Article

Expanding the human proteome with microproteins and peptideins

https://doi.org/10.1038/s41586-026-10459-x

Received: 7 September 2024

Accepted: 27 March 2026

Published online: 06 May 2026

Open access

Check for updates

Eric W. Deutsch[1,60], Leron W. Kok[2,3,60], Jonathan M. Mudge[4,60], Cristian F. Valls[5,6,60], Irwin Jungreis[7,8,60], Jorge Ruiz-Orera[9], Zhi Sun[1], Ulrike Kusebauch[1], Ivo Fierro-Monti[4,10], Jennifer G. Abelin[8], M. Mar Alba[11,12], Julie L. Aspden[13], Sreejan Bandyopadhyay[14], Kaushik Banerjee[5,6], Pavel V. Baranov[15], Ariel A. Bazzini[16,17], Francis Bourassa[18], Elspeth A. Bruford[19], Lorenzo Calviello[20], Steven A. Carr[8], Anne-Ruxandra Carvunis[21,22,59], Sonia Chothani[23,24], Jim Clauwaert[5], Kellie Dean[15], Pouya Faridi[25,26], Adam Frankish[4], Amy Goodale[8], Thomas Green[8], Norbert Hubner[9,27,28,29], Nicholas T. Ingolia[30], Manolis Kellis[7,8], Michele Magrane[4], Maria Jesus Martin[4], Thomas F. Martinez[31,32,33], Gerben Menschaert[34], Uwe Ohler[35,36], Sandra Orchard[4], Alisa Potter[2,3,37], Owen J. L. Rackham[38], Matthew G. Rees[8], David E. Root[8], Jennifer A. Roth[8], Xavier Roucou[39], Fernando J. Sialana[14], Sarah A. Slavoff[40,41,42], Michał I. Świrski[43], Jack A. S. Tierney[4], Félix-Antoine Trifiro[18], Eivind Valen[44], Valeriia Vasylieva[18], Aaron Wacholder[21,22,59], Shengbo Wang[4], Li Wang[8], Jonathan S. Weissman[45,46,47,48], Wei Wu[49,50], Zhi Xie[51], Jyoti S. Choudhary[14], Michal Bassani-Sternberg[52,53,54], Juan Antonio Vizcaíno[4], Nicola Ternette[55,56], Marie A. Brunet[18,57,58], Robert L. Moritz[1,61]✉, John R. Prensner[5,6,61]✉ & Sebastiaan van Heesch[2,3,61]✉

A major scientific drive is to characterize the protein-coding genome, which is a primary basis for studying human health. But the fundamental question remains of what has been missed in previous analyses. Over the past decade, the translation of non-canonical open reading frames (ncORFs) has been observed across human cell types and disease states[1–3], with major implications for biomedical science. However, a key gap in knowledge has been which ncORFs produce small microproteins or alternative protein molecules that contribute to the human proteome. Here we report the collaborative efforts of the TransCODE Consortium[4] to produce a consensus landscape of protein-level evidence for ncORFs. We show that about 25% of a set of 7,264 ncORFs gives rise to detectable peptides in a large-scale analysis of 95,520 proteomics experiments. We develop an annotation framework for ncORF-encoded microproteins as human proteins and codify the new conceptual model of 'peptideins' as microproteins that have indeterminate potential as functional proteins. To probe the biological implications of peptideins, we create an evolutionary analysis approach, termed ORF relative branch length (ORBL), and determine that evolutionary constraint is common and associates with observation of ncORF-derived peptides. We then characterize a pan-essential cellular phenotype for one peptidein from the *OLMALINC* long non-coding RNA. Overall, we generate public research tools supported by GENCODE and PeptideAtlas and advance biomedical discovery for understudied components of the human proteome.

Whether the human genome encodes substantially more than the approximately 19,500 canonical protein-coding genes has sparked a spirited debate in recent years. Protein-coding genes are the bedrock of biomedical investigations, including the overwhelming majority of drug development programmes. Therefore, any wholesale addition of protein-coding genes creates ripple effects across human bioscience.

Curation and maintenance of these genes is the task of reference annotation projects, such as Ensembl-GENCODE (hereafter, GENCODE) and UniProtKB/Swiss-Prot (hereafter, UniProt), the work of which builds on the Human Genome Project. Although the number of canonical protein-coding genes has been refined continuously over time, it was felt to be largely stable until recent evidence of translation of small polypeptide or protein sequences from thousands of unannotated ncORFs, which are variably referred to as microproteins, small ORF-encoded peptides (SEPs) or micropeptides (hereafter, microproteins). These ncORFs and their encoded polypeptides are now reported widely as part of a 'dark proteome', and their promise to advance medical science is manifested in their contributions to the genetic basis of disease[5,6], mechanisms of cancer biology[7], and cancer-restricted and HLA-presented cryptic antigens targetable by immunotherapy[8].

A list of affiliations appears at the end of the paper.

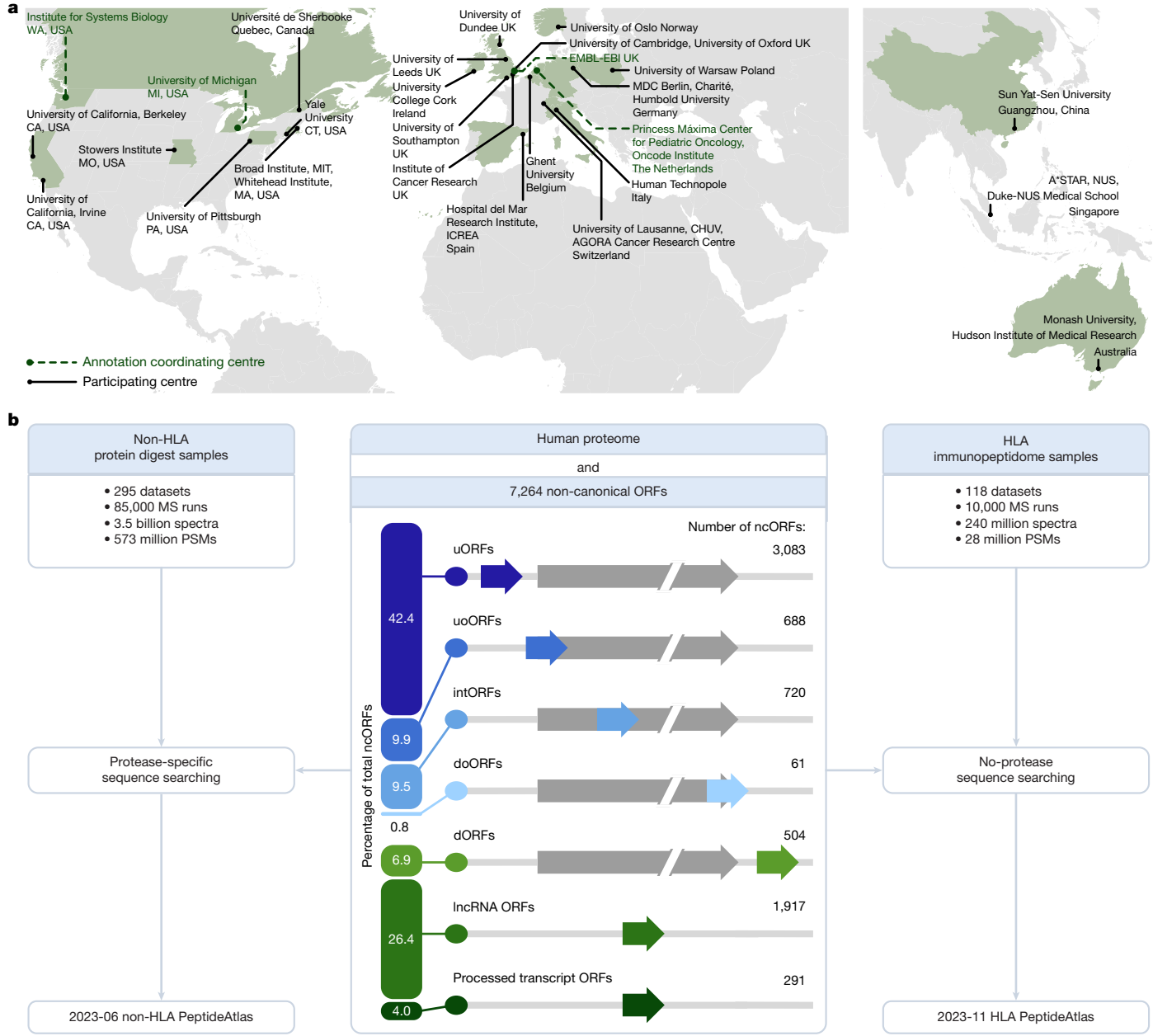


Fig. 1 | Overview of the centres participating in the annotation effort and the PeptideAtlas framework for protease-digested sample (mostly trypsin) MS and immunopeptidomics builds. a, Map of the participating institutions included in the annotation effort. Coordinating centres are highlighted. **b**, Overview of the datasets included in the non-HLA and HLA builds. The biotypes of the 7,264 ncORFs are shown in the middle. dORF, downstream ORF.

Yet, the number of ncORFs that represent true protein-coding genes has been a subject of controversy. To date, few microproteins have been annotated as canonical proteins by reference annotation catalogues (such as GENCODE and UniProt) because their uncertain structure and low evolutionary constraint complicate their classification as conventional proteins. At the same time, peptides resulting from cryptic translation have become an emerging area for therapeutic targeting discovery in cancer and other disorders[2,8–10].

In 2022, we launched the international TransCODE Consortium[4] with the goal of defining standards for the reference annotation of ncORFs and their encoded microproteins, including members of GENCODE[11], PeptideAtlas[12], the Human Proteome Organization-Human Proteome Project (HUPO-HPP)[13] and the HUPO-Human ImmunoPeptidome Project (HUPO-HIPP)[14] (Fig. 1a). Here we develop a pathway for microproteins to be annotated as reference human proteins when annotation-quality proteomics support is present. To bring formal reference gene annotation status to less-well-characterized microproteins, we introduce 'peptidein' as a classification scheme, recognized by our consortia, to exist alongside conventional proteins. To illustrate that further characterization of a peptidein may elevate its classification, we use functional genomics and evolutionary constraint to pinpoint examples that exhibit a signature consistent with a protein-coding gene. Lastly, we propose a research agenda based on consensus among the multiconsortium group, intended to guide future efforts to bring ncORFs, microproteins and peptideins from research discoveries to biological, societal and biomedical impact through ongoing standardized annotation.

A microprotein annotation workflow

We sought to provide reference annotation-quality evidence for ncORF-encoded microproteins that meet criteria as human proteins.

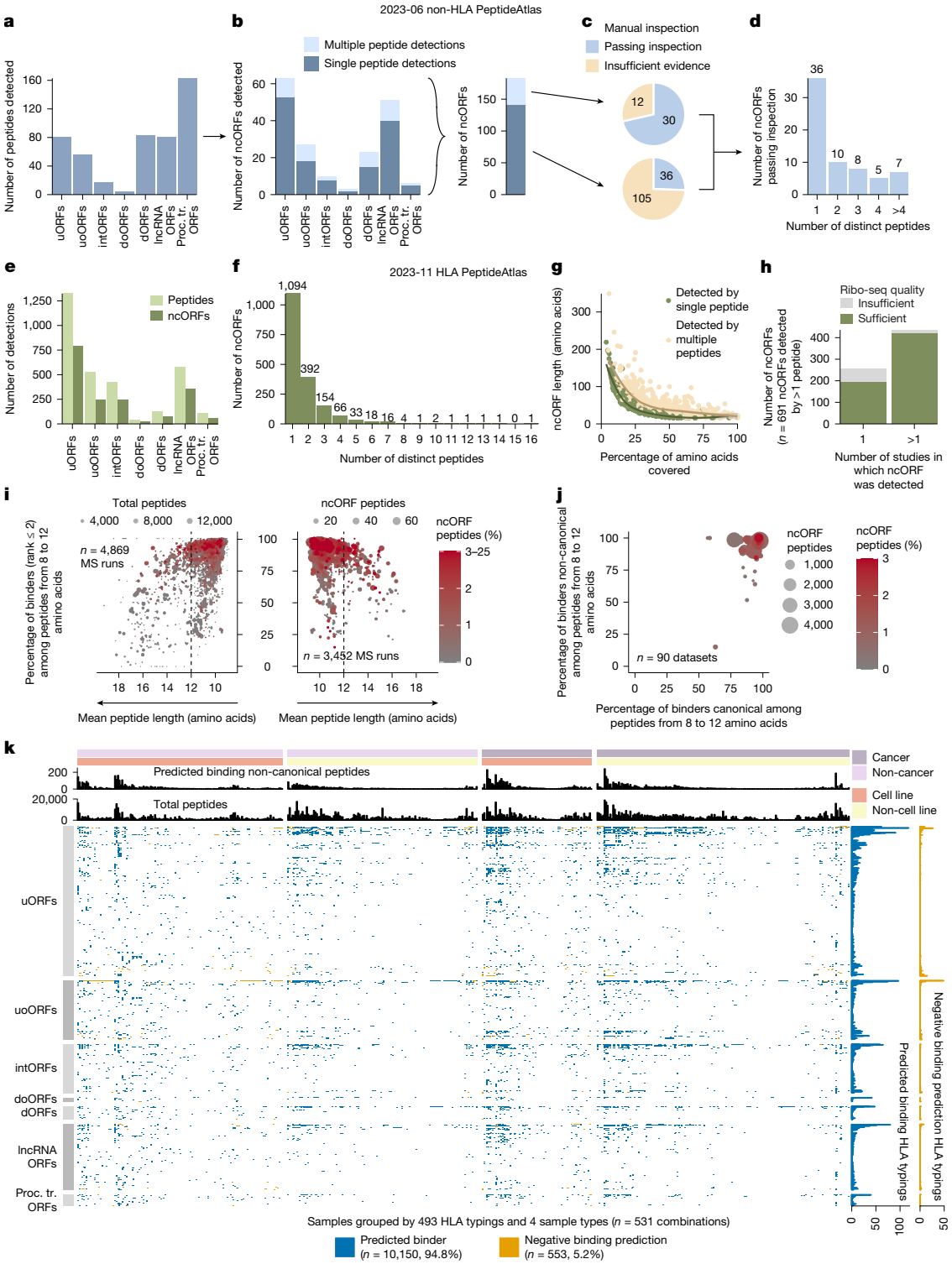


Fig. 2 | The non-HLA and HLA PeptideAtlas builds. a–**d**, The 2023-06 non-HLA PeptideAtlas analysis. **a**, Detected peptides ($n = 484$) categorized by biotype. **b**, The number of ncORFs detected by one or more peptides categorized per biotype or summed to a total. $n = 183$. **c**, The number of ncORFs with peptides passing quality filters for ncORFs detected by multiple peptides (top) or by a single peptide (bottom). **d**, The number of ncORFs passing inspection. **e**–**k**, The 2023-11 HLA PeptideAtlas analysis. **e**, The number of distinct detected peptides ($n = 3,116$) and ncORFs ($n = 1,785$) grouped by biotype. Proc. tr. ORFs, processed transcript ORFs. **f**, The number of distinct peptides by which an ncORF was detected. **g**, The percentage of the total ncORF sequence covered by HLA peptides, plotted against ncORF length. Confidence intervals of fitted lines are shown in grey. **h**, The number of ncORFs for which the Ribo-seq data quality was sufficient or insufficient. **i**, Visualization of the correlation between mean peptide length and the percentage of predicted binders (8–12 amino acids). On the left, the dot size indicates the total number of peptides per each of 4,869 MS runs with known HLA typing. On the right, the dot size shows the count of ncORF-derived peptides across 3,452 MS runs with at least one detected ncORF-derived peptide. One outlier MS run (average length, 22.75 amino acids) is not shown. **j**, Contrast of the percentage of predicted canonical and non-canonical binders (NetMHCpan rank ≤ 2) per dataset. **k**, NetMHCpan ncORF binding verifications portioned by sample type. Top, detected non-canonical peptides predicted to bind to HLA alleles within a typing and the total distinct peptides associated with it. Right, the total counts of positive and negative predictions for each peptide.

