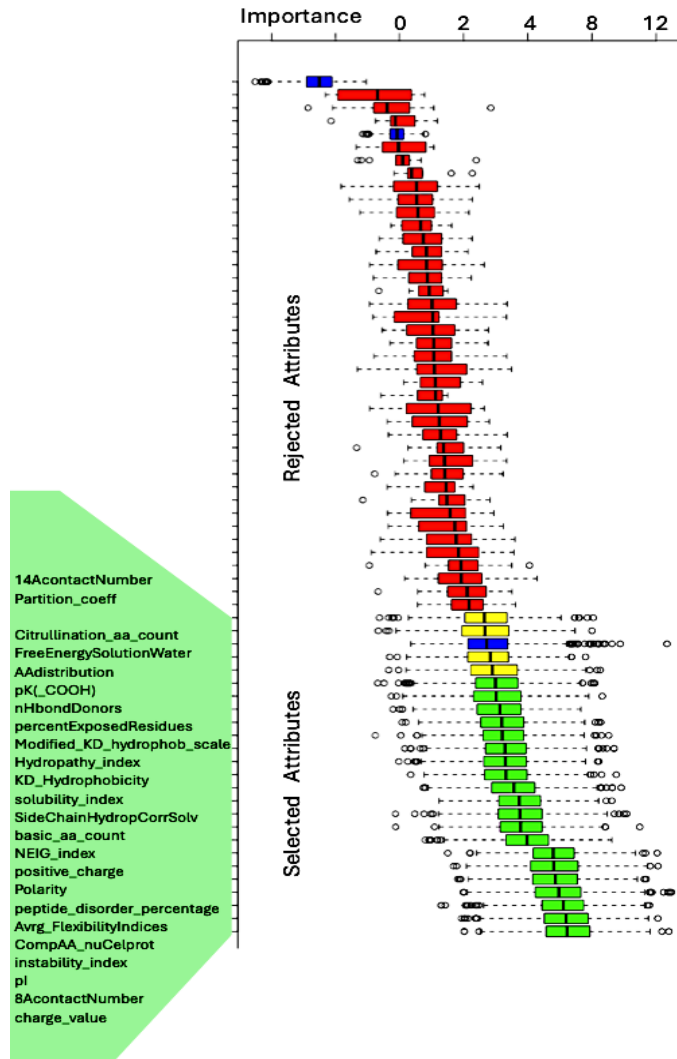


Figure 7. Boruta selected attributes. As described in the previous section, features that consistently have higher Z-scores than the shadow features are considered important, while those with lower Z-scores are considered unimportant and are removed from the model. Blue box plots correspond to minimal, average and maximum Z score of a shadow attribute. Red and green box plots represent Z scores of respectively rejected and confirmed, and yellow boxplots are considered tentative attributes.

Boruta selected and confirmed 20 attributes (**Figure 7**, green boxplots), 4 were assessed as tentative (yellow boxplots), all 24 are shown on the left side green panel. 39 attributes were rejected (red boxplots) as being not useful in discriminating between detected and undetected. We further selected all confirmed 20 (green boxplots) and added 2 tentative attributes. The following 22 attributes associated with these ncORF peptides were finally utilised to test a statistical learning Multilayer Perceptron¹⁰⁴ model, (MLPClassifier):

1. Partition_coeff - Partition coefficient (Garel et al., 1973)
2. Citrullination_aa_count - Number of residues that can be modified by Citrullation
3. FreeEnergySolutionWater - Free energy of solution in water, kcal/mole (Charton-Charton, 1982)
4. pK(_COOH) - pK-a(RCOOH) (Fauchere et al., 1988)
5. nHbondDonors - Number of hydrogen bond donors (Fauchere et al., 1988)
6. percentExposedResidues - Percentage of exposed residues (Janin et al., 1978)
7. Modified_KD_hydrophob_scale - Modified Kyte-Doolittle hydrophobicity scale (Juretic et al., 1998)
8. Hydrophathy_index - Hydropathy index (Kyte-Doolittle, 1982)
9. KD_Hydrophobicity - Hydrophobicity (Kyte & Doolittle, 1982)
10. solubility_index - Bio.SeqUtils.ProtParam
11. SideChainHydropathyCorrectSolvation - Side chain hydropathy, corrected for solvation (Roseman, 1988)
12. basic_aa_count - number of basic amino acid residues.
13. NEIG_index - NNEIG index (Cornette et al., 1987)
14. positive_charge - Positive charge (Fauchere et al., 1988)
15. Polarity - Polarity (Grantham, 1974)
16. peptide_disorder_percentage - Bio.SeqUtils.ProtParam
17. Avrg_FlexibilityIndices - Average flexibility indices (Bhaskaran-Ponnuswamy, 1988)
18. CompAA_nuCelprot - Composition of amino acids in nuclear proteins (percent) (Cedano et al., 1997)
19. Instability_index - Bio.SeqUtils.ProtParam
20. pl - Isoelectric point (Zimmerman et al., 1968)
21. 8AcontactNumber - 8 A contact number (Nishikawa-Ooi, 1980)
22. Charge_value - Bio.SeqUtils.ProtParam

A dataset with these attributes served as a basis to generate an MLP Classifier model using Python 3, and software/libraries: pandas, numpy, matplotlib, sklearn. This involved data preparation, model initialisation and tuning, model fitting, prediction, and evaluation. First, the features of the dataset were separated from the target variable, and the data was split into training and testing sets, allocating 80% for training and 20% for testing, while ensuring reproducibility by setting the random state to 42.

Before fitting the model, the features were standardised using the StandardScaler, a preprocessing step that removes the mean and scales the features to have unit variance. Next, an MLP Classifier model was initialised with a maximum of 8000 iterations and a random state of 42. The model was then tuned using grid search with cross-validation, exploring various hyperparameters including the hidden layer sizes, activation function, and regularisation parameter. Grid search with cross-validation tuned the model using the specified hyperparameters: Hidden layer sizes: (280,) Activation function: 'tanh', and Regularization parameter (alpha): 0.01.

Upon completion of the tuning process, the best performing model was identified based on the grid search results. This model was used to make predictions on the previously untouched test set, allowing for an assessment of its predictive capabilities. Finally, a range of evaluation metrics including ROC AUC, F1 score, accuracy, precision, recall, and the confusion matrix were calculated to gauge the model's performance on the test set.

Evaluation of the model:

ROC AUC Score: 0.694

F1 Score: 0.732

Accuracy Score: 0.699

Precision Score: 0.737

Recall Score: 0.727

Confusion Matrix: $\begin{bmatrix} 39 & 20 \\ 21 & 56 \end{bmatrix}$

Considering all the metrics results, the model demonstrates relatively balanced performance. The accuracy score of 0.699 indicates that approximately 70% of the predictions were correct. The precision and recall scores, both around 0.73, suggest that the model is reasonably good at correctly identifying positive cases and minimising false positives.

Looking at the confusion matrix, the model made 39 correct predictions for the positive class (detected) and 56 correct predictions for the negative class. However, it also misclassified 21 negative instances as positive and 20 positive instances as negative.

Figure 8 shows a ROC plot for the test set of 677 ncORF peptides.

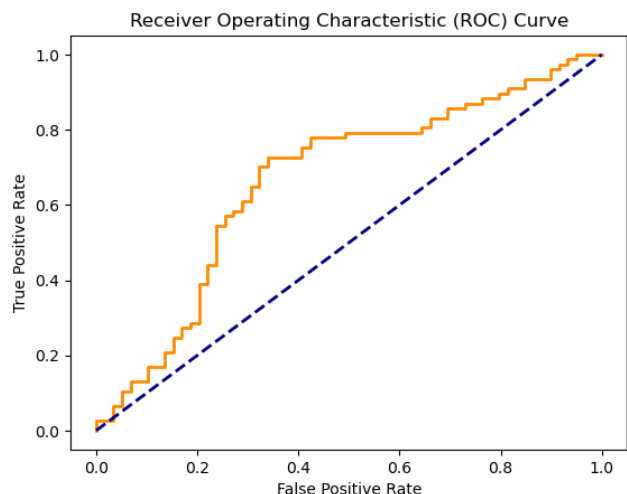


Figure 8. MLPClassifier's output ROC AUC curve. The ROC AUC score is an indicator of how well the MLP Classifier model distinguishes between detected and undetected ncORF peptides. A value of **0.69** suggests that the model has some ability to differentiate between the classes. The ROC AUC curve (orange) specifically shows the relationship between the true positive rate (sensitivity) and the false positive rate (1-specificity) for the MLP binary classification model. Here, the model's performance changes as the discrimination threshold is adjusted, providing insights

into its ability to correctly classify positive and negative instances. Then, the AUC value quantifies the overall predictive accuracy of the MLP model, with a higher AUC indicating better performance in distinguishing between detected and undetected peptides.

Supplementary Table 10 provides a complete list of all 677 peptides, including their sequence, best allele, the 22 features that the model used for training, and the output probabilities from the model.

Figure 9 shows the highest importance computed property, Instability Index versus the model-predicted probability of detection. Instability index is usually calculated using various algorithms that consider factors such as amino acid composition, secondary structure, and other physicochemical properties. **Figure 9** suggests an increase in ncORF HLA-I peptide detectability correlates with an increase in stability.

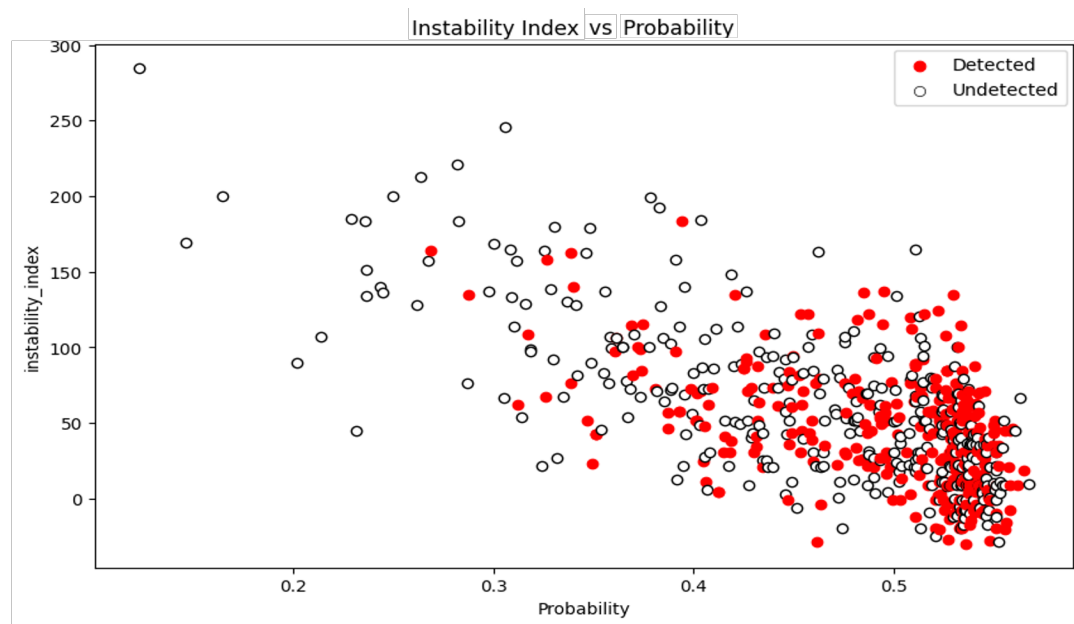


Figure 9. Scatter plot of the probability of detection for each ncORF HLA-I predicted peptide based on the Instability Index. The detected ORF peptides are depicted in red.

Figure 10 shows the Variable Importance scores for the attributes associated with the Instability Index according to the MLP model.

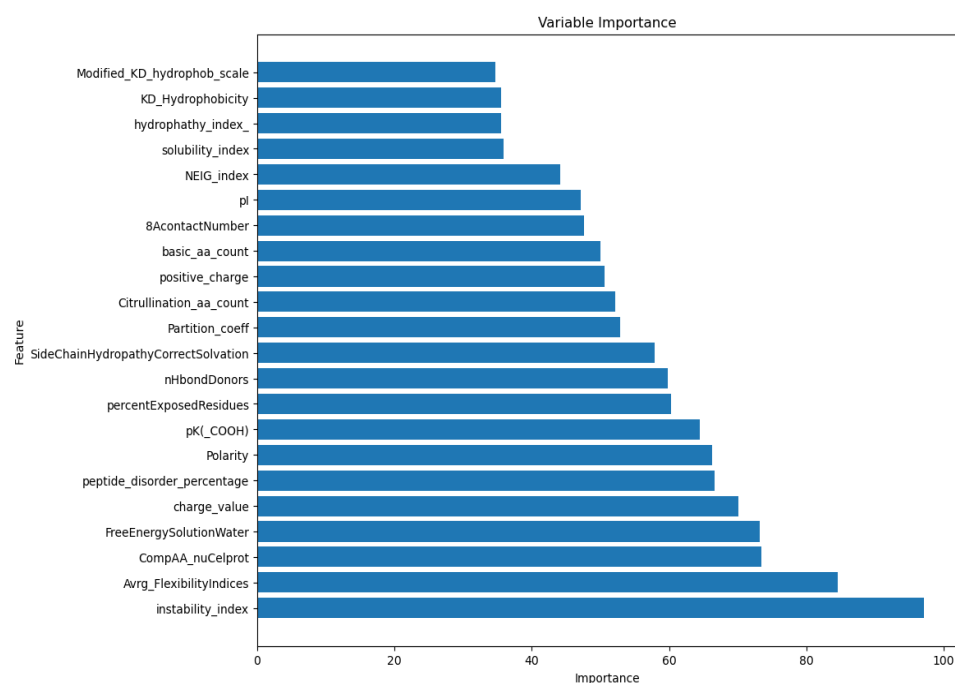


Figure 10. Variable Importance scores bar plot.

While the model showed moderate discriminatory ability (as indicated by the ROC AUC score) and achieved decent accuracy, precision, and recall scores, it also misclassified a notable number of instances. The model performance may improve e.g. when data on ncORFs HLA-I peptide abundance or other relevant attributes become available.

Although the performance of the model is not optimal to classify peptides as detectable or undetectable, the most important features appear to be correlated, with the peptide Instability Index as the most important attribute for their potential detectability.

A lower instability index suggests that a peptide is more stable, whereas a higher index indicates greater instability. The usefulness of this variable in the model can be explained by several key factors:

Peptide Degradation: peptides with high instability indices are more prone to degradation. In MS-based experiments, unstable peptides may degrade during sample preparation, storage, or analysis, leading to reduced detectability. Stable peptides are more likely to remain intact, ensuring they reach the mass spectrometer and produce reliable signals.

Ionisation Efficiency: the stability of a peptide can affect its ionisation efficiency. Unstable peptides might undergo partial degradation or modifications that alter their ionizable groups, impacting their ionisation efficiency and, consequently, their detectability by MS.

Fragmentation Patterns: in MS/MS (tandem MS), peptides are fragmented to provide sequence information. Peptides with a high instability index might produce unpredictable or incomplete

fragmentation patterns, complicating their identification and reducing confidence in their detection.

Protein Expression and Processing: the stability of peptides can also reflect their processing and expression within the cell. Stable peptides are likely to be present in higher quantities and more consistently processed, leading to higher chances of detection.

Biological Relevance: The Instability Index might correlate with other biologically relevant properties such as protein half-life, subcellular localisation, and interaction with other cellular components. These factors collectively influence the overall abundance and accessibility of peptides for MS analysis.

(3) Detection of ncORFs in human ubiquitinated proteomics datasets

Selection of datasets

Ubiquitination-enriched datasets were manually selected from the PRIDE database and reanalyzed. The datasets were selected based on the following criteria: (i) human-derived samples enriched for lysine ubiquitination; (ii) data generated using Thermo Fisher Scientific instruments; (iii) availability of metadata, either through the original publication or by direct communication with the authors; and (iv) derived from human cell lines (e.g., HEK293, HeLa, HCT116, see **Supplementary Table 5**) enriched for ubiquitination using the Gly-Gly/UbiSite methods.

A total of 11 ubiquitinated datasets (**Supplementary Table 5**) were selected after manual curation from the preliminary selection of 27 datasets. Most of them were supplemented with a proteasome inhibitor. When the dataset contained different experiment settings, they were reanalysed separately. For instance, the dataset PXD037009 was split into 5 experiments because of the different experimental conditions.

Proteomics raw data processing and post-processing

MS raw files were processed using a standardized protocol for PTM-enriched datasets. Datasets were analysed separately, raw files from each dataset were converted to mzML format using ThermoRawFileParser (version 1.3.4) and analyzed independently. The search database included the UniProt human reference proteome (one protein per gene, downloaded on April 4, 2024), the complete list of 7,264 ribo-Seq ORFs, and cRAP protein sequences as a contaminants database (<https://www.thegpm.org/crap/>, obtained April 2024). Decoys were generated using the reverse decoy method to the FASTA file via FragPipe (<https://fragpipe.nesvilab.org/>). Semi-tryptic digestion was employed to account for incomplete digestion, expanding the range of detectable peptides. Peptide and protein identification, including ubiquitination sites, was performed using the Comet search engine (version 2024) on a Linux-based high-performance computing cluster. Default parameters were applied, with the following exceptions: missed cleavages were set to 4, and variable modifications included oxidation of methionine and N-terminal protein acetylation (excluding N-terminal peptide acetylation).