To do this, we expanded the purview of the PeptideAtlas platform, which is the basis for certification of human protein-coding genes through HUPO and the HPP. Using the ProteomeXchange mass spectrometry (MS) data repository, we assembled the human non-HLA PeptideAtlas 2023-06 build (constituting 3.5 billion protease-digested MS/MS spectra) and the Human HLA PeptideAltas 2023-11 build (constituting 240 million MS/MS spectra from human leukocyte antigen (HLA) datasets). These were used to query 7,264 ncORFs supported by GENCODE (Fig. 1b; Methods).

To ensure high-quality detections, we set a decoy-estimated false-discovery rate (FDR) of <0.1% at the protein level and required protein identifications to adhere to the guidelines established by the HUPO-HPP, that is, two distinct, uniquely mapping peptides of length 9 or more residues and a minimum protein coverage of 18 residues[15]. This approach led to a peptide-level FDR of 0.0009% for the non-HLA build and 0.0041% for the HLA build (Methods), which is more conservative than many studies[16] because annotation-level proteomics evidence requires higher stringency. As expected, we observed near saturation in the power to identify canonical human proteins in the non-HLA build, adding around 1 protein per million peptide spectrum matches (PSMs) with incorporation of additional datasets (Extended Data Fig. 1a–d).

Microproteins in digest MS/MS datasets

We first explored which of the 7,264 GENCODE ncORFs encoded detectable microproteins using conventional peptide data (96.3% of experiments are digested with trypsin; Supplementary Table 1). We found 484 peptides passing FDR thresholds that map to 183 out of the 7,264 ncORFs (around 2.5%) (Fig. 2a,b and Supplementary Tables 2 and 3). Additional search engines only marginally increased the total number of identifications at a considerable increase in computational expense (Extended Data Fig. 2a,b), and using more permissive FDR thresholds preferentially increased false positives (Extended Data Fig. 2c,d).

As reference annotation efforts require certainty in proteomic identifications, we next manually inspected MS spectra and ribosome-sequencing (Ribo-seq) data for all 183 ncORFs in the non-HLA build to eliminate false positives. We validated 30 out of 42 ncORFs with two unique supporting peptides but only 36 out of 141 ncORFs with one verified peptide (Fig. 2c,d, Extended Data Fig. 2e and Supplementary Tables 2 and 3). We also validated the spectra of 29 of 30 PSMs by matching synthetic peptide spectra with the original spectra (Supplementary Table 4). We further confirmed the endogenous expression of ncORF-derived microproteins by parallel reaction monitoring (PRM) with isotopically labelled synthetic peptides spiked into cultured cell lysate tryptic digests (Extended Data Fig. 3 and Supplementary Figs. 1 and 2). We also found microproteins exhibiting various post-translational modifications (PTMs) (Supplementary Tables 2 and 5, Extended Data Fig. 2f and Supplementary Results).

We then hypothesized that microproteins may face an adverse bias with tryptic peptide detection driven by their small size. Indeed, using a manually curated set of small GENCODE proteins[17], we found that only 2 out of 36 known proteins under 50 amino acids (5.6%) satisfy benchmarks for HUPO-HPP verification. By comparison, in a single dataset[18], incorporation of alternative proteases increased both the total number of microprotein identifications and their coverage (Extended Data Fig. 2g), although the overall numbers were small.

Microproteins as HLA-I-presented peptides

We next searched for ncORF-encoded microproteins in the Human HLA PeptideAtlas 2023-11 build. Overall, we found 3,116 peptides mapping to 1,785 out of 7,264 Ribo-seq ncORFs (24.6%) (Fig. 2e,f, Extended Data Fig. 4a and Supplementary Tables 6 and 7). Almost all of these peptides (2,937 out of 3,116; 94.3%) were found presented by HLA class-I (HLA-I) alone, with few observed in HLA class-II (HLA-II) datasets (Extended Data Fig. 4a–f). This indicates that peptides produced from ncORF-encoded microproteins are most often sourced from the intracellular pool of protein translation products and are less likely from extracellular sources, in contrast to canonical proteins[9,19,20].

Several factors influenced microprotein detectability in HLA data, including length (increased peptide detection with longer coding sequences) and the position of the ncORF in the translated RNA molecule (Extended Data Fig. 5a–d). However, we did not observe significant differences in microprotein detectability between cancer or non-cancer datasets, and their distribution between cancer or non-cancer samples was not influenced by peptide mass, hydrophobicity (Kyte–Doolittle) or isoelectric point (Extended Data Fig. 5e,f).

To extend confidence for HLA build identifications, we manually inspected 859 HLA-I MS spectra and 691 matching Ribo-seq profiles, focusing the latter on ncORFs with at least two uniquely mapping peptides in the HLA build (Fig. 2g). Overall, we validated the Ribo-seq signal in 88.7% (613 out of 691) of ncORFs and observed that ncORFs found in multiple published studies exhibited a higher rate of validation (96.1% (419 out of 436) verified) compared with ncORFs reported in a single study (76.1% (194 out of 255) verified) (Methods and Fig. 2h). We conclude that a great preponderance of HLA-I peptide identifications for ncORF-encoded microproteins is well supported.

HLA-I-presented peptide characteristics

The concordance between HLA-I peptide binding predictions and immunopeptidomics data can be used to augment peptide identification of rare source proteins, such as microproteins. We therefore curated the HLA types of samples used in 4,870 of the 6,479 MS runs (Supplementary Table 8) and collectively assessed in silico HLA-I binding predictions for 2,711 microprotein peptides that fell within a required length of 8–12 amino acids. For 4,308 out of the 4,870 (88.5%) analysed HLA-I MS runs, >70% of detected HLA-I peptides were predicted as binders (percentage rank score <2%) (Fig. 2i), and microprotein peptides were equally likely to have predicted binding to annotated HLA types compared to canonical proteins (Fig. 2j). As cells have up to six classical HLA-I alleles, we next used the binding predictions to assign each microprotein peptide to the most likely HLA allele reported or predicted for a dataset. We then checked the individual binding predictions per HLA typing, ORF biotype and source material. We observed a strong concordance (94.8%) between predictions and detected peptides across ORF biotypes and independent of the source material (for example, cancerous or non-malignant cell lines or tissues) (Fig. 2k).

We next examined whether there are key determinants that would make a certain microprotein, or part of the microprotein, more likely to be detected in HLA-I immunopeptidomics data. Overall, we found that the amino acid sequence, length and tissue expression pattern all have a role in the ability to observe microprotein peptides in HLA-I datasets (Fig. 3a–e, Extended Data Fig. 6a–h, Supplementary Tables 9 and 10 and Supplementary Results). Notably, the isoelectric point was increased for detected microproteins compared with undetected microproteins, while detected canonical proteins displayed the opposite pattern (Fig. 3a). However, we found no difference in overall sequence hydrophobicity (Fig. 3a), in contrast to recent reports[21]. While C-terminal hydrophobicity varied across ncORF biotypes—probably reflecting sequence context (Fig. 3b)—these differences did not account for microprotein detectability, as detected and undetected microproteins showed similar C-terminal hydrophobicity (Extended Data Fig. 6c,d). We did notice that the C-terminal parts of microproteins were preferentially sourced for HLA presentation, and that this enrichment is stronger than for canonical proteins (20.3-fold versus 7.2-fold, respectively, $P = 9.8 \times 10^{-9}$, Fisher's exact test) (Fig. 3c).

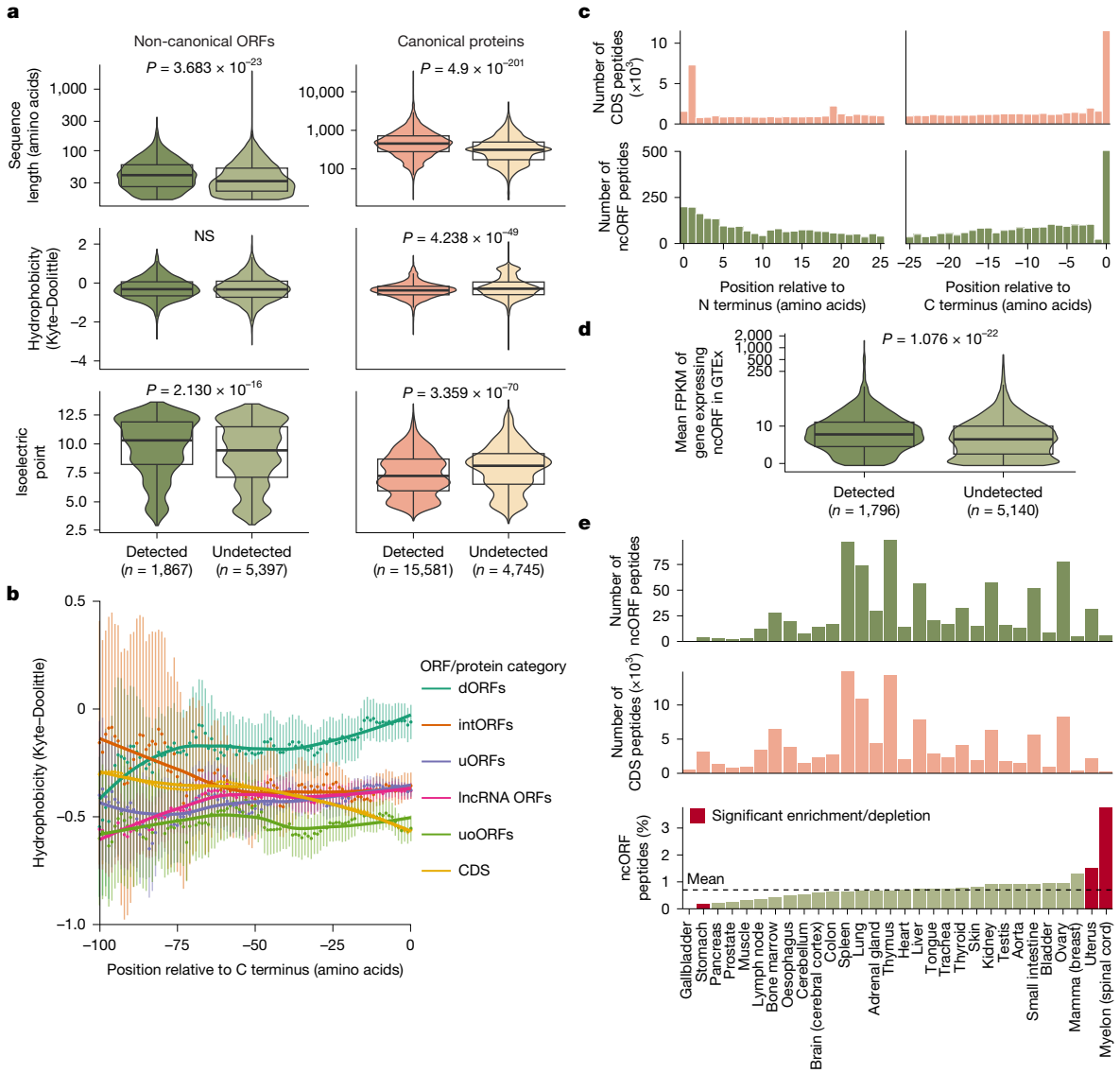


Fig. 3 | Determinants of ncORF peptide detection in the HLA build. **a**, Comparison of different sequence properties between detected and undetected ncORFs and canonical proteins: sequence length, hydrophobicity based on the Kyle–Doolittle scale and the isoelectric point. The box plots show the median (centre line) and the 25th (Q1) and 75th (Q3) percentiles (box limits); the whiskers extend to the most extreme values within 1.5× the interquartile range from Q1 and Q3. *P* values were determined using two-sided Wilcoxon rank-sum tests with Holm–Bonferroni correction. NS, not significant. **b**, Hydrophobicity profiles by ORF biotype ($n = 6{,}912$ ncORFs; 20,326 CDS). Each dot represents the mean hydrophobicity of the amino acid at a given position plus the preceding 14 residues per ncORF biotype or CDS. The vertical bars indicate the 95% confidence intervals of the fitted lines. doORFs ($n = 61$) and processed transcript ORFs ($n = 291$) were excluded due to low abundance. **c**, The distribution of detected peptide positions within CDS (top) and ncORF (bottom) sequences. The left histograms show the distance from the start codon to peptide start (42,787 CDS peptides; 2,344 ncORF peptides). The right histograms show the distance from peptide end to the last amino acid (33,095 CDS peptides; 2,199 ncORF peptides). **d**, Expression levels of detected versus undetected ncORFs. The *y* axis shows mean Genotype-Tissue Expression (GTEx) FPKM values of genes expressing ncORFs (pseudo-log scale). The box plots were generated as described in **a**. Significance was assessed using a two-sided Wilcoxon rank-sum test (unadjusted). **e**, HLA ligand atlas data by tissue. Top, counts of ncORF-derived ($n = 837$) and canonical ($n = 118{,}701$) peptides per tissue. Bottom, the percentage of ncORF peptides relative to total peptides per tissue. Significant differences (two-sided Fisher's exact test with Holm–Bonferroni correction) are highlighted in red. The dashed line indicates the mean ncORF percentage.