To ensure high-confidence identifications, the searching result files from each dataset were combined and processed using PeptideProphet, iProphet and PTMProphet from the Trans-Proteomic Pipeline (TPP version 7.1.0). False localization rates (FLRs) were estimated using a decoy amino acid (Alanine)⁷⁵, to calculate site-specific FLRs. A 1% FDR threshold was applied to filter high-confidence peptide-spectrum matches (PSMs).

Results

A summary of the total results of the proteomics analysis in terms of total number of Peptide Spectrum Matches (PSMs) and PSMs corresponding to ubiquitinated-peptidodforms is provided in **Supplementary Table 5**.

Of those, 151 PSMs including ubiquitinated peptides, corresponding to 19 RiboSeq ORFs, were detected (**Supplementary Table 5**, including Universal Spectrum Identifiers, USIs). Detections were checked by visualising the spectra in the PRIDE Spectra USI Archive (<https://www.ebi.ac.uk/pride/archive/usi>). Six of them were detected by more than one peptidoform:

- 3 peptidoforms: c17norep138 (7 PSMs in total, only detected in the PeptideAtlas HLA build), c7norep167 (12 PSMs, undetected in PeptideAtlas), and c7riboseqorf100 (23 PSMs, only detected in the PeptideAtlas HLA build);
- 2 peptidoforms: c15norep52 (10 PSMs, detected in both PeptideAtlas HLA & non-HLA builds), c3riboseqorf10 (6 PSMs, only detected in the PeptideAtlas HLA build), and c5norep142 (22 PSMs, detected in both the PeptideAtlas HLA & non-HLA builds).

The rest of the RiboSeq ORFs (13 of them) were detected by one peptidoform. Three of these are undetected in PeptideAtlas, six are only detected in the HLA build, and four in both the HLA and non-HLA builds. The number of PSMs for each RiboSeq ORF (for each peptidoform in this case) varied between just one (peptide DIPHTLK[Ubiq]QISFR, c11riboseqorf38 (undetected in PeptideAtlas); LKGTAAVK[Ubiq]K, c6riboseqorf130) and 15 PSMs (SDGVSPK[Ubiq]HVGR, c10riboseqorf41 (detected in the PeptideAtlas HLA & non-HLA builds)).

This analysis demonstrates the detection of RiboSeq ORFs using ubiquitination-enriched proteomics datasets, complementing previous findings from HLA-enriched and standard tryptic digestion experiments.

(4) ORBL: motivation, methodology, validation, and limitations

Motivation and starting point

ORBL, an acronym for Open reading frame (ORF) Relative Branch Length, is an approach to measuring evolutionary conservation and constraint on the translatability of a segment of an RNA transcript, regardless of whether there is conservation or constraint on the hypothetical peptide produced by such translation. It was motivated by the existence of uORFs whose peptide product is not believed to serve any cellular function but for which the act of translation is an important regulator of translation of the downstream protein-coding ORF^{105,106}. Computational tools designed to detect protein-coding genomic regions, such as PhyloCSF²⁴ will often fail to recognize translation of such uORFs because they focus on substitutions and amino acid frequencies typical of protein-coding regions, which are indicative of function of the translated protein rather than the act of translation. Since we lack experimental information about actual translation in most other species, ORBL estimates translatability by the presence of a start codon, a stop codon, and a reading frame uninterrupted by other stop codons, which together we refer to as “ORFness”.

Measuring evolutionary constraint on the ORFness of an ORF within a clade requires three things: determining if the ORF is conserved in each particular species in the clade, combining that information into a score that measures conservation in the clade, and comparing that score to a null model to estimate how likely it is that the ORF would have had the same or greater level of conservation if there were no evolutionary constraint to preserve its ORFness, which is necessary to rule out the possibility that the conservation is due to chance rather than constraint.

ORBL takes as input the genomic coordinates of an ATG-initiated open reading frame in a reference species. It then uses a multi-species whole-genome alignment to find syntenic orthologous regions in other species in a clade. An alternative would have been to use BLAST or another search tool to find all homologs in the other species. However, that could include non-syntenic aligned sequences, which is undesirable because these could be recent paralogs that have lost their function, whereas using the whole-genome alignment has the advantage that it is more likely to align the ORF to a syntenic sequence.

ORBLv conservation score

We consider the ORFness of an ATG-initiated ORF to be conserved in another species if there is an ATG codon aligned to the ORF start codon, a stop codon aligned to the ORF stop codon, and a total length that is a multiple of 3 with no intermediate in-frame stop codons, even if there are insertions or deletions that change the reading frame for part of the ORF. Several considerations went into our choice to define ORFness conservation in the way we have. First, since many ncORFs begin with a non-ATG start codon, it would be desirable to remove the requirement that the start codon be ATG. However, that would require determining whether conservation of function requires the same aligned non-ATG codon, any near cognate codon, or some subset of

near cognate codons. We hope to remove this restriction in the future. Next, requiring the ATG and stop codon to be aligned to the original ones might be too restrictive. For example, it might be that an ORF in which there is an ATG or stop codon within a certain distance of the ones in the reference species would maintain the function of the ORF. However, our knowledge of the complete range of possible functions of translation of an ncORF is limited, so we do not know which movement might damage the function. Consequently, we have taken the conservative approach and only consider the ORFness to be conserved if the start and stop codons are exactly aligned. Finally, we do not require any conservation of the internal sequence of the ORF, other than keeping the reading frame open, although features of this sequence could influence its functionality. For example, in some cases the length of the ORF (which can be changed substantially by insertions and deletions), the presence of certain codons that slow down translation, or codons that translate to amino acids that interact with the ribosome exit channel are important for the functionality of a ncORF^{106,107}. However, without a complete understanding of the mechanisms by which translation can function, and which mutations would disrupt that function, we have not imposed any other requirements on the ORF interior.

To measure conservation of the ORF in the clade, we have defined ORBLv (v for conservation) to be the relative branch length of the conserved species. That is, we start with the phylogenetic tree of all species in the whole genome alignment, calculate the branch length of the smallest subtree that includes all the conserved species, and divide by the branch length of the whole tree. Note that for the latter we include all species in the whole genome alignment, even ones that are not present in the local alignment of the ORF, because an ORF that is well conserved in a small subset of closely related species but that has no alignment beyond that subset should not be considered conserved in the whole clade.

Relative branch length has been used previously for measuring conservation of other properties. Examples include regulatory motifs¹⁰⁸, the “bls” branch length score of PhyloCSF, which is used for the PhyloCSF Power track in the PhyloCSF track hub for display in genome browsers¹⁰⁹, and the most recent adaption of BLS for ncORF conservation by Dr. Hong Zhang and colleagues¹¹⁰. Relative branch length has several advantages over other methods of measuring conservation in a clade, such as counting the number of conserved species: it takes into account evolutionary distance so an ORF that survived over a longer evolutionary time gets a higher score; it accounts for redundancy, in that conservation of the ORF in two closely related species only adds a small increment to the score compared to only one of them; and it mitigates the apparent loss of conservation due to sequencing, assembly, or alignment errors, because the presence of the ORF in closely related species can rescue most of the branch length.

ORBLq constraint score

What we would really like to know is not whether an ncORF is conserved, but whether it is translated and the translation has served a conserved function. If so, then we would expect purifying selection to have imposed constraint to preserve it as an ORF. The mere fact that an orthologous ORF is present in some distantly related species is not in itself a sufficient indication of this, because that conservation could be due to chance. For example, very short ORFs are less

likely than longer ORFs to gain interior stop codons or frame-changing insertions and deletions, making them more likely to be conserved by our definition simply due to chance. Alternatively, conservation of an ncORF could be due to purifying selection to preserve some other property of that DNA. For example, purifying selection to preserve the reading frame of a CDS will have a similar effect on any overlapping intORFs, and selection to preserve the amino acid sequence of the CDS would eliminate some of the mutations that might otherwise have disrupted the start or stop codon of the overlapping ncORF.

We would like to determine if an ncORF has undergone purifying selection *specifically* for ORFness. If it has, then it should be significantly more conserved than a typical untranslated ORF of the same length having the same constraints, other than constraint on ORFness. To this end, we have compiled a list of ORFs that are not believed to be translated. Then, we compared the ORBLv score of each ncORF to the scores of a subset of these untranslated ORFs that have similar length and similar constraints, other than constraint on ORFness. While we cannot match all possible constraints, both because not all are known and because with a finite list of untranslated ORFs there might not be enough of them that match our ncORF to get a statistically significant result, we can at least match biotype and, in the case of overlapping biotypes (uoORF, intORF, and doORF), the frame of overlap. The latter is important because overlap in different frames imposes different constraint on the start and stop codons, as well as on potential premature stop codons.

We defined our measure of evolutionary constraint on ORFness, ORBLq (q for quantile), to be the quantile of an ncORF's ORBLv score among the ORBLv scores of these matched ORFs, i.e., those from our list of untranslated ORFs having the same biotype, same frame of overlap, and similar length.

Because our analysis of evolutionary constraint is strongly influenced by overlap with CDS, we first redid the biotype determination for the ncORFs using the most recent GENCODE annotations at the time we began our analysis, and applied strict criteria to determine ORFs with a “pure” biotype. The particular requirements for each biotype were:

- uORF: Contained in the 5'-UTR of some transcript, and does not overlap any CDS on the same strand.
- uoORF: Starts in the 5'-UTR of some containing transcript and extends into its CDS, without any other overlap with any CDS on the same strand in any reading frame (other than transcripts that share CDS in the same frame as the containing transcript).
- intORF: Entirely within the CDS of some containing transcript, and does not overlap any other CDS on the same strand in a different reading frame.
- d/doORF: Mirror the requirements of u/uoORF but in the 3'-UTR instead of the 5'-UTR.
- lncRNA-ORF: no intersection with any mRNA on the same strand. We have continued to call these lncRNA-ORF for consistency with Mudge *et al*⁴, although some were on lncRNA transcripts when that list was compiled but were not contained in any lncRNA (or other) transcript in the later annotations.

- For intORF, uoORF, and doORF, the same-strand coding intersection must be contiguous on the transcript and in only one reading frame, which is not the coding frame of any overlapping CDS.

If an ncORF did not satisfy any of these criteria, for example, an ORF that overlaps two CDSs from different transcripts in different reading frames; we considered these to have a “mixed” biotype and did not compute an ORBLq constraint score for them.

To generate the list of “untranslated” ORFs from which matched ORFs were chosen, we began with every ATG-initiated open reading frame of any length in any protein-coding or lncRNA transcript, provided that the ORF satisfied the above criteria for one of the biotypes. Since we needed a large number of such ORFs of various sizes, we did not require them to be maximal, so, for example, if an ORF included a downstream ATG the list would include another ORF starting at that downstream ATG and ending at the same stop codon. In order to minimize the chance that the ORF is translated, or is partly constrained by overlap with a translated ORF, we excluded ORFs that overlapped one of the 7264 ncORFs, an mRNA on the opposite strand, or a pseudogene on either strand. We also excluded dORFs that overlapped some 5'-UTR because we had found that 5'-UTRs were generally more conserved than 3'-UTRs. We then computed the ORBLv constraint score for each of these untranslated ORFs

To choose the ORFs to match with a given ncORF, we begin with all of our untranslated ORFs that have the same biotype and reading frame, and have exactly the same length. In some cases, there were very few such ORFs, so we needed to include some ORFs of similar but different length. This requires a compromise between the need to match to ORFs having lengths as close as possible to that of the ncORF (since ORFs of different lengths will have a different probability of being conserved due to chance) and the need to have enough matched ORFs to achieve statistical significance. We resolved this by adding all ORFs of length one more or one less, two more or two less, and so on, until there were at least 1000 matched ORFs.

As noted above, some ORFs can be conserved simply due to chance even if there has been no constraint on their ORFness. By defining ORBLq using a comparison to matched untranslated ORFs, we assure that a high ORBLq score indicates not just that an ORF is conserved, but that it is more conserved than neutrally evolving ORFs, and thus indicates that the conservation is due to evolutionary constraint rather than chance.

Limitations

There are several limitations of our ORBLv conservation score. We define an ORF to be conserved only if it has start and stop codons exactly aligned to those in human, though it is possible that the function would persist despite slight shifts. For now, we have not defined ORBLv for non-AUG starts, something that is not required for the current list of ncORFs but likely will be in the future. We do not consider conservation of other features that contribute to whether an ORF is translated, such as the Kozak context, presence of a 5' ATG in a different frame, proximity to the 5' cap, etc., though those could provide additional clues as to conserved function. We do not

require any conservation of the internal sequence of the ORF, other than keeping the reading frame open, although features of this sequence could influence its functionality. Some species in which the ORF is actually conserved might be treated as not conserved due to sequencing, assembly, or alignment errors, though that will be mitigated if there is a closely related species without such errors, so the missing branch length will be small. Our scores for primates lack statistical power because the primate tree has only about one tenth of the branch length of the placental mammal tree. Finally, recently evolved ORFs will get low ORBLv scores even if they are functional.

There are also several limitations of our ORBLq constraint score. In some cases, we do not have many untranslated ORFs of similar length to match to a given ORF, so the ORBLq score is not reflective of the actual ORF length; this is particularly a problem for long ncORFs and ones with less common biotypes, such as uoORFs and doORFs. For uoORFs and doORFs we try to match the overall ORF length, but we do not try to match the fraction that overlaps the CDS, which is related to the amount of “free” conservation the ncORF gets from constraint on the CDS. When ORBLq is close to 1, say, greater than 0.99, small differences are highly relevant but there will be high statistical uncertainty because with as few as 1000 matched untranslated ORFs there will be very few with larger ORBLv. Currently, we do not compute ORBLq scores for the roughly 6% of ncORFs that have “mixed” biotype. There are reasons other than constraint on ORFness or constraint on an overlapping CDS that could inflate the ORBLv and ORBLq scores of individual ORFs; for example, if the ORF start or stop lies in an enhancer motif it might be highly conserved even if there is no constraint for ORFness. If such alternative constraints were more common among the untranslated ORFs than the ncORFs, or vice versa, that would bias the ORBLq scores. Finally, some of our “untranslated” ORFs might actually be translated; that would dilute any differences between the ncORFs and matched controls and result in ORBLq scores that understate the true level of ORFness constraint.

Correlating ORBL with functional readouts

Ideally, we would validate ORBL by measuring its ability to distinguish known functional ncORFs from non-functional ones. However, currently there is no sufficiently large “gold standard” list of functional ncORFs. Instead, we measured the correlation of ORBL scores with four functional readouts that are themselves expected to correlate with functionality (**Figure 1**): (i) our Tier-based classification system for ncORFs, (ii) our pooled CRISPR cumulative genetic dependency data, (iii) conservation of ORF translation across species, as measured by ribosome profiling data, and (iv) HLA-I presentation of ncORFs. Our results are shown below. We found a statistically significant positive correlation between ORBL scores and each of the four measures.

1. ncORF Tier classifications

We have plotted the log-transformed ORBLq scores for each Tier level. The median ORBLq scores are highest for Tiers 1A and 1B, and lowest for Tier 3. Spearman

correlation shows a statistically significant relation between ORBLq scores and higher Tier levels ($\rho = -0.099$, $p = 2.4 \times 10^{-16}$).

2. CRISPR cumulative genetic dependency

We used the CRISPR screen data of 8 cell lines for 2,196 ncORFs to correlate the log-transformed ORBLq scores with the number of cell lines in which an ncORF was found to affect cell fitness (chronos score < -0.5 ; cumulative genetic dependency). Spearman correlation shows a statistically significant relation between ORBLq scores and cumulative genetic dependency ($\rho = 0.087$, $p = 6.3 \times 10^{-6}$).

3. Ribo-seq translation

As a third strategy, we assessed ORBL relationship with evolutionary conservation of Ribo-seq translation. Specifically, we have compared ORBL metrics with mammalian Ribo-seq translation levels in a subset of 1,183 ncORFs translated in the heart, as defined in Ruiz-Orera et al. 2024⁸⁴. We estimated the translation age of these cardiac ncORFs by integrating Ribo-seq data from human, chimpanzee, macaque, mouse, and rat, which gives an estimate of the translatability of these sequences across primate and rodent mammalian lineages. This evolutionary analysis showed that ncORFs with evidence of translation across multiple mammalian species consistently display significantly higher ORFness constraint values as measured by ORBLq (Pairwise Wilcoxon tests with adjusted p-values displayed in **Figure 1c**). This supports the idea that ORBL captures biologically meaningful signals linked to evolutionary persistence of translation.

4. HLA-I presentation

The correlation of ORBLq scores with HLA-I detection of ncORF-derived peptides is included in the main manuscript as Extended Data Figure 9f.