This preference for C-terminal HLA-I peptides from microproteins probably results from fewer cleavages required to process peptides from the termini.

Lastly, we also found that expression level and tissue type may influence microprotein detection. For example, RNA expression was significantly higher for detected microproteins compared with undetected ones (14.3 fragments per kilobase of transcript per million mapped reads (FPKM) versus 10.7 FPKM, respectively; $P = 1.1 \times 10^{-23}$; two-sided Wilcoxon rank-sum test) (Fig. 3d and Extended Data Fig. 6e,f). Using the HLA Ligand Atlas data[22], we compared the relative proportions of microprotein-derived peptides and annotated protein-coding sequence (CDS)-derived HLA-I peptides per tissue (Extended Data Fig. 6g). We found that the proportion of ncORF-encoded peptides in stomach tissue showed a subtle decrease compared to other tissues (−0.6%, $P = 1.7 \times 10^{-4}$, Fisher's exact test) (Fig. 3e), while the spinal cord and uterus showed mild enrichments (0.8%, $P = 1.9 \times 10^{-3}$; 3.1%, $P = 0.029$) (Fig. 3e). These observations were not explained by differences in RNA transcript expression (Extended Data Fig. 6h). Although modest in effect size, these results may point to tissue-specific regulation of ncORF translation and presentation in the immunopeptidome.

Evolutionary insights to interpret ncORFs

The biological interpretation of ncORFs and ncORF-encoded microproteins has historically been framed by their lack of clear evolutionary constraint as protein-coding genes[23]. However, it is also possible that conventional methods[24] do not sufficiently capture their constraint as ORFs, which may inform their evolutionary provenance. To address this issue, we created ORBL, which uses multispecies whole-genome alignments to quantify an evolutionary signature of conservation of 'ORFness': conservation across species of the initiation codon, the termination codon and the 'openness' of the reading frame without regard to conservation of the amino acid sequence (Fig. 4a, Methods and Extended Data Fig. 7a). This conservation score is denoted as ORBLv (Supplementary Results).

We calculated ORBLv for human protein-coding CDSs and GENCODE ncORFs (Supplementary Table 11), observing several expected trends: CDSs, particularly short ones, exhibit higher scores overall, and scores calculated based on branch lengths within the primate clade are higher than those from the placental mammal clade (Extended Data Fig. 7b,c). Likewise, ncORFs with biotypes that overlap annotated CDSs (upstream overlapping ORFs (uoORFs), internal ORFs (intORFs) and downstream overlapping ORFs (doORFs)) have higher ORBLv scores, presumably due to constraint to preserve the CDS (Extended Data Fig. 7b,c).

Yet, conservation as measured by ORBLv may be confounded if the nucleotide sequence is retained by chance, as might happen for short ORFs. We therefore developed an ORBL constraint score, ORBLq, as the quantile of the ORBLv score among untranslated ORFs of the same biotype and similar length. Using ORBLq, we observed that a large percentage of ncORFs exhibit evolutionary constraint (Fig. 4b–e, Extended Data Fig. 7d,e and Supplementary Table 11). For example, 2,211 of the 7,264 ncORFs (30.4%), including 1,335 of the 2,915 upstream ORFs (uORFs, 45.8%), had a placental mammal ORBLq > 0.9, as compared to 10% expected for untranslated ORFs ($P < 2.3 \times 10^{-50}$, binomial test) (Fig. 4b,c). The excess for uORFs and uoORFs suggests the presence of a large number of conserved functional regulatory upstream ORFs. By contrast, few ncORFs score positively for metrics of amino acid conservation, as only 143 of the 7,264 ncORFs (2.0%), including 74 out of 2,915 uORFs (2.5%), have a PhyloCSF[24] score of >10 (Fig. 4f).

We next examined whether ORBL can inform reference annotation by determining whether there is any association between evolutionary constraint and peptide-level detectability. We found that ORBLq scores of ncORFs encoding microproteins detected by HLA-I peptides were significantly higher than those of undetected ones (n = 1,735 versus 5,081, $P = 1.38 \times 10^{-12}$, two-sided Wilcoxon rank-sum test) (Fig. 4e and Extended Data Fig. 7f), particularly for the uORF and intORF subsets (n = 759 versus 2156, $P = 2.66 \times 10^{-5}$, and n = 247 versus 496, P = 0.032, respectively, with Holm–Bonferroni correction for six hypotheses) (Fig. 4g). c8riboseqorf102 provides an instructive example, as this ncORF has a high ORBLq score (0.98) demonstrating ORF-level constraint across placental mammals but a negative PhyloCSF score (−30), and its corresponding microprotein is detected by immunopeptidomics (Fig. 4h). Even so, recently emerged ncORFs can also present with HLA-I peptide evidence, such as c11norep1 in *BET1L* (Fig. 4i). We conclude that signatures of evolutionary constraint according to ORFness are associated with the detection of HLA-I peptides from ncORF-encoded microproteins.

Annotation of protein-coding ncORFs

A primary goal of this work is to develop a standardized analytical framework and nomenclature system for assigning evidence to ncORFs[25]. Using proteomics, immunopeptidomics and Ribo-seq as complementary techniques, we designed a tier-based classification of ncORFs with the intent of streamlining their annotation schema and biological interpretation (Fig. 5a and Methods).

From this system, we highlight several ncORFs undergoing a change in status towards a conventional protein-coding gene. These ncORFs encode microproteins within the tier 1A classification, which indicates sufficient support in conventional proteomics and Ribo-seq data to satisfy HUPO-HPP guidelines for protein verification[15]. After manual inspection of MS peptide data, we identified 20 candidates for tier 1A status, but further scrutiny reduced this list to 15 ncORFs due to pseudogenic sequences, a GRCh38 assembly error and insufficient Ribo-seq evidence (Fig. 5b,c, Supplementary Table 3 and Supplementary Results).

GENCODE have so far annotated three of the tier 1A ncORFs as protein-coding genes: c12norep105 in *CYP27B1*, c21norep46 in *ERVH48-1* and c11riboseqorf4 in *PIDD1*. The latter is a 171-amino-acid uoORF (Extended Data Fig. 8a) that has recently been investigated as a functional protein[26]. Notably, *PIDD1* uoORF peptides are found in non-malignant tissue samples, cancer samples and cell lines, suggesting a physiological role for the protein.

Peptideins: candidates of unclear status

While tier 1A candidates clearly warrant consideration as potential new protein-coding genes, our tier system further prioritizes other ncORFs with strong experimental data as a translated microprotein. However, these often fall short of the threshold for protein annotation at this time. To advance these ncORFs in biological inquiry, we invoke the emerging umbrella term of peptidein, which we define as an ORF with experimentally confirmed RNA translation and protein synthesis, but for which the data are currently insufficient to claim conventional protein-coding gene status. Factors that influence a peptidein designation are the number of observed non-HLA MS peptides, whether the MS peptides are observed exclusively in cancer samples, whether evolutionary constraint at the amino acid level is present and whether mechanistic biological inquiry has elucidated a function for the ncORF-encoded microprotein (Fig. 5a).

We identified three subclasses of ncORFs that may qualify as peptideins (Fig. 5d). Tiers 1B and 2B encompass ncORFs from which microproteins exhibit with high-confidence HLA-I evidence as presented HLA ligands, confirming protein synthesis at these sites. In fact, three tier 1B candidates were annotated as protein coding by GENCODE previously[4] (uoORF c14riboseqorf117 in *EIF5*, uORF c1riboseqorf55 in *PTP4A2* and uORF c3riboseqorf98 in *CGGBP1*), in each case on the basis of the evolutionary profile. Tier 2A ncORFs have one tryptic MS peptide supporting its microprotein, which may therefore capture candidates that are too short to generate multiple peptides. Indeed, 21 out of 39 tier 2A ncORFs demonstrated a validated tryptic peptide, Ribo-seq data and additional HLA peptides (Supplementary Tables 3 and 12 and Fig. 5e). For example, c17norep146, a uoORF in the *PSMC5* gene, is well supported by HLA-I and HLA-II peptides, and demonstrates Ribo-seq translation far exceeding that of the annotated main CDS in the same gene locus (Extended Data Fig. 8b). Such peptideins are worthy of close monitoring, as future data generation may reclassify them as protein-coding genes.

Functional genomics augments annotation

We hypothesized that incorporation of functional genomics may highlight a subset of ncORFs with roles in cellular biology[7,27,28], potentially illuminating candidate protein-coding genes and/or peptideins. We therefore sought to integrate knockout phenotype evidence to define candidates encoding potential pan-essential proteins, because such proteins are central to core cell functions[29] and common drug targets[30].

To pursue this, we developed a step-wise approach (Fig. 6a) using (1) loss-of-function CRISPR–Cas9 screening for over 2,000 ncORFs across 8 human cell lines; (2) filtration of putative hits for adequate

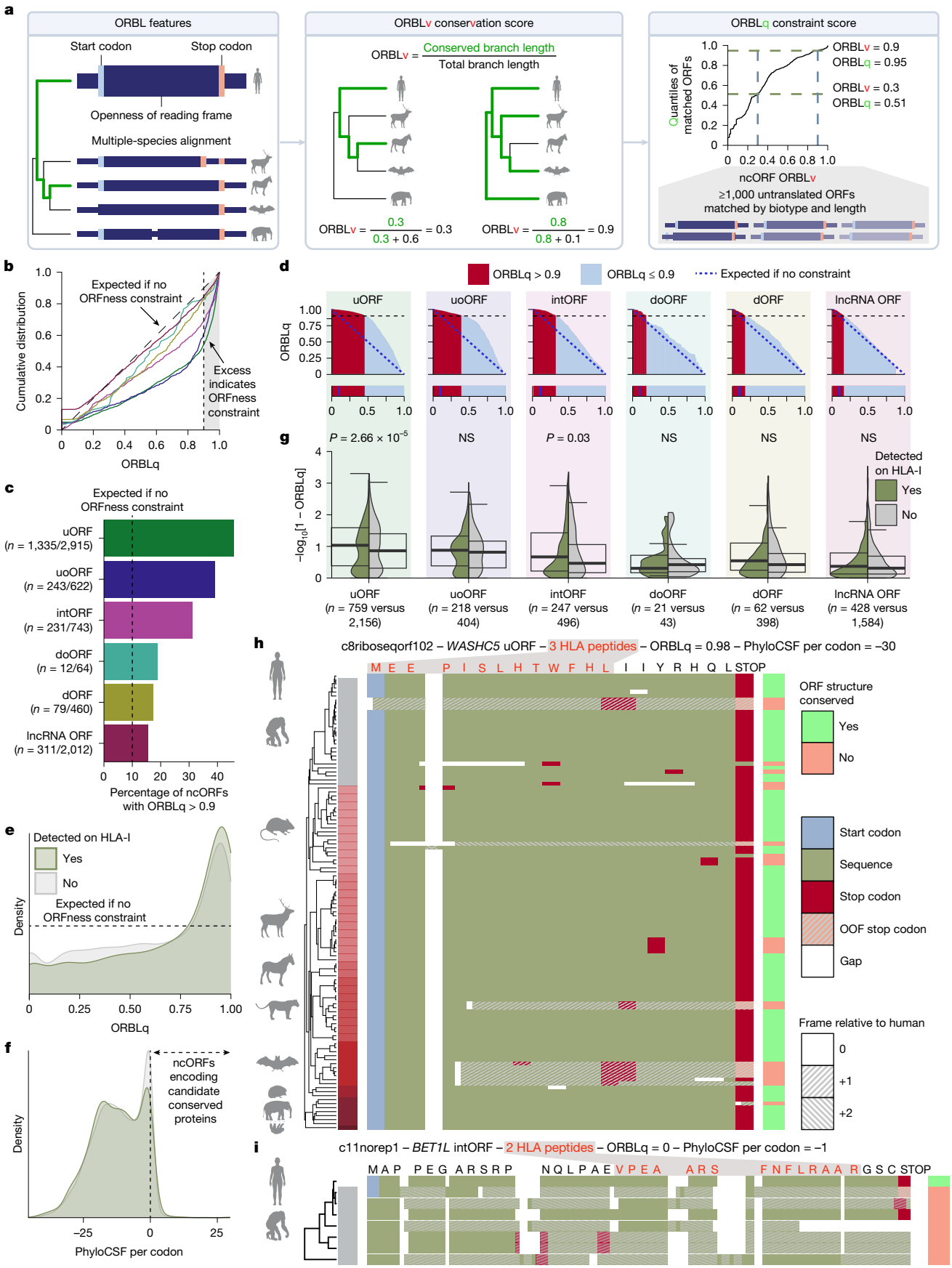


Fig. 4 | See next page for caption.