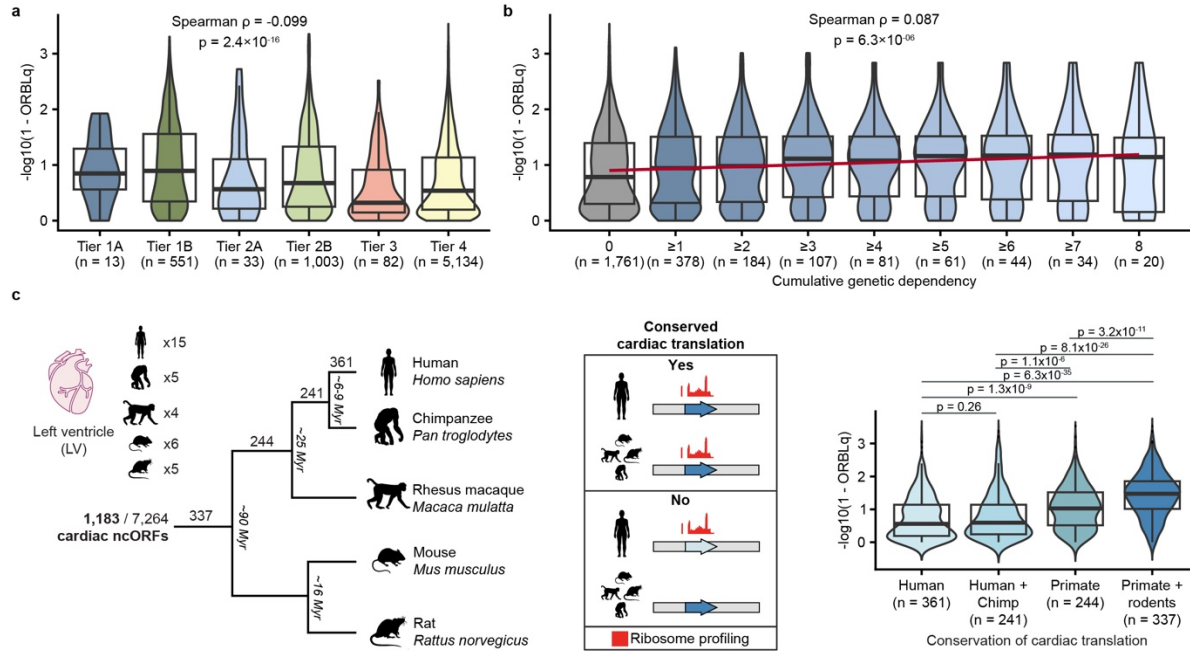


Figure 1. Correlation of placental ORBLq with functional readouts. **(a)** Violin plots showing the $-\log_{10}(1 - \text{ORBLq})$ transformed ORBLq distribution for each Tier level. Only ncORFs without a mixed biotype in Gencode V45 are included **(b)** Violin plots correlating the $-\log_{10}(1 - \text{ORBLq})$ transformed ORBLq with ncORF cumulative genetic dependency. Note that for categories $\geq 1 - 8$, ncORFs are represented cumulatively. Cumulative genetic dependency is based on the CRISPR screen of 8 cell lines. 2,139 of the 2,196 measured ncORFs are included based on having no mixed biotype in Gencode V45. **(c)** Left: Schematic of the conservation of translation of 1,183 cardiac ncORFs across five mammalian species. Data was retrieved from Ruiz-Orera et al. 2024. Right: Violin plots correlating the $-\log_{10}(1 - \text{ORBLq})$ transformed ORBLq with the conservation of cardiac translation of the 1,183 ncORFs. Statistical significance between groups was assessed using two-sided Wilcoxon tests, with adjusted p-values displayed on the plot.

Benchmarking ORBL and BLS

There are conceptual similarities between ORBL and the BLS tool developed by Dr. Hong Zhang and colleagues¹¹⁰ so we compared ORBLq to the most recent adaptation of BLS for ncORF conservation.

This adaptation considers an ORF structure to be conserved in an orthologous or reconstructed ancestral sequence if at least one of the first three codons is an ATG, CTG, GTG, or TTG codon and if the region orthologous to at least 70% of the 5' region of the reference ncORF is free of in-frame stop codons. To be able to compare both metrics in a meaningful way, we calculated the BLS score using the 116-placental mammal alignments previously used for calculating ORBL. We calculated both local and global BLS scores, using both Ancestral Sequence Reconstruction (ASR) and a "naive" method that depends only on sequences in extant species. For the ASR scores, we first predicted the sequence at each ancestral node using PRANK v.170427¹¹¹ in codon mode, with ancestral states and evolutionary events inferred along a species tree formed

by pruning the 116-mammal tree to the species present in the local alignment for each ncORF. We later ran the script `orf_bls.py` implemented by Chang *et al.*¹¹⁰ available at <https://github.com/gxelab/scripts>. First, we used the script to determine which extant and ancestral sequences had conserved ORF structure as defined above. Based on this information, we predicted an origination node for each ncORF. For ASR scores, the origination node was the most ancient ancestor in which the reconstructed sequence had a conserved ORF structure, whereas for the naive scores the origination node was the most recent ancestor whose descendants included all extant species in which the ORF structure was conserved. All BLS scores were computed as a ratio of a conserved branch length to a total branch length. For **ASR scores**, the conserved branch length was the sum of the lengths of all branches whose start and end both had conserved ORF structures, whereas for **naive scores** the conserved branch length was the branch length of the minimal tree connecting all extant species having conserved ORF structure. For **global scores**, the total branch length was the branch length of the complete 116-mammal tree, whereas for **local scores**, the total branch length was the branch length of the subset of the 116-mammal tree descended from the origination node. The naive global scores are the ones whose calculation is most similar to that of ORBLv, though with a much looser definition of what constitutes a conserved ORF structure.

As there is no established “true positive” set of constrained non-canonical ORFs, either on the peptide level or on the openness of the ORF, benchmarking these tools is challenging. We therefore evaluated the sensitivity of both scoring metrics against three evolutionary and functional readouts (see below). Across all three readouts, ORBLq scores relative to 116 placental mammals recover a higher proportion of ncORFs than global or local BLS, with or without ASR, at similar quantile cut-offs:

1. Preservation of translation

We initially assessed placental ORBL and BLS relationship with evolutionary conservation of Ribo-seq translation. To this end we investigated a subset of 337 cardiac ncORFs with conserved primate and rodent translation published in Ruiz-Orera *et al.* 2024⁸⁴ (**Figure 1c**). ORBL identifies a greater proportion of ncORFs that are translated across the five species (**Figure 2**). At the cutoff selected in our publication (ORBLq > 0.9), ORBLq identifies evolutionary conservation in 80% of ncORFs showing persistent conserved translation across primates and rodents.

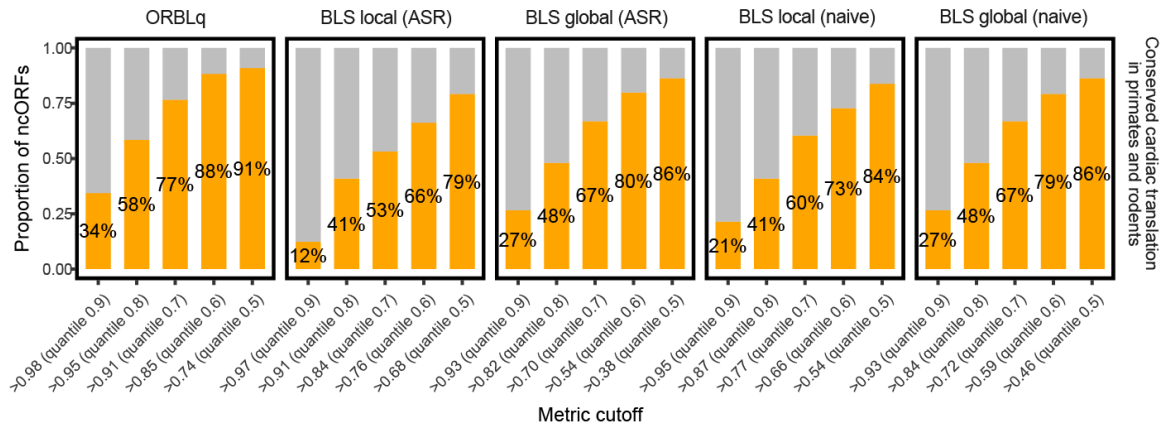


Figure 2. Bar plots displaying the percentage of ncORFs with persistent conserved translation in primates and rodents, among ncORFs having ORFness conservation above each of five quantile levels, as measured by placental ORBLq, local BLS and global BLS, with (ASR) and without (naive) using Ancestral Sequence Reconstruction alignments.