Fig. 4 | Overview of the ORBL tool, using placental mammal scores. **a**, Schematic of ORBL scores, measuring the conservation and evolutionary constraint of ORFness through conservation of start codon, stop codon and reading frame openness. The ORBLv conservation score is the branch length of species having a conserved ORF divided by the branch length of all in the whole-genome alignment. To distinguish ORFness constraint from CDS constraint or chance conservation, the ORBLq constraint score reports the quantile of an ncORF's ORBLv among ORBLv scores of matched untranslated ORFs having the same biotype and similar length. **b**, Cumulative distributions of ncORF ORBLq scores by biotype (excluding the 'mixed' biotype because ORBLq is undefined). Colour designations are as described in **c**. The diagonal line shows the expected null distribution for ORFs without ORFness constraint. **c**, The percentage of ncORFs by biotype having ORBLq > 0.9, far exceeding the 10% expected without ORFness constraint (vertical dashed line), indicating that many ncORFs experienced purifying selection to preserve ORFness. **d**, The distributions of ORBLq scores by biotype. The red bars indicate scores > 0.9. The diagonal lines show expected distribution without ORFness constraint. **e**,**f**, Density plots of ORBLq scores (**e**) and PhyloCSF-per-codon scores (**f**) for ncORFs detected by HLA-I immunopeptidomics (green) or undetected (grey). Higher ORBLq scores than the expected distribution (dashed line) indicate evolutionary constraint on ORFness, particularly for detected ncORFs. **g**, Transformed ORBLq scores grouped by ncORF biotype and HLA-I immunopeptidomics detection status. Detected ncORFs have significantly higher scores than undetected among uORFs and intORFs ($P = 2.66 \times 10^{-5}$ and 0.03, respectively, two-sided Wilcoxon rank-sum test adjusted for six hypotheses using Holm–Bonferroni correction). The box plots show the median (centre line) and 25th (Q1) and 75th (Q3) percentiles; the whiskers extend to the most extreme values within 1.5× the interquartile range from Q1 to Q3. **h**,**i**, Conservation of immunopeptidomics-detected ncORFs: c8riboseq102 (high ORBLq, low PhyloCSF) (**h**) and c11norep1 (low ORBLq) (**i**).

expression (RNA-seq) and translation (Ribo-seq) in the indicated cells; (3) inspection of HLA peptide evidence; (4) hit prioritization through a meta-analysis of 25 CRISPR screens including saturation mutagenesis efforts; and (5) assessment of evolutionary constraint through ORBL (Methods, Supplementary Table 13 and Extended Data Fig. 9a–g). We additionally ensured that the phenotypic effect of uORF knockout was distinct from that of the adjacent coding region on that mRNA. Lastly, we ruled out the possibility of uORF knockout resulting in cellular toxicity through upregulation of an anti-proliferative adjacent CDS using CRISPR activation (CRISPRa) screening data from a matching cell line[31], which showed little evidence for anti-proliferative CDSs adjacent to essential uORFs (Extended Data Fig. 9e,f).

Overall, we identified 51 ncORFs exhibiting a pan-essential knockout signature (Supplementary Table 13). From these, six ncORFs (c2riboseqorf47, c14riboseqorf118, c2riboseqorf55, c3riboseqorf106, c10riboseqorf92, c6norep15) qualified as candidate peptideins or protein-coding genes based on HLA peptide evidence for their encoded microproteins. These six segregated based on high ORBLq values exhibiting evolutionary constraint (ORBLq > 0.9; c2riboseqorf47, c14riboseqorf118, c2riboseqorf55, c3riboseqorf106), or low ORBLq values indicating little ORF-level constraint (ORBLq < 0.7; c10riboseqorf92, c6norep15) (Extended Data Fig. 9g).

From this set, c2riboseqorf47 (a tier 1B uORF in the *GMCL1* gene) further warranted inspection as a protein-coding gene (Fig. 6b): in addition to a loss-of-function phenotype and a high ORBLq score, c2riboseqorf47 had a positive PhyloCSF score, evidence of translation in non-human mammals (Extended Data Fig. 9h), and both HLA-I and HLA-II peptides, suggesting production of a stable microprotein that can be presented from both intracellular and extracellular sources (Supplementary Table 6). Thus, GENCODE have now annotated c2riboseqorf47 as protein-coding gene ENSG00000310604, exemplifying the integration of functional genomics, evolutionary analyses and immunopeptidomics to nominate protein-coding genes in the absence of tryptic MS support.

OLMALINC produces an essential peptidein

We next looked at whether CRISPR screening could further narrow which of the other peptideins (Fig. 6a) may demonstrate roles in cell biology. Using published ncORF saturation mutagenesis screens[7,28], we defined a functional enrichment score for ncORF essentiality compared with the local background signal, which may contain other adjacent ncORFs or CDSs (Fig. 6c). This analysis revealed that both c10riboseqorf92 and c3riboseqorf106 exhibited a signature of selective loss of fitness (Fig. 6d and Extended Data Fig. 9i). As c3riboseqorf106 (located in the *ZBTB11-AS1* transcript) has been previously inspected[28], we focused on c10riboseqorf92.

c10riboseqorf92 is a 123 amino acid sequence located on the *OLMALINC* transcript (also known as *LINC00263*)—an RNA that has six ncORFs recognized by GENCODE annotations. Yet, only c10riboseqorf92 scored as a pan-essential genetic dependency (Fig. 6e and Supplementary Table 13), which was also supported by CRISPR–Cas13 RNA degradation screening of long non-coding RNAs (lncRNAs) (Extended Data Fig. 9j). We confirmed that re-expression of the c10riboseqorf92 coding sequence rescued the loss-of-viability phenotype observed after *OLMALINC* knockdown, indicating an ORF-specific function (Fig. 6f and Extended Data Fig. 9k,l).

To nominate potential biological roles for c10riboseqorf92, we used both correlation analyses for its genetic knockout as well as transcriptome profiling. First, we performed c10riboseqorf92 knockout in over 485 cell lines, observing a loss-of-viability phenotype in 415 cell models (85.6%), consistent with pan-essentiality (Extended Data Fig. 9m,n). Using these data, we correlated its knockout pattern with other pan-essential genes tested in the Dependency Map, which demonstrated an enrichment for genes that are involved in mitosis and DNA damage regulation (Fig. 6g and Supplementary Table 13).

Second, because *OLMALINC* may have biological roles both as a non-coding RNA as well as through the c10riboseqorf92 peptidein, we performed transcriptome profiling after *OLMALINC* knockdown in isogenic A375 cells expressing either GFP or the c10riboseqorf92 coding sequence. We found that 513 genes showed increased abundance and 456 genes showed decreased abundance in c10riboseqorf92-expressing cells compared with GFP-expressing cells treated with *OLMALINC* knockdown, with 14 genes exhibiting a statistically significant interaction *P* value of <0.05 (Fig. 6h and Extended Data Fig. 9o). Differentially regulated genes were associated with Gene Ontology (GO) terms related to cell metabolism (hypoxia, glycolysis) as well as DNA damage response through ultraviolet light exposure and TNF signalling (Fig. 6h), supporting our correlative analyses of c10riboseqorf92 knockout with Dependency Map phenotypes (Fig. 6g).

To generalize these findings, we performed multiplexed single-cell RNA-seq (scRNA-seq) analysis after c10riboseqorf92 knockout[32], in which we determined scRNA-seq profiles for 12 different cell lines after knockout. Consistent with our data in A375 cells, we observed induction of mitosis- and chromosome-related processes with c10riboseqorf92 knockout across cell lines, with downregulation of translation and metabolism-related processes (Fig. 6i,j and Extended Data Fig. 9p). Despite the cumulative data supporting c10riboseqorf92, we note that it remains annotated as a peptidein because it does not possess clear evidence of function in normal physiology, as its evidence remains restricted to transformed cell lines or cancer.

An agenda for microprotein research

A unique capacity of our multi-consortium collaboration is the ability to develop consensus on the key challenges that we believe the ncORF field needs to address. Moving forward, we see seven areas where the research community should be engaged:

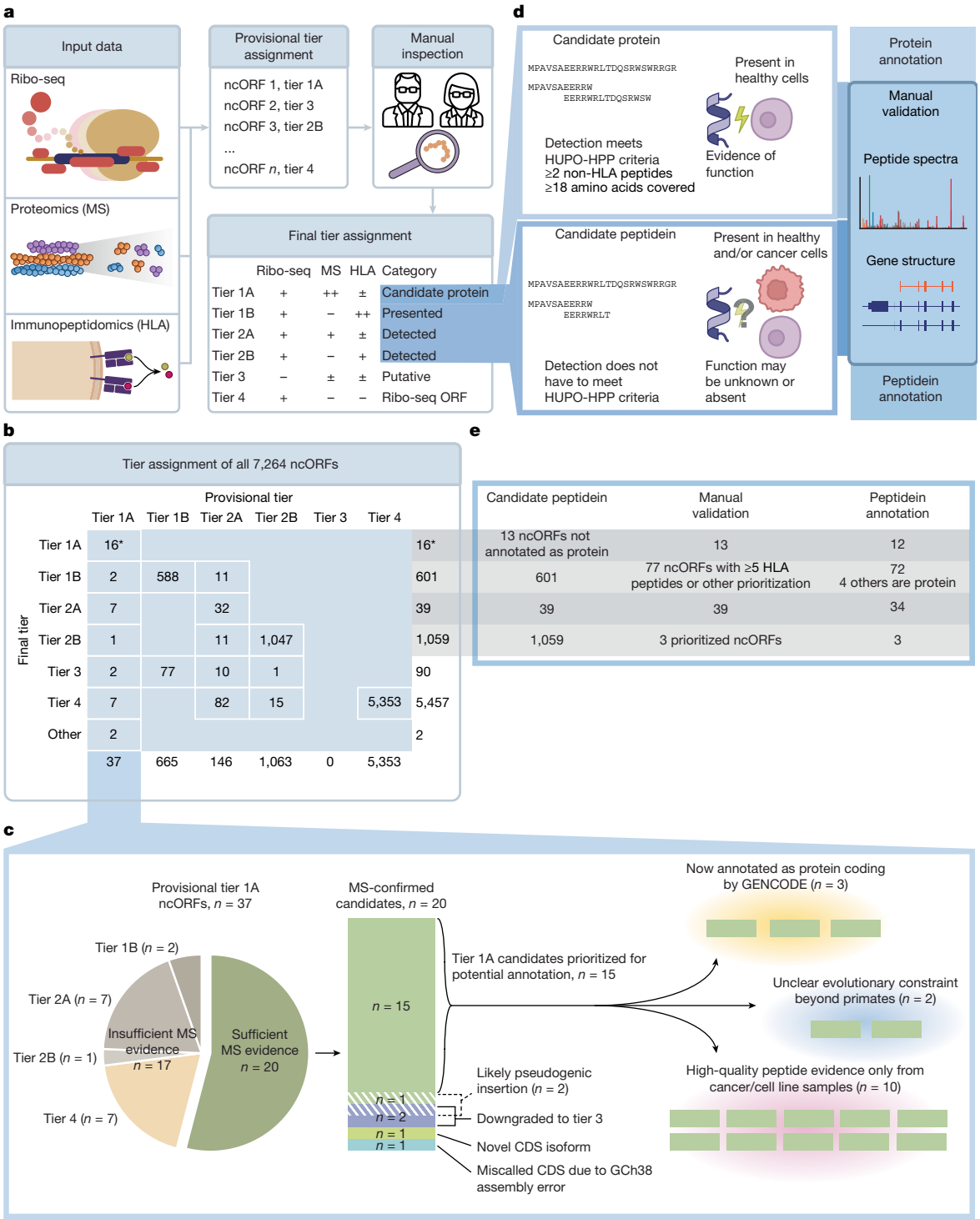


Fig. 5 | Overview of the tier system. a, Schematic of how provisional and final tiers can be assigned to ncORFs. First, Ribo-seq, proteomics and immunopeptidomics data can be computationally integrated to assign provisional tiers based on the quality of each data entity. Manual inspection of each data entity is then necessary to assign a final tier to each ncORF. '+' denotes detection, '++' denotes abundant detection, '±' denotes either presence or absence of detection, and '−' denotes absence of detection. **b**, The results of the provisional and final tier assignment for the 7,264 ncORFs analysed for this study. The asterisks indicate the inclusion of one likely pseudogenic sequence. **c**, Overview of the curation process for the provisional tier 1A ncORFs. **d**, The criteria for ncORFs to be annotated as proteins or peptideins. For annotation as a protein, the detection of the ncORF has to follow HUPO-HPP criteria in normal cells and evidence of function is required. For annotation as a peptidein, detection of the ncORF in normal or cancer cells may be enough. **e**, The results from protein and peptidein annotation of the 7,264 ncORFs considered in this study. Only tier 1A and tier 2A ncORFs, and tier 1B ncORFs detected by ≥5 HLA peptides were considered for manual validation.

(1) Are HUPO-HPP guidelines for protein verification suitable for ncORF-encoded microproteins? These require two peptides of length 9 amino acids or more, and spanning at least 18 amino acids of the ORF[15]. Yet, many ncORFs are smaller than 18 amino acids[4,23,33,34], and 28.3% (2,059 out of 7,264) of ncORFs in this study are <25 amino acids, making it inherently difficult to meet these guidelines.

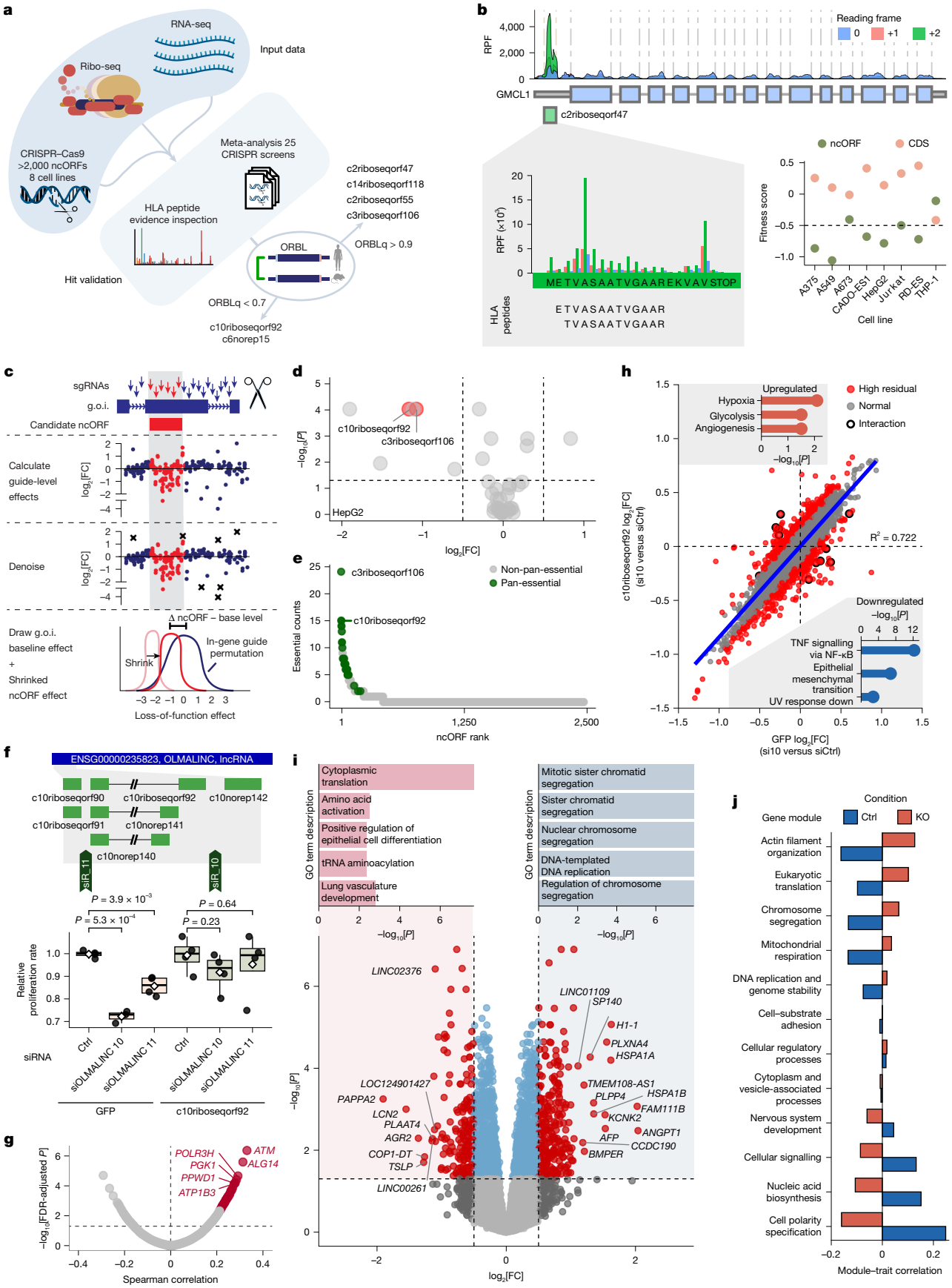


Fig. 6 | See next page for caption.

Fig. 6 | Function-based refinement of ncORF annotation. a, Schematic. **b**, Evidence supporting c2riboseqorf47. Top, frame-coloured ribosome profiling-derived P-site coverage. Bottom right, loss of fitness from perturbing the annotated CDS (orange) or the ncORF (green). RPF, ribosome protected fragments. **c**, Workflow for enrichment scores in CRISPR tiling screens. g.o.i., gene of interest; sgRNA, single guide RNA. **d**, CRISPR tiling screen scores in HepG2 cells. Significance was determined using a gene-specific permutation test (two-sided *P* values, see Methods). **e**, Across 25 CRISPR datasets, ncORFs were ranked by the number of datasets showing a fitness score of <−0.5. Green dots show ncORFs that are essential in ≥60% of evaluated samples. **f**, c10riboseqorf92 rescues the phenotype ($n = 4$ biological replicates per condition) induced by *OLMALINC* transcript silencing. Top, transcript positions of c10riboseqorf92 and siRNA target sites. Bottom, proliferation values of A375 cell expressing GFP or c10riboseqorf92 after siRNA treatment. The box plots show the median and the 25th (Q1) and 75th (Q3) percentiles; the whiskers extend to 1.5× the interquartile range from Q1 and Q3. *P* values were determined using the two-sided Welch *t*-test. **g**, Spearman correlation of c10riboseqorf92 versus 17,110 genes across 485 pooled CRISPR screen cell lines (DepMap). The top 0.5 percentile is highlighted ($\rho > 0.3$). **h**, Correlation between transcriptional responses to *OLMALINC* knockdown in A375 cells expressing GFP control or c10riboseqorf92 ($n = 3$ per condition). Each point represents a gene's $\log_2$-transformed fold change (FC) in GFP (*x* axis) versus c10riboseqorf92-expressing cells (*y* axis). The blue line shows the linear regression fit ($R^2 = 0.722$). High-residual genes (top 5%) are shown in red. Significant interaction hits are circled. UV, ultraviolet. **i**, Pseudobulk differential expression ($n = 12$ cell lines) comparing knockout (KO) to control from scRNA-seq data. Statistical analysis was performed using a limma-voom linear model with Benjamini–Hochberg adjustment. Significant genes (FDR < 0.05; $|\log_2[FC]| > 0.5$) are highlighted. GSEA results are shown, ranked by significance and normalized effect size (NES). **j**, Module–trait correlation of co-expression gene modules derived from a network built using all conditions and filtered cell lines in scRNA-seq.

(2) Should HLA immunopeptidomics be used as evidence that a ncORF encodes a protein-coding gene? 1,785 out of 7,264 ncORFs are observed with HLA data, including 24 ncORFs with 1 peptide in tryptic MS data suitable for potential annotation. For example, our identification of a protein-coding gene from c2riboseqorf47 was informed by HLA data.
(3) Should peptides detected in cancer samples or immortalized cell lines support protein-coding gene annotation? 2.36 billion out of 3.53 billion (66.9%) MS2 spectra searched in the non-HLA PeptideAtlas are from cancer tissue or cancer cell line samples. Proteins supported by such data are potentially cancer-specific products, which has implications for their annotation as peptideins.
(4) What is the role of evolutionary inference in annotation for ncORFs? Most of the 7,264 Ribo-seq ncORFs apparently lack constraint on their amino acid sequence[23,35], even though thousands exhibit ORF-level constraint according to ORBL. Thus, whether lack of amino acid constraint argues against function remains debatable[36,37].
(5) Which alternative forms of experimental analysis could support protein-coding gene annotation? Clearly, it would make sense for any ncORF to be annotated as protein coding if evidence is provided not only for the existence of the protein, but also the nature of at least one biological function. Notably, immune recognition of a peptide is not currently considered a biological function by annotation projects.
(6) How should we annotate microproteins for which function can be neither demonstrated nor inferred? The GENCODE, UniProtKB, HGNC, RefSeq and HUPO-HPP annotation projects have decided to classify such products as peptideins, as described in this work. Precise annotation guidelines are currently being prepared.
(7) Should deep learning approaches inform gene or protein annotation? While annotation is historically rooted in manual inspection, advancements in deep learning may offer an opportunity to classify high-quality MS spectra and Ribo-seq data for future annotation efforts[38,39].

Discussion

The extent of the undiscovered proteome is one of the central questions in human biomedicine. This work reflects the multi-consortium collaboration between the TransCODE Consortium, the HUPO-HPP/PeptideAtlas project, the HIPP immunopeptidomics project and the GENCODE gene annotation group to coalesce a generalizable approach towards understanding which ncORFs can be understood as encoding proteins.

This work further helps to resolve the tension between the concepts of protein identification and protein-coding gene annotation, which are distinct. While protein identification refers to the experimental detection of a polypeptide molecule, protein-coding gene annotation is historically rooted in the idea that the translated protein imparts a biological function. ncORFs and their encoded microproteins present a paradox in this realm: despite their prevalence, few have conventional metrics supporting a protein-coding gene annotation. To solve this paradox, we have invoked the annotation concept of the peptidein: a translation product confidently detected endogenously, but for which a role in normal physiology cannot be verified at the present time. Peptideins also include potentially transient products of cellular stress or defective ribosome translation[40]. Here we classify 121 initial peptidein annotations (Supplementary Table 12), and full guidelines for this process are forthcoming.

Another innovation of our work is the ability to adapt new lines of evidence to delineate ncORFs as protein-coding genes. Our development of the ORBL method to analyse ORF evolutionary constraint, our assessment of CRISPR-based functional inference and our prioritization of HLA immunopeptide support were central to the establishment of c2riboseqorf47 as a protein-coding gene (*GMCL1* uORF), despite there being no tryptic MS peptides and there being ambiguous amino acid evolutionary constraint when using conventional approaches. Thus, some HLA-detected ncORFs lack evidence in tryptic MS proteomics for either technological or biological reasons, such as the small size or amino acid composition of their encoded microproteins.

Our work also points toward emerging technological and methodological innovations optimized for small proteins that help with their identification by MS. Assessing synthetic peptide standards, PTMs or multi-protease digestions, we find unambiguous evidence for some ncORF-encoded microproteins that were not visible in routine tryptic MS studies.

At the same time, many ncORFs appear to generate immunopeptides but not tryptic peptides, as we and others have observed[2,3,9,19,27,41]. One potential explanation for this observation implicates lower stability for many ncORF-derived polypeptides, perhaps due to BAG6-mediated degradation in the proteasome[19,21]. Whether structure-based analyses of ncORFs, such as those facilitated by AlphaFold or ESMFold, can be used to parse which HLA-supported ncORFs are more likely to encode a stable protein remains an open question (Extended Data Fig. 10, Supplementary Table 14 and Supplementary Results).

We also emphasize that annotation of some ncORFs as peptideins based on HLA peptides may enable their further study and medical relevance, potentially opening a path to their annotation as true protein-coding genes. Our work here highlights c10riboseqorf92 (in the *OLMALINC* transcript): while we do not yet have sufficient evidence that this ncORF encodes a bona fide protein, its CRISPR-based phenotypes in the context of cancer cells are intriguing. Second, some peptideins may have direct biomedical relevance even if they have no physiological basis in normal biology, which is exemplified by targeting such cryptic peptides with cancer immunotherapy[10,42,43]. Third, the human genetics

community has intensified scrutiny on how variants impacting ncORFs and their translation products contribute to human genetic disease[5,44].

Lastly, we highlight four limitations to our current efforts. First, the role of sample type in protein-coding gene annotation remains problematic. Thus, despite multiple tryptic MS peptides for ncORFs in *STK11*, *ZNF219* and *CIRBP*, they remain peptideins because all supporting peptides are derived either from cancer samples or immortalized cell lines. Second, our work has focused on data-dependent acquisition for MS; data-independent acquisition, especially when coupled with targeted validation through PRM, may provide increased sensitivity to detect ncORF-derived proteins in specific contexts[45]. Third, our efforts emphasize large-scale manual inspection of both peptide data and ribosome profiling data for annotation purposes. However, manual inspection of thousands of candidates is not feasible for most researchers, who may use tools assessing peptide retention time prediction and ion mobility prediction[38,46–49] for scalable, high-throughput research studies. Finally, there are several limitations of our ORBL evolutionary analysis, which are reported in the Supplementary Results.

In summary, we report the collaborative efforts of the TransCODE consortium with stakeholders in proteomics (PeptideAtlas/HUPO-HPP), immunopeptidomics (HIPP) and gene annotation (GENCODE) to develop a consensus understanding of protein-level evidence for 7,264 ncORFs. Through our efforts, we bring microproteins and alternative protein molecules into reference gene annotation by defining them as either a protein-coding gene or a peptidein, a new concept referring to confirmed protein molecules of indeterminate consequence. Finally, we make all ncORFs, peptides and spectra publicly available through PeptideAtlas (https://peptideatlas.org/builds/human/#ncORFs).

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-026-10459-x.