2. Protein-coding constraints

As a second strategy, we estimated selective protein-coding constraints acting on ncORFs using three metrics:

a) PhyloCSF: Scores were calculated using local alignments extracted from the 116-placental mammal whole-genome alignments. ncORFs with $\text{PhyloCSF}^{24} > 10$ were defined as under constrained protein-coding evolution, following previously published studies on ncORFs.

b) Nonsynonymous to synonymous substitution rate ratios (Ka/Ks): We calculated Ka/Ks for 7,264 ncORFs using codon-level multiple sequence alignments, restricting the analysis to 116 placental mammal species, following the approach used with ORBL. Ka/Ks ratios were estimated using the codeml program from the PAML¹¹² package under a single-ratio model (model = 0, NSsites = 0), which assumes a constant Ka/Ks across all branches. Codon frequencies were estimated using the F3×4 model, with transition/transversion rates and Ka/Ks parameters freely estimated from the data. Resulting Ka/Ks estimates were used as a measure of selective constraint acting on each ncORF across placental mammals. ncORFs with $\text{Ka/Ks} < 0.5$ were considered under strong purifying selection at the protein-coding level. Values between 0.5 and 1 can also indicate purifying selection, but were not used as a threshold due to the high variability and potential false positives in short ncORFs.

For ncORFs overlapping annotated CDSs in an alternative reading frame and with $\text{Ka/Ks} < 0.5$, we additionally assessed selective constraints using OLGene¹¹³. We included cases with at least one full codon overlap, excluding non-overlapping ncORFs and those overlapping only by one or two nucleotides. Overlaps derived from lncRNAs, PTs, uORFs, and dORFs were also included if they overlapped CDSs from alternative isoforms. For each overlapping ncORF, the appropriate alternative frame (+1 or +2) relative to the CDS was selected. Two selection statistics were estimated: dNN/dSN and dNS/dSS, which classify substitutions according to their synonymous

or nonsynonymous effects in both the ncORF and the overlapping CDS. dNN/dSN contrasts substitutions nonsynonymous in both frames (NN) with substitutions synonymous in the ncORF but nonsynonymous in the CDS (SN), controlling for selective constraints imposed by the canonical CDS. dNS/dSS contrasts substitutions nonsynonymous in the ncORF but synonymous in the CDS (NS) with substitutions synonymous in both frames (SS), measuring selection specifically in the ncORF while accounting for the overlapping CDS. To evaluate whether ncORFs are under stronger selective constraint than the overlapping CDS, we compared dNS/dSS values of the ncORFs with dNN/dSN values of the overlapping CDS.

c) PhyloP values: We assessed conservation by extracting 150 exonic bases upstream and downstream of each feature to determine whether observed nucleotide conservation by PhyloP^{113,114} (470way) was local to the feature or regional. Features were classified hierarchically: mean PhyloP ≤ 0.5 was marked as non-conserved; if either flank had a score >1 , the feature was considered regionally conserved, provided they did not overlap annotated CDSs. Features with non-conserved flanks and mean PhyloP between 0.5 and 1.5 were marked as weakly conserved. Features with mean PhyloP >1.5 were considered candidates for restricted nucleotide conservation.

We next measured the ability of ORBL to identify constrained protein-coding sequences based on these three metrics. Specifically, we compared it to the global and local BLS scores with (ASR) and without (naive) ASR implementation. Across all comparisons, placental ORBLq consistently detects a larger fraction of constrained sequences, independent of the specific evolutionary metric or conservation cutoff applied (quantiles ranging from 0.6 to 0.9, **Figure 3**). At the cutoff selected in our publication (> 0.9), ORBLq identifies conservation for 54-69% of constrained ncORFs (**Supplementary Table 11**). These results provide evidence that ORBL exhibits greater sensitivity in identifying ncORFs with constrained protein-coding sequences. However, we would like to emphasize that these comparisons should not be interpreted as a validation of ORBL, because each metric captures different patterns of sequence evolution. ORBL is designed to measure conservation and constraint of ORFness, reflecting the potential for translation; it does not examine protein-level constraint.

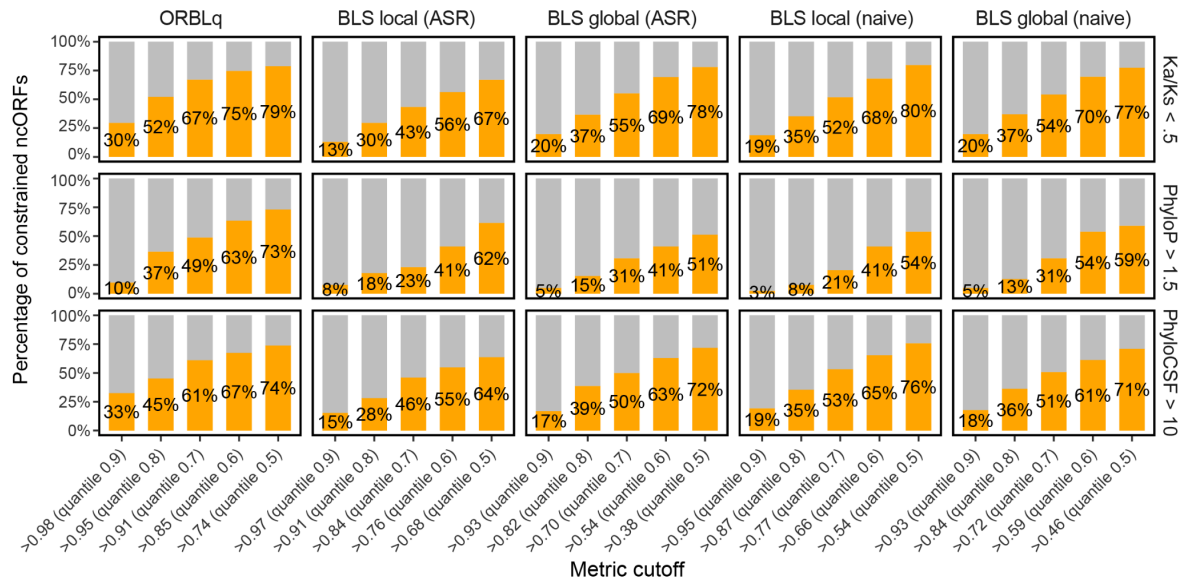


Figure 3. Bar plots displaying the percentage of ncORFs exhibiting evolutionary protein-coding constraint as assessed by Ka/Ks (>0.5, overlapping CDS cases corrected by in-frame overlap), PhyloP (>1.5, corrected by flanking regions) and PhyloCSF (>10), among ncORFs having ORFness conservation above each of five quantile levels, as measured by placental ORBLq, local BLS and global BLS, with (ASR) and without (naive) using Ancestral Sequence Reconstruction alignments.

3. CRISPR cumulative genetic dependency

Lastly, we used the CRISPR screen data of 8 cell lines for 2,139 ncORFs to compare ORBLq to BLS, after excluding 57 cases having mixed biotype in Gencode V45. Cumulative genetic dependency was defined as the number of cell lines in which an ncORF was found to affect cell fitness (chronos score < -0.5). Placental ORBLq shows greater sensitivity in identifying ncORFs with cumulative genetic dependency (**Figure 4**). Using the cutoff established in our study (ORBLq > 0.9), ORBLq captures 44% of ncORFs with phenotype in 1-4 cell lines and 58% of those with phenotype in 5-8 cell lines.

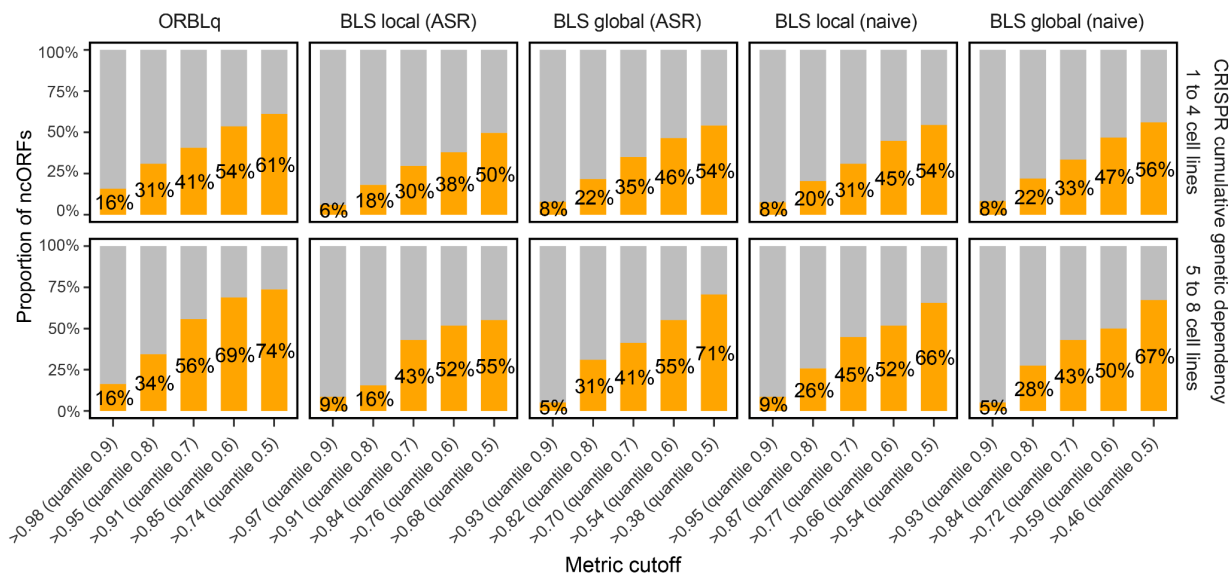


Figure 4. Bar plots displaying the percentage of ncORFs with CRISPR cumulative genetic dependency across 1-4 or 5-8 cell lines, among ncORFs having ORFness conservation above each of five quantile levels, as measured by placental ORBLq, local BLS and global BLS, with (ASR) and without (naive) using Ancestral Sequence Reconstruction alignments.