1. van Heesch, S. et al. The translational landscape of the human heart. *Cell* **178**, 242–260 (2019).
2. Ouspenskaia, T. et al. Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. *Nat. Biotechnol.* **40**, 209–217 (2022).
3. Martinez, T. F. et al. Accurate annotation of human protein-coding small open reading frames. *Nat. Chem. Biol.* **16**, 458–468 (2020).
4. Mudge, J. M. et al. Standardized annotation of translated open reading frames. *Nat. Biotechnol.* **40**, 994–999 (2022).
5. Whiffin, N. et al. Characterising the loss-of-function impact of 5′ untranslated region variants in 15,708 individuals. *Nat. Commun.* **11**, 2523 (2020).
6. Oz-Levi, D. et al. Noncoding deletions reveal a gene that is critical for intestinal function. *Nature* **571**, 107–111 (2019).
7. Hofman, D. A. et al. Translation of non-canonical open reading frames as a cancer cell survival mechanism in childhood medulloblastoma. *Mol. Cell* **84**, 261–276 (2024).
8. Ely, Z. A. et al. Pancreatic cancer–restricted cryptic antigens are targets for T cell recognition. *Science* **388**, eadk3487 (2025).
9. Chong, C. et al. Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat. Commun.* **11**, 1293 (2020).
10. Huang, D. et al. Tumour circular RNAs elicit anti-tumour immunity by encoding cryptic peptides. *Nature* **625**, 593–602 (2024).
11. Frankish, A. et al. GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res.* **51**, D942–D949 (2022).
12. van Wijk, K. J. et al. Detection of the *Arabidopsis* proteome and its post-translational modifications and the nature of the unobserved (dark) proteome in PeptideAtlas. *J. Proteome Res.* **23**, 185–214 (2024).
13. Deutsch, E. W. et al. The 2025 Report on the Human Proteome from the HUPO Human Proteome Project. *J. Proteome Res.* **25**, 539–555 (2026).
14. Caron, E., Aebersold, R., Banaei-Esfahani, A., Chong, C. & Bassani-Sternberg, M. A Case for a Human Immuno-Peptidome Project Consortium. *Immunity* **47**, 203–208 (2017).
15. Deutsch, E. W. et al. Human Proteome Project mass spectrometry data interpretation guidelines 3.0. *J. Proteome Res.* **18**, 4108–4116 (2019).
16. Wacholder, A. et al. Community benchmarking and evaluation of human unannotated microprotein detection by mass spectrometry based proteomics. *Nat. Commun.* **17**, 1241 (2026).
17. Whited, A. M. et al. Biophysical characterization of high-confidence, small human proteins. *Biophys. Rep.* **4**, 100167 (2024).
18. Wang, D. et al. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol. Syst. Biol.* **15**, MSB188503 (2019).
19. Cuevas, M. V. R. et al. Most non-canonical proteins uniquely populate the proteome or immunopeptidome. *Cell Rep.* **34**, 108815 (2021).
20. Abelin, J. G. et al. Workflow enabling deepscale immunopeptidome, proteome, ubiquitylome, phosphoproteome, and acetylome analyses of sample-limited tissues. *Nat. Commun.* **14**, 1851 (2023).
21. Kesner, J. S. et al. Noncoding translation mitigation. *Nature* **617**, 395–402 (2023).
22. Marcu, A. et al. HLA Ligand Atlas: a benign reference of HLA-presented peptides to improve T-cell-based cancer immunotherapy. *J. Immunother. Cancer* **9**, e002071 (2021).
23. Sandmann, C.-L. et al. Evolutionary origins and interactomes of human, young microproteins and small peptides translated from short open reading frames. *Mol. Cell* **83**, 994–1011 (2023).
24. Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275–i282 (2011).
25. Prensner, J. R. et al. What can ribo-seq, immunopeptidomics, and proteomics tell us about the noncanonical proteome? *Mol. Cell. Proteom.* **22**, 100631 (2023).
26. Comtois, F. et al. Noncanonical altPIDD1 protein: unveiling the true major translational output of the *PIDD1* gene. *Life Sci. Alliance* **8**, e202402910 (2025).
27. Chen, J. et al. Pervasive functional translation of noncanonical human open reading frames. *Science* **367**, 1140–1146 (2020).
28. Prensner, J. R. et al. Noncanonical open reading frames encode functional proteins essential for cancer cell survival. *Nat. Biotechnol.* **39**, 697–704 (2021).
29. Funk, L. et al. The phenotypic landscape of essential human genes. *Cell* **185**, 4634–4653 (2022).
30. Chang, L., Ruiz, P., Ito, T. & Sellers, W. R. Targeting pan-essential genes in cancer: challenges and opportunities. *Cancer Cell* **39**, 466–479 (2021).
31. Sanson, K. R. et al. Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. *Nat. Commun.* **9**, 5416 (2018).
32. McFarland, J. M. et al. Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. *Nat. Commun.* **11**, 4296 (2020).
33. Chothani, S. P. et al. A high-resolution map of human RNA translation. *Mol. Cell* **82**, 2885–2899 (2022).
34. Chothani, S. et al. An expanded reference catalog of translated open reading frames for biomedical research. *Nucleic Acids Res.* **54**, gkag234 (2026).
35. Broeils, L. A., Ruiz-Orera, J., Snel, B., Hubner, N. & van Heesch, S. Evolution and implications of de novo genes in humans. *Nat. Ecol. Evol.* **7**, 804–815 (2023).
36. Carvunis, A.-R. et al. Proto-genes and de novo gene birth. *Nature* **487**, 370–374 (2012).
37. Keeling, D. M., Garza, P., Nartey, C. M. & Carvunis, A.-R. The meanings of "function" in biology and the problematic case of de novo gene emergence. *eLife* **8**, e47014 (2019).
38. Adams, C. et al. Fragment ion intensity prediction improves the identification rate of non-tryptic peptides in timsTOF. *Nat. Commun.* **15**, 3956 (2024).
39. Clauwaert, J. et al. Deep learning to decode sites of RNA translation in normal and cancerous tissues. *Nat. Commun.* **16**, 1275 (2025).
40. Yewdell, J. W. & Hollý, J. DRiPs get molecular. *Curr. Opin. Immunol.* **64**, 130–136 (2020).
41. Laumont, C. M. et al. Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Transl. Med.* **10**, eaau5516 (2018).
42. Barczak, W. et al. Long non-coding RNA-derived peptides are immunogenic and drive a potent anti-tumour response. *Nat. Commun.* **14**, 1078 (2023).
43. Zeng, L. et al. An epitope encoded by uORF of RNF10 elicits a therapeutic anti-tumor immune response. *Mol. Ther. Oncolyt.* **31**, 100737 (2023).
44. Lim, Y. et al. Multiplexed functional genomic analysis of 5′ untranslated region mutations across the spectrum of prostate cancer. *Nat. Commun.* **12**, 4217 (2021).
45. Martinez, T. F. et al. Profiling mouse brown and white adipocytes to identify metabolically relevant small ORFs and functional microproteins. *Cell Metab.* **35**, 166–183 (2023).
46. Bouwmeester, R., Gabriels, R., Hulstaert, N., Martens, L. & Degroeve, S. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nat. Methods* **18**, 1363–1369 (2021).
47. Wen, B., Li, K., Zhang, Y. & Zhang, B. Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis. *Nat. Commun.* **11**, 1759 (2020).
48. Tibbo, A. J. et al. Phosphodiesterase type 4 anchoring regulates cAMP signaling to Popeye domain-containing proteins. *J. Mol. Cell. Cardiol.* **165**, 86–102 (2022).
49. Declercq, A. et al. TIMS2Rescore: a data dependent acquisition-parallel accumulation and serial fragmentation-optimized data-driven rescoring pipeline based on MS2Rescore. *J. Proteome Res.* **24**, 1067–1076 (2025).

[1]Institute for Systems Biology (ISB), Seattle, WA, USA. [2]Princess Máxima Center for Pediatric Oncology, Utrecht, The Netherlands. [3]Oncode Institute, Utrecht, The Netherlands. [4]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, UK. [5]Department of Pediatrics, Division of Pediatric Hematology/Oncology, University of Michigan Medical School, Ann Arbor, MI, USA. [6]Department of Biological Chemistry, University of Michigan Medical School, Ann Arbor, MI, USA. [7]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. [8]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [9]Cardiovascular and Metabolic Sciences, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany. [10]Biozentrum, University of Basel, Basel, Switzerland. [11]Hospital del Mar Research Institute, Barcelona, Spain. [12]Catalan Institute for Research and Advanced Studies (ICREA), Barcelona, Spain. [13]School of Molecular and Cellular Biology, Faculty of Biological Sciences, University of Leeds, Leeds, UK. [14]Functional Proteomics Group, Institute of Cancer Research, Chester Beatty Labs, London, UK. [15]School of Biochemistry and Cell Biology, University College Cork, Cork, Ireland. [16]Stowers Institute for Medical Research, Kansas City, MO, USA. [17]Department of Molecular and Integrative Physiology, University of Kansas Medical Center, Kansas City, KS, USA. [18]Pediatrics Department, University of Sherbrooke, Sherbrooke, Quebec, Canada. [19]HUGO Gene Nomenclature Committee (HGNC), Department of Haematology, University of Cambridge School of Clinical Medicine, Cambridge, UK. [20]Human Technopole, Milan, Italy. [21]Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA. [22]Pittsburgh Center for Evolutionary Biology and Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA. [23]Centre for Computational Biology and Program in Cardiovascular and Metabolic Disorders, Duke-NUS (National University of Singapore) Medical School, Singapore, Singapore. [24]Genome Institute of Singapore (GIS), Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [25]Centre for Cancer Research, Hudson Institute of Medical Research, Clayton, Victoria, Australia. [26]Monash Proteomics & Metabolomics Platform, Department of Medicine, School of Clinical Sciences, Monash University, Clayton, Victoria, Australia. [27]Charité-Universitätsmedizin Berlin, Berlin, Germany. [28]Helmholtz-Institute for Translational AngioCardioScience (HI-TAC) of the Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC) at Heidelberg University, Heidelberg, Germany. [29]DZHK (German Center for Cardiovascular Research)–Partner Site Berlin, Berlin, Germany. [30]Department of Molecular and Cell Biology, Center for Computational Biology, University of California, Berkeley, Berkeley, CA, USA. [31]Department of Pharmaceutical Sciences, University of California, Irvine, Irvine, CA, USA. [32]Department of Biological Chemistry, University of California, Irvine, Irvine, CA, USA. [33]Chao Family Comprehensive Cancer Center, University of California, Irvine, Irvine, CA, USA. [34]Biobix, Lab of Bioinformatics and Computational Genomics, Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Ghent, Belgium. [35]Department of Biology, Humboldt University Berlin, Berlin, Germany. [36]Berlin Institute of Medical Systems Biology (BIMSB), Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany. [37]Biomolecular Mass Spectrometry and Proteomics, Bijvoet Centre for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands. [38]School of Biological Sciences, University of Southampton, Southampton, UK. [39]Department of Biochemistry and Functional Genomics, University of Sherbrooke, Sherbrooke, Quebec, Canada. [40]Department of Chemistry, Yale University, New Haven, CT, USA. [41]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA. [42]Institute for Biomolecular Design and Discovery, Yale University, West Haven, CT, USA. [43]Institute of Genetics and Biotechnology, Faculty of Biology, University of Warsaw, Warsaw, Poland. [44]Department of Biosciences, University of Oslo, Oslo, Norway. [45]Whitehead Institute for Biomedical Research, Cambridge, MA, USA. [46]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. [47]Howard Hughes Medical Institute, Whitehead Institute, Cambridge, MA, USA. [48]David H. Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA, USA. [49]Singapore Immunology Network (SIgN), Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [50]Department of Pharmacy & Pharmaceutical sciences, National University of Singapore (NUS), Singapore, Singapore. [51]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China. [52]Ludwig Institute for Cancer Research, University of Lausanne, Lausanne, Switzerland. [53]Department of Oncology, Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland. [54]Agora Cancer Research Centre, Lausanne, Switzerland. [55]School of Life Sciences, Division Cell Signalling and Immunology, University of Dundee, Dundee, UK. [56]Centre for Immuno-Oncology, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [57]Centre de Recherche du Centre hospitalier universitaire de Sherbrooke (CRCHUS), Sherbrooke, Quebec, Canada. [58]Cancer Research Institute, University of Sherbrooke (IRCUS), Sherbrooke, Quebec, Canada. [59]Present address: Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland. [60]These authors contributed equally: Eric W. Deutsch, Leron W. Kok, Jonathan M. Mudge, Cristian F. Valls, Irwin Jungreis. [61]These authors jointly supervised this work: Robert L. Moritz, John R. Prensner, Sebastiaan van Heesch. ✉e-mail: rmoritz@systemsbiology.org; prensner@umich.edu; s.a.a.vanheesch@prinsesmaximacentrum.nl

Methods

PeptideAtlas database construction and searching

The human non-HLA PeptideAtlas 2023-06 build contains 295 ProteomeXchange datasets (PXDs)[50] split into 1,172 different experiments that comprise a total of 3.5 billion MS/MS spectra. Sequence database searching was performed using MSFragger[51] v.3.7 using search parameters appropriate for each dataset, depending on alkylation, labelling, fragmentation type, instrument, enrichment strategy and more. All datasets were searched with semi-enzymatic settings (typically semi-tryptic). The search database was 2023-02 THISP level 4 database[52] (https://peptideatlas.org/thisp/), which included the 7,264 Ribo-seq ORFs from ref. 4 as well as other contributed sequences that might be translated. All datasets were searched with generic artifactual variable modifications methionine oxidation, protein N-terminal acetylation, peptide N-terminal pyro-glutamic acid from glutamic acid or glutamine, and asparagine and glutamine deamidation. The alkylation modification was set as a fixed modification (typically carbamidomethylated cysteine).

Statistical validation of the results for each experiment was performed using the Trans-Proteomic Pipeline (TPP)[53,54] v.7.0 tools PeptideProphet[55], iProphet[56] and PTMProphet[57], and the results were mapped to the human proteome using ProteoMapper[58], taking known variants into account, and to the genome using the ENSEMBL[59] toolkit as previously described[60]. A complete list of datasets used and a summary of the search results in each build are available online (https://peptideatlas.org/builds/human/non-hla/). Supplementary Table 15 provides FDR metrics at the PSM-, peptide- and protein-levels, as well as for certain subsets of proteins, including the neXtProt core proteome, the 7264 Ribo-seq ncORFs, as well as all CONTRIB sequences, many of which are putative ncORFs.

The Human HLA PeptideAtlas 2023-11 build comprises a set of 118 HLA immunopeptide-enriched publicly available PXDs, which we split into 592 separate experiments, containing 240 million MS/MS spectra from 9,776 MS runs (Supplementary Table 16). Sequence database searching was performed with MSFragger v.3.7 using search parameters appropriate for each dataset, depending on sample handling. All datasets were searched in no-enzyme mode. While some HLA peptides have a lysine or arginine on the C terminus, and therefore exhibit fragmentation patterns typical of tryptic peptides, many HLA peptides do not have such characteristics and their spectra may therefore have strong b ions and internal fragmentation ions, rather than strong y ions, which are customary in tryptic peptide spectra.

To estimate the FDR of ncORF detections, we used the target–decoy entrapment approach[61,62]. We generate fake protein sequences at a 1:1 ratio with target sequences by scrambling the sequence of each target protein and adding it to the sequence database. This gives the search engine and post-processing software the opportunity to assign some spectra to decoy sequences (presumed to be wrong), and then the number of target-sequence mistaken assignments can be estimated as 1:1 with the number of decoy-sequence mistaken assignments. The entrapment part of the approach is to ensure that the software pipeline is not given knowledge of the decoys. The approach is not perfect as decoys do not model target sequences perfectly; in one direction, the similarity of some real protein sequences to one another may cause a small bias toward making errors to target sequences, while, in the other direction, scrambled sequences may yield an observable sequence (not already in the target database) occasionally by pure chance. Despite small imperfections, the technique is widely regarded as sufficiently accurate and is the standard in the community. Supplementary Table 16 provides FDR metrics at the PSM, peptide and protein levels, as well as for certain subsets of proteins, including the neXtProt core proteome, the 7,264 Ribo-seq ncORFs, as well as all CONTRIB sequences, many of which are putative ncORFs.

The search database was the 2023-07 THISP level 4 database[52] (https://peptideatlas.org/thisp/), which included the 7,264 Ribo-seq ORFs from ref. 4 as well as other contributed sequences that might be translated. 299 common contaminants based on the list from ref. 63 minus the human proteins are included in the search database (available at https://peptideatlas.org/thisp/). All datasets were searched with generic artifactual variable modifications methionine oxidation, cysteine cysteinylation, protein N-terminal acetylation, peptide N-terminal pyro-glutamic acid from glutamic acid or glutamine, and asparagine and glutamine deamidation. Static carbamidomethylation of cysteine was set for experiments that used Iodoacetamide. For samples that were treated with tandem mass tag or SILAC or enriched for phosphorylated peptides, appropriate mass modifications were applied. Statistical validation was performed as described above by the TPP. A complete list of datasets used and a summary of the search results in each build are available online (https://peptideatlas.org/builds/human/hla/).

Protein identifications and categories

Peptides are preferentially mapped using ProteomeMapper[58] (in TPP v.7.0) to the 20,389 entries (core proteome) and their isoforms of the 2023 version of neXtProt[64] taking into account all single amino acid variants encoded in neXtProt. Proteins that have 2 or more uniquely mapping non-nested (contained completely within the other) peptides of length 9 or more amino acids, together covering at least 18 amino acids are categorized as canonical by PeptideAtlas. If a protein entry meets the above two-peptide criteria with peptides that cannot be mapped to the core proteome, they are termed non-core canonical. There are nine additional categories, including indistinguishable representative, indistinguishable, representative, marginally distinguished, subsumed, weak, insufficient evidence for various scenarios of ambiguous and redundant evidence. Finally, the categories 'identical' are assigned to entries that are sequence-identical to another entry, and proteins that have no peptide evidence whatsoever are categorized as not detected. See ref. 65 for an extensive description of the PeptideAtlas protein categories. For reasons of integration with the HPP annual metrics[66–68], only sequence entries that belong to the core set of around 20,389 neXtProt[64] and UniProtKB/Swiss-Prot[69] protein-coding genes can achieve canonical status.

Manual inspection of ORF MS spectra

Despite extraordinary efforts to minimize false positives, both builds do contain some false positives, and they are most easily found mapping to proteins that are unlikely to be detected. For gene annotation purposes, manual inspection is therefore crucial to ensure that few false positives are reported for extraordinary detections, as described extensively previously[15]. We manually inspected each of the peptides corresponding to ncORFs and provided a manual categorization as well as a commentary. The manual categories are as follows: excellent (highly compelling evidence that the peptide identification is completely correct); good (the PSM is likely correct but lacks sufficient quality and coverage of the residues to provide highly compelling evidence); false positive; close but false positive (the PSM has many matching ions and is likely to be almost the correct peptidoform, but slight discrepancies indicate that the true identification is very close but not quite the listed sequence); low information (the ions that are detected are compatible with the identification, but coverage is too low to be compelling). The best peptide-spectrum match is also listed in the Supplementary Tables 2 and 6 as a USI that can be resolved and viewed online (https://proteomecentral.proteomexchange.org/usi/), or in cases in which a USI cannot be achieved, a direct URL for the spectrum in the PeptideAtlas web interface. In any case, all protein entries, peptides and spectra may be browsed using the PeptideAtlas web interface starting at the URLs provided above.

Procedure for manually validating PSMs

(1) Obtain a listing of PSMs for a given peptide in PeptideAtlas. (2) Examine PSMs until at least one PSM provides excellent evidence, and record its USI (Universal Spectrum Identifier) if available, or PeptideAtlas spectrum viewer URL if a USI is not available. For spectra without a PXD number associated with the dataset, a USI is usually not available. This is most common in Clinical Proteomic Tumor Analysis Consortium (CPTAC) datasets, for which a PXD has not been assigned. PSMs with USIs should be checked at https://proteomecentral.proteomexchange.org/usi/.

(3) Evaluate the PSM as follows. To obtain the 'excellent' rating: (i) the combination of b ion and y ion series must yield nearly complete coverage of the proposed peptidoform explanation. For tryptic or tryptic-like peptides (Arg or Lys residue on the C terminus), this will typically mean a nearly complete y ion series and a b ion series that begins at b2 and at least meets the y ion series. For the rules above and below for tryptic-like ions, swap y ion for b ion when there is a basic residue instead on the N terminus. (ii) If there are any prominent peaks beyond the last matching ion peak, suggesting that the sequence should extend with different residues, the PSM is not excellent. (iii) Any gaps in the y or b ion series must not have a plausible unannotated candidate in the gap, implying that the true identification is slightly different than the proposed identification. Such a plausible unannotated candidate must have a mass defect between the ions before and after the gap. (iv) Gaps should have a plausible explanation for low intensity, such as a y ion C-terminal to a proline. (v) For tryptic-like peptides, the y ions N-terminal to a proline should be more intense than surrounding ions, although confounding factors such loss of sensitivity at the high m/z end or other nearby prolines should also be considered. (vi) Strong b2 and corresponding a2 diketopiperazine ions are preferred in HCD spectra. There may be a gap at b1 ions as these are usually not visible unless there is an N-terminal mass modification. (vii) Internal fragmentation ions should be considered when annotating peaks, especially for peptides without basic residues at either terminus. (viii) Mass modifications should be kept to a minimum. (ix) For long peptides especially, there must not be a substantial region with no ions. (x) There should be no prominent unannotated peaks that suggest contamination or misassignment. Internal fragmentation ions and neutral losses should be considered for peaks that are not attributable to ordinary b and y ions.

A flowchart that describes the decision process and the outcome groups is shown in Extended Data Fig. 2e.

Gene annotation

The gene annotation work in this study has been carried out as part of the ongoing GENCODE project using existing workflows[11].

Annotating immunopeptidomics MS runs

All HLA-I MS runs were annotated for the source material (cancer versus non-cancer and cell line versus non-cell line) (Supplementary Table 8). These annotations were largely based on what was documented by PeptideAtlas but, for several instances, the category was changed based on data in the publication corresponding to the MS run. HLA typings of MS runs were determined by manually searching the publications corresponding to each MS run. For 4,879 MS runs, the full four-digit HLA typing could be retrieved.

Categorizing HLA peptides

Starting with the 865,922 peptides from the Human HLA PeptideAtlas 2023-11 build, 99 peptides starting with LLLLLLL, PPPPPPP or QQQQQQQ were filtered out. Mappings to entries starting with DECOY, CONTRIB_smORFs_Cui, CONTRIB_sORFs, CONTRIB_Fedor, CONTRIB_Bazz, CONTRIB_HLA, CONTRIB_GENCODE_nearcognate were ignored. All peptides with a length of at least 8 amino acids, mapping to UniProtKB/Swiss-Prot entries with at most 30 distinct mappings were considered to be derived from canonical proteins. These criteria are less strict than those used by PeptideAtlas to avoid mapping canonical protein derived peptides to ncORFs. Peptides with a length of at least eight amino acids, mapping to ncORFs and not canonical proteins, with at most ten distinct mappings were considered to be derived from ncORFs. For peptides with mappings against multiple ncORFs, one ncORF was selected based on the first one alphanumerically. All remaining peptides (those not assigned to canonical proteins or ncORFs) were put in the 'other peptides' category. This 'other peptides' category contains mainly peptides that map to more than 30 canonical proteins, but also includes among others peptides mapping to more than 10 ncORFs and peptides with a length of 7 amino acids.

ncORF expression in cancer tissues

To determine whether ncORFs were preferentially expressed in cancer or non-cancer tissues, each ncORF peptide was categorized to originate exclusively from MS runs from cancer samples, exclusively from MS runs from non-cancer samples or from both. Moreover, each ncORF (and corresponding peptides) was classified to originate from a cancer gene based on the Cancer Gene Census genes (accessed 4 January 2024)[70].

ncORF expression in enriched ubiquitination datasets

Public enriched datasets for ubiquitination generated from human cell lines using Thermo Scientific instruments were manually selected from the PRIDE database: PXD003936, PXD006201, PXD019692, PXD019854, PXD020909, PXD022367, PXD023218, PXD023889, PXD025890, PXD027328 and PXD037009. Datasets were reanalysed independently. The search database included the UniProt human reference proteome (one protein per gene, downloaded in April 2024), the 7,264 GENCODE Ribo-seq ORFs and cRAP protein sequences as a contaminants database. Peptide and protein identification was performed using the Comet search engine (v.2024). Default parameters were applied, with the following exceptions: semitryptic digestion, missed cleavages were set to 4 and variable modifications included ubiquitination (Gly–Gly enrichment method), oxidation of methionine and N-terminal protein acetylation. A reversed decoy database was also used to estimate the FDR. Postprocessing of the results was performed using PeptideProphet, iProphet and PTMProphet from the TPP (v.7.1.0). False-localization rates were estimated using a decoy amino acid approach (alanine)[71].

HLA binding predictions

Binding predictions were performed using NetMHCpan (v.4.1)[72,73]. Predictions were done for MS runs with a known four-digit HLA typing. For nine MS runs with A24:01, B43:01 or C12:01 as one of the alleles, no predictions could be made because these alleles were not known to NetMHCpan. Supplementary Table 8 shows an overview of MS runs for which binding predictions were made. Peptides were predicted to bind to an allele if the rank score was smaller than or equal to 2. If the HLA typing of an MS run consisted of multiple alleles, the peptide was assigned to the allele with the lowest predicted rank score, irrespective of whether this rank score was smaller than 2 or not.

Detectability determinants

Canonical proteins were categorized as detected and undetected based on whether they were detected by a single peptide. Canonical proteins shorter than 16 amino acids and proteins with amino acid symbol 'U' in their sequence were filtered out. ncORFs sequences were categorized similarly to the canonical proteins. Contrary to most other analyses, peptides were not exclusively assigned to a single ncORF, due to which the number of detected ncORFs was larger than in Extended Data Fig. 4b. For the ncORF analysis taking into account only the first or last 30% of the sequence, the requirement was that this 30% was again

16 amino acids long. Significance was determined using a two-sided Wilcoxon rank-sum test.

Hydrophobicity analysis

For the hydrophobicity analysis, all sequences were aligned by the C terminus. Starting at that position and moving towards the N terminus, the average hydrophobicity of the 15 previous amino acids across the sequences was determined. For every position, only sequences long enough to still contain 15 amino acids before the position were taken into account. A line was fit through measurements using local polynomial regression fitting. 95% confidence intervals were determined using a two-sided *t*-test.

Expression analysis

To compare the expression of detected and undetected ncORFs, we used data from GTEx[74]. The mean FPKM of all genes per tissue (excluding testis) was used. Tissues from the same organ (for example, all brain-derived tissues) were grouped together. For each ncORF, the expression was determined using the gene IDs. Contrary to most other analyses, peptides were not exclusively assigned to a single ncORF, due to which the number of detected ncORFs was larger than in Extended Data Fig. 4b. However, for 326 ncORFs the associated gene ID was not present in GTEx, so these were excluded.

Tissue comparison

For comparing the expression of ncORFs in tissues, the data from the HLA Ligand Atlas (PXD019643) was used[22]. Tissue names were extracted from the MS run file names. For each tissue, the number of distinct ncORF and CDS peptides was determined, as well as the percentage of ncORF peptides. Statistical significance was determined using multiple Fisher's exact tests and Holm–Bonferroni multiple testing correction. Gene expression levels were determined using mean FPKM values per gene across tissues from GTEx[74]. Only genes that expressed ncORFs in the HLA Ligand Atlas that were present in GTEx were considered. A selection of GTEx tissues that showed resemblance to the HLA Ligand Atlas tissues was used. Resemblance was based on the similarity of the HLA Ligand Atlas and GTEx tissue names.

Analysis of Ribo-seq data

We manually inspected Ribo-seq data for 183 ncORFs with at least one peptide nominated in the non-HLA build and 699 ncORFs with at least one peptide nominated in the HLA build. We used the GWIPS-viz browser[75] to assess evidence of ncORF translation with a publicly accessible web portal that enables the research public to examine our assessment of these ncORFs independently. The GWIPS-viz parameters were as follows: the elongating ribosomes (A-site) with the global aggregate track on 'full', which reflects native Ribo-seq; and the initiating ribosomes (P-site) with the global aggregate track on 'full', which represents Ribo-seq signal from ribosomes enriched at initiation sites. We independently evaluated the native Ribo-seq and 'initiation' Ribo-seq data. For each of these data types, we classified the data as insufficient, sufficient or excellent for supporting the translation of a given ncORF. A given ncORF was considered to be verified at the level of Ribo-seq data if either the elongating ribosome or initiating ribosome track data were sufficient or excellent. We defined excellent if there were four sequential clearly identified Ribo-seq peaks in-frame within the first 100 nucleotides of the ncORF. We defined sufficient if there were three sequential clearly identified Ribo-seq peaks in-frame within the first 100 nucleotides of the ncORF. We defined insufficient if there were not clearly sequential in-frame reads. We additionally collated selected ncORFs in the GENCODE set that were first identified in refs. 1,76 but considered insufficient in the GWIPS-viz database. As GWIPS-viz does not include the data for these two studies, we evaluated the raw data for these selected ncORFs in the primary datasets and categorized their support according to these data. We additionally calculated the percentage of in-frame ribosome footprints (PIF) and uniformity of ribosome coverage for each of the ncORFs supported by one or more peptides in each PeptideAtlas build, as observed in the human body map.

Subsequently, a second phase of manual inspection was conducted on the subset of candidates that were assigned tier 1A supporting evidence. The smaller scale of this second phase effort enabled a more nuanced assessment of Ribo-seq support for these candidates where, alongside GWIPS-viz, we explored each candidate in the transcriptome browser, Trips-Viz[77], and the more richly populated RiboCrypt (RiboCrypt.org), which contains over 3,600 ribo-seq libraries (available in the RiboSeq.Org portal)[78] aligned to the human genome. Both additional browsers apply dynamically calculated read-length-dependent offsets leading to periodic signals in the profiles that is clearer than that in GWIPS-viz where fixed offsets are applied for all lengths. These assessments looked to validate the support for the translation initiation and termination sites for each candidate while accounting for transcript isoform complexity and regions of poor genomic mappability. Ribo-seq area plots were obtained using RiboCrypt with the sliding-window function calculating moving average of length 9 of P-site coverage to make informative concise graphics. Zoomed-in regions were obtained with the 'columns' setting and no sliding window to maintain single-nucleotide resolution. Through this assessment, higher confidence can be placed in the supporting translation evidence for these tier 1 annotations.