Other applications

Finally, we note that in addition to measuring constraint on the ORFness of ORFs whose translation is functional but that do not encode functional proteins, ORBL can complement tools such as PhyloCSF, which focus on amino acid conservation but that do not consider reading frame or stop codon conservation, to help in determining if an ORF is protein coding. Something similar to ORBL might also be helpful in distinguishing cases where a 5' extension of a coding ORF is coding or the region from the stop codon of a coding ORF to the next in-frame stop codon is translated through stop-codon readthrough. We hope to provide variants of ORBL to handle these cases in the future.

Availability of ORBL

A utility for computing ORBLv and ORBLq is available at https://github.com/iljungr/ORBL_tools. The ORBL utility can calculate ORBLv conservation scores for any alignment available in CodAlignView (listed here <https://data.broadinstitute.org/compbio1/cav.php?Alnsets>), as well as some subclades of the CodAlignView alignments. These include many other reference species in addition to human (mouse, rat, zebrafish, fly, worm, etc). The full list of alignments available for ORBLv, as well as which of them are available for ORBLq, can be found using the --alignmentSets option to the orbl utility. Currently, ORBLq constraint scores are only available for human because our method for constructing the list of “untranslated” ORFs used by ORBLq depends on excluding ORFs that overlap the ncORF list; when lists of ncORFs of similar quality become available we intend to make ORBLq available for other species as well.

(5) Structural predictions of ncORF-derived proteins

Motif and structure prediction are important steps in understanding the potential functionality of identified protein-coding regions. However, these analyses present significant challenges for ncORFs, which are often short in length and evolutionarily young. For instance, deep learning tools for structural predictions can produce inflated confidence scores for shorter sequences¹¹⁵. Moreover, these tools were primarily trained on well-conserved, globular proteins, leading to suboptimal performance for poorly conserved proteins^{84,116,117}. We examined the length distributions across our 7,264 Tier-classified ncORFs. The ncORFs in Tiers 1A and 2A—supported by ribosome profiling and mass spectrometry—showed significantly longer sequences than those in Tiers 1B, 2B, and 4 (**Extended Data Figure 10a**).

Impact of protein length and evolutionary constraint on ncORFs

We next assessed protein structure prediction confidence using AlphaFold3¹¹⁸, ESMFold¹¹⁹, and OmegaFold¹²⁰, focusing on the average per-residue predicted Local Distance Difference Test (pLDDT) scores across 7,264 ncORFs (**Supplementary Table 14**). Surprisingly, Tiers 1A and 2A—despite their strong experimental protein evidence—showed lower average pLDDT scores than other tiers, likely due to their longer sequence lengths (**Extended Data Figure 10b-d**). To evaluate potential length bias, we permuted the amino acid sequences of the 581 high-confidence ncORFs (pLDDT >90), generating five shuffled versions per sequence. Remarkably, 62% and 93% of these shuffled sequences still scored above 90 and 80, respectively (**Extended Data Figure 10e**), revealing that structure predictors can assign spuriously high confidence scores to short, poorly conserved, potentially non-functional sequences. Accordingly, we observed a clear linear relationship between ncORF length and predicted pLDDT values, with shorter ncORFs exhibiting higher pLDDT scores (**Extended Data Figure 10g**).

We only found 36 ncORFs in which none of the five shuffled sequence controls reached a pLDDT score above 90 (**Extended Data Figure 10g**). Despite this, the majority of shuffled sequences still produced scores near this threshold. Notably, for only six ncORFs, all shuffled variants yielded pLDDT scores below 80, indicating higher structural confidence (**Extended Data Figure 10h**). These six cases predominantly consist of predicted alpha-helical structures. Among them, only c19norep157 displays a well-defined complex structural fold. As previously reported by us²³, c19norep157 corresponds to a dORF that, in humans, has undergone truncation from a full protein and is now translated as an independent, pseudogenized ncORF.

Overall, these findings underscore the limitations of current structure prediction models, which are trained on long, evolutionarily conserved proteins, when applied to short ncORFs having evolutionarily unconstrained amino acid sequences. As such, structural predictions for ncORFs—especially short ones—should be interpreted with caution.

Analysis of disordered regions and linear motifs

We next identified eukaryotic Linear Motif (ELM) matches from the ELM database¹²¹ within predicted disordered regions by DSSP¹²² (**Supplementary Table 14**). We used the output from the previously predicted AlphaFold3 structures. We also searched for functional domains using the ScanProsite tool¹²³.

ncORFs in Tiers 1A and 2A exhibited higher fractions of disordered regions and greater densities of ELMs relative to Tiers 1B, 2B, and 3 (**Extended Data Figure 10i+j**). ELM densities in Tier 1A/2A were comparable to Tier 4. Only 76 proteins had one or more PROSITE hits, with the highest proportions observed in Tier 1A (12.5%) and Tier 4 (1.0%). These findings suggest that disordered regions and linear motifs may represent relevant functional features in certain ncORFs, particularly those with strong protein evidence.

Given the limitations of current structure prediction tools for short and poorly conserved sequences, and the observed biases in pLDDT scoring, we opted not to include structure-based functional interpretations for ncORFs in the main manuscript. Instead, we provide motif and disorder-based analyses that are more applicable to the evolutionary and structural nature of these proteins. These analyses strengthen the biological relevance of Tier 1A and 2A ncORFs and provide complementary evidence.

(6) *De novo* gene annotation

The annotation of *de novo* protein-coding genes - i.e., those originating in specific lineages and primarily from non-coding DNA - remains a major challenge. This is particularly true for short, lineage-specific ORFs, as signatures of protein-level constraint are harder to detect in shallow multispecies alignments. Furthermore, annotation projects remain circumspect when annotating proteins that lack these signatures, so very few are presently recognised as canonical. Nonetheless, *de novo* genes can contribute to cellular biology via functional innovation across taxa, and be relevant to human disease^{35,36,124}.

We therefore compared our set of translated ncORFs to four published lists of *de novo* genes, yielding a total of 192 genes, of which 147 were unique (**Figure 1, Supplementary Table 11**):

- **Xie et al. 2012**¹²⁵ – 24 genes
- **An et al. 2023**¹²⁶ – 74 genes, including a duplicated gene id.
- **Broeils et al. 2023**³⁵ – 82 genes; a recent, manually curated list of *de novo* protein-coding genes
- **Ruiz-Orera et al. 2024**⁸⁴ – 12 genes; a recent set of cardiac *de novo* protein-coding genes whose evolutionary status was confirmed by the absence of Ribo-seq translation in other species.

Among the 147 unique *de novo* genes, 48 are not represented in GENCODE v42 annotations. Of the 99 annotated genes, 84 are classified as lncRNAs, 4 as pseudogenes, and 11 as protein-coding genes with the same CDS. The *de novo* status of the 11 presumed protein-coding genes has been recently debated¹²⁷. To ensure accuracy, we manually curated their conservation, validating the *de novo* status of six genes while discarding five. Of these, *MYEOV* and *IL32* have canonical protein-coding evidence according to HPP standards.

Based on this curation, we identified 81 ncORFs mapping to 35 *de novo* genes from these publications: 79 ncORFs in 34 lncRNAs and two uORFs encoded by the protein-coding *de novo* gene *MYEOV*. Among these, four lncRNA-ORFs are classified as 1B, supported by multiple HLA peptide evidence; notably, the lncRNA *MOCS2-DT* is the only ncORF in this set currently manually curated as a peptide. An additional six cases are in scope for peptidein annotation following our efforts here. Further work by our project and parallel efforts from others will likely lead to reconciliation in the future.

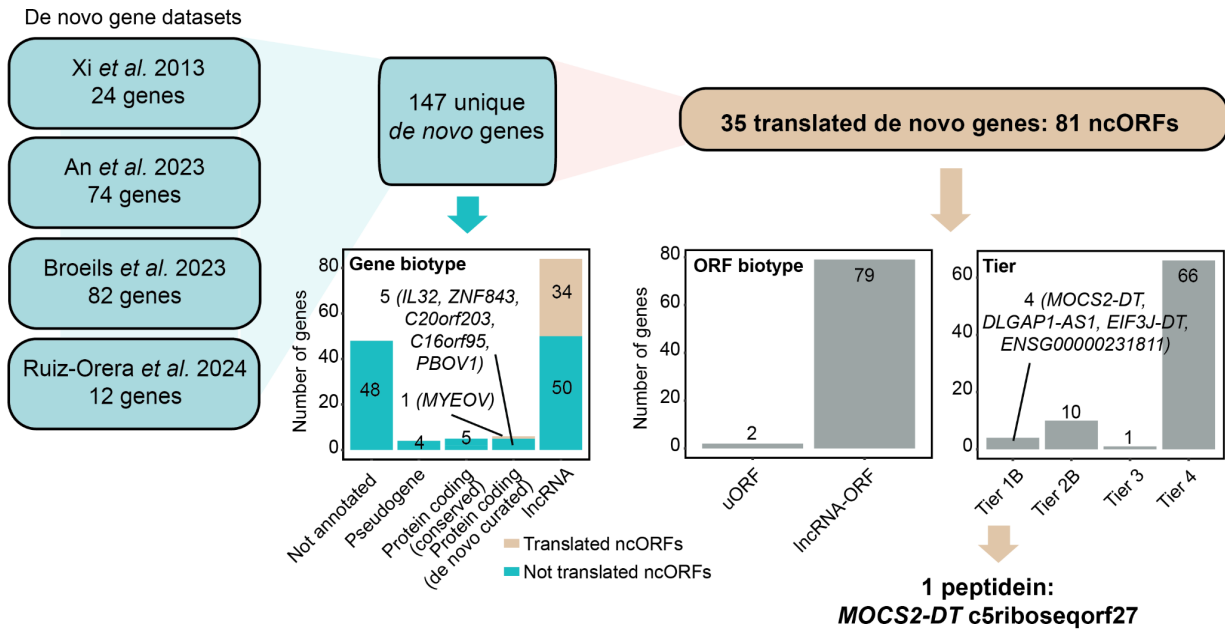


Figure 1. Schematic of 147 unique *de novo* genes compiled from four independent studies (Cai *et al.*, 2013; An *et al.*, 2023; Broeils *et al.*, 2023; Ruiz-Orera *et al.*, 2024). GENCODE v42 gene biotypes are indicated, including 48 unannotated genes and five protein-coding genes reclassified as conserved following manual curation. *De novo* protein-coding gene names are provided. Mapped *de novo* gene IDs were matched to the host genes in the ncORF catalog, resulting in 35 unique genes encoding 81 ncORFs. ORF biotype and classifications are shown, along with the gene names of four ncORFs in 1B.

References

95. Deutsch, E. W. *et al.* Universal Spectrum Identifier for mass spectra. *Nature Methods* **18**, 768–770 (2021).
96. Paten, B. *et al.* Cactus: Algorithms for genome multiple sequence alignment. *Genome Res.* **21**, 1512–1528 (2011).
97. Christmas, M. J. *et al.* Evolutionary constraint and innovation across hundreds of placental mammals. *Science* **380**, eabn3943 (2023).
98. Roberts, R. M. *et al.* Syncytins expressed in human placental trophoblast. *Placenta* **113**, 8–14 (2021).
99. Dyring-Andersen, B. *et al.* Spatially and cell-type resolved quantitative proteomic atlas of healthy human skin. *Nat. Commun.* **11**, 5587 (2020).
100. Vu, T. H., Chuyen, N. V., Li, T. & Hoffman, A. R. Loss of imprinting of IGF2 sense and antisense transcripts in Wilms' tumor. *Cancer Res.* **63**, 1900–5 (2003).
101. Okutsu, T. *et al.* Expression and Imprinting Status of Human PEG8/IGF2AS, a Paternally Expressed Antisense Transcript from the IGF2 Locus, in Wilms' Tumors1. *J. Biochem.* **127**, 475–483 (2000).
102. Zou, X.-D. *et al.* Long noncoding RNA ARRDC1-AS1 is activated by STAT1 and exerts oncogenic properties by sponging miR-432-5p/PRMT5 axis in glioma. *Biochem. Biophys. Res. Commun.* **534**, 511–518 (2021).
103. Abadi, M. *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv* (2016) doi:10.48550/arxiv.1603.04467.
104. Cybenko, G. Approximation by superpositions of a sigmoidal function. *Math. Control, Signals Syst.* **2**, 303–314 (1989).
105. Hinnebusch, A. G. Evidence for translational regulation of the activator of general amino acid control in yeast. *Proc. Natl. Acad. Sci.* **81**, 6442–6446 (1984).
106. Dever, T. E., Ivanov, I. P. & Hinnebusch, A. G. Translational regulation by uORFs and start codon selection stringency. *Genes Dev.* **37**, 474–489 (2023).
107. Dever, T. E., Ivanov, I. P. & Sachs, M. S. Conserved Upstream Open Reading Frame Nascent Peptides that Control Translation. *Annu. Rev. Genet.* **54**, 1–28 (2020).
108. Kheradpour, P., Stark, A., Roy, S. & Kellis, M. Reliable prediction of regulator targets using 12 Drosophila genomes. *Genome Res.* **17**, 1919–1931 (2007).
109. Mudge, J. M. *et al.* Discovery of high-confidence human protein-coding genes and exons by whole-genome PhyloCSF helps elucidate 118 GWAS loci. *Genome Res.* **29**, 2073–2087 (2019).
110. Chang, Y. *et al.* Evolutionary remodeling of non-canonical ORF translation in mammals. *eLife* (2025) doi:10.7554/elife.109128.1.
111. Löytynoja, A. Phylogeny-Aware Alignment with PRANK and PAGAN. *Methods Mol. Biol.* **2231**, 17–37 (2020).
112. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
113. Nelson, C. W., Arden, Z. & Wei, X. OLGene: Estimating Natural Selection to Predict Functional Overlapping Genes. *Mol. Biol. Evol.* **37**, 2440–2449 (2020).
114. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
115. Monzon, V., Haft, D. H. & Bateman, A. Folding the unfoldable: using AlphaFold to explore spurious proteins. *Bioinform. Adv.* **2**, vbab043 (2022).
116. Aubel, M., Eicholt, L. & Bornberg-Bauer, E. Assessing structure and disorder prediction tools for de novo emerged proteins in the age of machine learning. *F1000Research* **12**, 347 (2023).
117. Middendorf, L. & Eicholt, L. A. Random, de novo, and conserved proteins: How structure and disorder predictors perform differently. *Proteins: Struct., Funct., Bioinform.* **92**, 757–767 (2024).
118. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
119. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
120. Wu, R. *et al.* High-resolution de novo structure prediction from primary sequence. *bioRxiv* 2022.07.21.500999 (2022) doi:10.1101/2022.07.21.500999.
121. Gouw, M. *et al.* The eukaryotic linear motif resource – 2018 update. *Nucleic Acids Res.* **46**, gkx1077- (2017).
122. Touw, W. G. *et al.* A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.* **43**, D364–D368 (2015).
123. Castro, E. de *et al.* ScanProsite: detection of PROSITE signature matches and ProRule-associated functional

and structural residues in proteins. *Nucleic Acids Res.* **34**, W362–W365 (2006).

124. Xia, S., Chen, J., Arsala, D., Emerson, J. J. & Long, M. Functional innovation through new genes as a general evolutionary process. *Nat. Genet.* **57**, 295–309 (2025).

125. Xie, C. *et al.* Hominoid-Specific De Novo Protein-Coding Genes Originating from Long Non-Coding RNAs. *PLoS Genet.* **8**, e1002942 (2012).

126. An, N. A. *et al.* De novo genes with an lncRNA origin encode unique human brain developmental functionality. *Nat. Ecol. Evol.* **7**, 264–278 (2023).

127. Xiao, C. *et al.* Reply to: Identification of old coding regions disproves the hominoid de novo status of genes. *Nat. Ecol. Evol.* **8**, 1831–1834 (2024).