Cross-species comparison of *GMCL1* translation

To inspect the translation of *GMCL1* uORF (c2riboseqorf47) and its downstream CDS across mammals, we analysed a total of 35 Ribo-seq datasets from the cardiac left ventricle, along with their corresponding RNA-seq datasets, across five mammalian species[1,79]. Datasets were aligned to the appropriate Ensembl genomes and transcriptomes (release 98), including human (GRCh38/hg38), chimpanzee (Pan_tro_3.0/panTro5), macaque (Mmul_10/rheMac10), mouse (GRCm38/mm10) and rat (Rnor_6.0/rn6). Mapping was performed using STAR (v.2.7.3a)[80] with the standard settings and the following modified parameters: --outSAMtype BAM SortedByCoordinate, --outFilterMismatchNmax 2, --outFilterMultimapNmax 20, --alignSJDBoverhangMin 6, --alignSJoverhangMin 500, --outFilterType BySJout, --limitOutSJcollapsed 10000000, --limitIObufferSize 300000000 and --outFilterIntronMotifs RemoveNoncanonical. Later, RiboseQC[81] was used to extract P-site positions from the Ribo-seq data. Ribo-seq reads, RNA-seq coverage, and in-frame P-site distributions over the *GMCL1* locus were visualized using the UCSC Genome Browser[82] for the five selected mammalian species.

Use of the tier classification system

We used the tier classification system for ncORFs initially proposed previously[25]. Specifically, ncORFs were given an initial or provisional tier based on the information available from the large-scale MS search. After manual review of the nomination data, ncORFs were then assigned a final tier. Tiers were defined as follows:

- Tier 1A: two non-nested peptides in MS proteome data, with or without HLA immunopeptidomics data, with Ribo-seq data
- Tier 1B: two non-nested peptides in HLA immunopeptidomics data with Ribo-seq data
- Tier 2A: one peptide in MS proteome data, with or without HLA immunopeptidomics data, with Ribo-seq data
- Tier 2B: one peptide in HLA immunopeptidomics data with Ribo-seq data
- Tier 3: any HLA immunopeptidomics and/or tryptic proteome LC–MS/MS evidence without Ribo-seq evidence
- Tier 4: Ribo-seq evidence without proteomic evidence
- Tier 5: in silico prediction of an ORF on an expressed transcript without any Ribo-seq or proteomic evidence.

MLP classifier model

A dataset comprising 677 ncORF peptide sequences of 9 amino acids, each annotated with 22 attributes, was used to develop a multilayer perceptron (MLP) classifier model. The implementation was carried out using Python 3 and the following software libraries: pandas, numpy, matplotlib and scikit-learn (v.1.5/1.6). The dataset was processed by separating the features from the target variable. The data were then split into training and testing sets, with 80% allocated for training and 20% for testing. To ensure reproducibility, the random state was set to 42 during the split. Before model fitting, the features were standardized using StandardScaler. This preprocessing step involved removing the mean and scaling the features to have unit variance, thereby normalizing the data. The MLP classifier model was initialized with a maximum of 8,000 iterations and a random state of 42 to ensure reproducibility. The model was then subjected to hyperparameter tuning using grid search with cross-validation. The hyperparameters explored included: hidden layer sizes: (280); activation function: 'tanh'; and regularization parameter (alpha): 0.01. Grid search with cross-validation was used to systematically evaluate the performance of various hyperparameter combinations and identify the optimal configuration. The best-performing model, as determined by the grid search results, was selected and fitted to the training data. Subsequently, this model was used to make predictions on the test set, which had not been seen by the model during training. The performance of the model was assessed using standard evaluation metrics to determine its predictive capabilities.

TensorFlow Keras model

Due to 1,785 ncORFs being detected while 5,479 remain undetected, presenting an approximate ratio of 1:3, a balanced weight for imbalanced dataset was used to address the imbalance and a neural network analysis to build, train and evaluate a TensorFlow Keras model (v.2.18.0). The dataset included 7,264 ncORF amino acid sequences with a selection of 43 attributes. Using Python 3, we imported the necessary libraries including TensorFlow, Keras and various components of TensorFlow and Keras for building and evaluating the model. Using the train_test_split function from scikit-learn, we allocated 80% for training and 20% for testing after separating the training and testing sets from the target variable. The features were standardized using the StandardScaler. A sequential model consisting of multiple layers was built with an input of 16 neurons, ReLU activation, and L2 regularization. To prevent overfitting, we added batch normalization and dropout layers after each hidden layer. The output layer consisted of a single neuron with sigmoid activation for binary classification. We compiled the model using the Adam optimizer with a learning rate of 0.001, binary cross-entropy loss function, accuracy as the metric and it was trained on the training data. During training, it was run for a total of 60 epochs with a batch size of 12. The target variable for the test set was predicted by the model.

Evolutionary conservation and constraint (ORBL)

To compute the ORBLv conservation score of an ORF for the placental and primate clades, we extracted the local alignment of the ORF from the 120 mammal whole-genome alignments with the hg38 reference[83] downloaded from https://bds.mpi-cbg.de/hillerlab/120MammalAlignment/Human120way/, and restricted to the 116 placental mammal or 26 primate subsets, respectively. ORBL scores for additional clades, as included in Supplementary Table 11, used whole-genome alignments of 100 vertebrates (downloaded from https://hgdownload.gi.ucsc.edu/goldenPath/hg38/multiz100way/), 470 mammals (downloaded from https://hgdownload.soe.ucsc.edu/goldenPath/hg38/multiz470way) and 447 mammals (downloaded from https://cgl.gi.ucsc.edu/data/cactus/447-mammalian-2022v1.fix2.maf.gz). We considered the start and stop to be conserved in a species if the bases (possibly more than three) aligned to the start and stop of the ORF in the reference species included three consecutive bases that were ATG or any stop codon, respectively. We considered the reading frame to be conserved if the aligned bases between the aligned ATG and aligned stop were a multiple of three in total (even if there were insertions or deletions that shifted part of the reading frame), and did not include any in-frame stop codons or gaps in the alignment. We considered the ORF to be conserved in a species if the start, stop and reading frame were conserved. We defined ORBLv to be the phylogenetic branch length of the conserved species divided by the branch length of all 116 placental or 26 primate species in the whole-genome alignments (not just the ones present in the local alignment). In Extended Data Fig. 7b,c plotting ORBLv distributions of ncORFs and MANE Select CDS, we excluded selenoproteins and MANE Select CDS having non-ATG starts, as these would not be considered ORFs by our definition.

As our analysis of evolutionary constraint is strongly influenced by overlap with CDS, we redid the biotype determination for the ncORFs using GENCODE v42 annotations (the original biotypes from ref. 4 used v35), and applied strict criteria to determine ORFs with a 'pure' biotype (details are provided in the Supplementary Results).

There were 448 of the 7,264 ncORFs that did not satisfy any of these criteria, for example, ORFs that overlapped CDS from two different transcripts in different reading frames; we considered these to have 'mixed' biotype and did not compute an ORBLq constraint score for them.

The list of 'untranslated' ORFs from which matched ORFs were chosen for the computation of ORBLq was created as follows. We began with every ATG-initiated ORF of any length in any protein-coding or lncRNA transcript in GENCODE v42 that did not overlap a protein-coding CDS of any transcript in the same frame. 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. We then excluded ORFs that did not have a 'pure' biotype as defined in the Supplementary Results; that overlapped one of the 7,264 ncORFs, an mRNA on the opposite strand, or a pseudogene on either strand; or was a dORF that overlapped some 5′-UTR. We segregated uoORFs, intORFs and doORFs by the frame in which they overlapped the main ORF (+1 or +2) as constraint on the main frame amino acid sequence imposes different ORFness constraint on the two overlapping frames. We were left with 1,717,927 untranslated ORFs. The counts by biotype were: uORF (63,795), uoORF+1 (3,231), uoORF+2 (2,940), intORF+1 (320,823), intORF+2 (60,135), doORF+1 (16,910), doORF+2 (5,946), dORF (653,983) and lncRNA-ORF (590,164). We then computed the ORBLv conservation score for each of these untranslated ORFs.

To compute ORBLq of an ncORF, we selected a matched set of at least 1,000 untranslated ORFs having the same biotype as the ncORF (and the same frame of overlap for uoORFs, intORFs and doORFs), starting with the untranslated ORFs of the same length as the ncORF. If there were fewer than 1,000 of the same length and biotype (which was common for longer ncORFs and ones having less frequent biotypes such as uoORF), we added in all ORFs of the same biotype having length one more or one less than the ncORF, length two more or two less, and so on, until there were at least 1,000 matched ORFs. We then estimated the probability that a similar untranslated ORF would have at least as much conservation as our ncORF, a kind of P value, as

$$P = \frac{(\text{Number of matched ORFs having ORBLv} \geq \text{ORBLv*}) + 1}{\text{Number of matched ORFs}}$$

where ORBLv* is the ORBLv of our ncORF, and the pseudocount of 1 is added to prevent P values of 0 (it was not added if the result would be more than 1). We then calculated the ORBL quantile, ORBLq, as $1 - P$.

For some analyses, we used an information content measure, $-\log_{10}[1 - \text{ORBLq}]$, rather than ORBLq itself, to linearize values near 1.

Lentiviral transduction for CRISPR screens

Optimal infection conditions were determined in each cell line to achieve around 30–50% infection efficiency, corresponding to a multiplicity of infection (MOI) of about 0.3–0.5. Final transductions were performed in 12-well plate format with 3×10^6 cells per plate using a polybrene concentration of 4 μg ml^{-1} and different virus volumes to achieve the desired MOI. Approximately 24 h after infection, cells were trypsinized and approximately 2×10^5 of A375 and A673 cells, 3×10^5 of CADO-ES1, Jurkat and THP-1 cells, and 1×10^6 of HepG2 and RD-ES cells from each infection were seeded in 12 wells of 6-well plates, each with complete medium, one supplemented with 1 μg ml^{-1} of puromycin for A673, Jurkat, RD-ES and THP-1 cells, 1.5 μg ml^{-1} of puromycin for A375 and HepG2 cells, and 3 μg ml^{-1} of puromycin for CADO-ES1 cells. Cells were counted 4–5 days after selection to determine the infection efficiency, comparing survival with and without puromycin selection. Volumes of virus that yielded around 30–50% infection efficiency were used for screening.

CRISPR screening

We selected 2,196 ncORFs from the GENCODE Phase I catalogue[4] and 1,245 ncORFs from a previous ncORF gRNA library[7]. ncORFs were selected according to the following rules: (1) there are a maximum of three ncORFs per gene selected. Exclude all intORFs as well as doORFs and uoORFs that have ≥25% overlap with the main CDS. (2) Minimum ncORF size of 12 amino acids to enable sufficient gRNA targeting sites. (3) ncORFs with <4 predicted targeting gRNAs were excluded. (4) The overall ncORF distribution in the final library was as follows: tier 1A (n = 4), tier 1B (n = 224), tier 2A (n = 13), tier 2B (n = 373), tier 3 (n = 13), tier 4 (n = 34) and no tier (n = 1,535).

The lentiviral barcoded library contained 27,464 sgRNAs, which were designed using the CRISPick program (https://portals.broadinstitute.org/gppx/crispick/public) from the Broad Institute Genomic Perturbation Platform, using settings for the reference genome Human GRCh38 (Ensembl v.108) for 'CRISPRko' with enzyme 'SpyoCas9 (NGG)' with the following modifications: (1) for ncORFs with ≥2 exons, all gRNAs could not come from the same exon. (2) A target of 6 gRNAs per ncORF were designed, and a target of 3 gRNAs per parental CDS were designed. (3) The spacing requirement for gRNA separation was reduced to 1% across the total target length for ORFs and maintained at 5% for parental CDSs. (4) A 2:1 on-target to off-target ratio was used. (5) gRNAs with ≥5 predicted mapping sites to the human genome were removed. (6) A previously published[7] set of 471 pan-essential gRNAs, 503 non-targeting gRNAs without genome cutting and 497 non-targeting gRNAs with genome cutting were included.

Lentiviral infections were performed in biological triplicate with sufficient cells to achieve a representation of 500 cells per gRNA after puromycin selection. Then, 24 h after infection, cells were selected with 1.5 μg ml^{-1} of puromycin for 7 days, and then approximately $1–2 \times 10^7$ cells were collected for assessing the initial abundance of the library. Cells were passaged every 3–4 days and collected about 14 days after infection. For all genome-wide screens, genomic DNA (gDNA) was isolated using Midi or Maxi kits for the validation screens gDNA was isolated using Midi kits according to the manufacturer's protocol (Qiagen). After PCR amplification of barcodes, the samples were sequenced on the HiSeq2000 or NextSeq (Illumina) system. For analysis, the read counts were normalized to reads per million and then $\log_2$ transformed. The $\log_2$[FC] of each sgRNA was determined relative to the initial timepoint for each biological replicate.

Analysis of CRISPR screening data: sgRNA mapping and hit calling

We began by standardizing the sgRNA nomenclature across each library. For sgRNA targeting locus assignment, we used a two-step computational approach to ensure accurate and comprehensive mapping of guide RNAs to their genomic targets. Initially, Bowtie2 v.(2.5.4) performed sequence alignment using stringent parameters (N=0, gbar=1, ma=2, mp=0,0), optimized for short RNA sequences. This allowed for up to 100 potential mapping locations per sgRNA while maintaining high specificity. Subsequently, SAMtools (v.1.20) and BEDTools (v.2.31.1) processed these alignments to generate genomic coordinates and intersect them with gene annotations. These mapping results were then integrated with GENCODE v45 gene annotations and GENCODE Phase 1 ORF definitions to create a harmonized dataset. During this integration, sgRNAs mapping to more than five genomic coordinates were filtered out, and ORFs with fewer than three guides were excluded, yielding the final mapping files. For multi-ORF sgRNAs, a hierarchical naming scheme prioritized primary targets, while shared transcript annotations were preserved.

We removed sgRNAs with low counts less than 3 s.d. below the total counts. Additional quality-control measures included evaluating replicate consistency and log-transformed fold change. Chronos (v.2.0.8) was used to process the raw counts from the CRISPR screen, incorporating copy-number variation (CNV) data. CNV information was retrieved from https://depmap.org; if not accessible, the CNV value was set to 1 (ref. 84). To assess the quality of the CRISPR screen, the mean median absolute deviation (MMAD) was calculated for each library, measuring the distance between positive and negative controls. Chronos outputs fitness scores for targeted loci, with loss-of-function hits defined by scores greater than 0.5 or less than −0.5.

For more stringent hit selection, CRISPR screen results were cross-validated with RNA-seq and ribosome profiling data for the 8 overlapping cell lines. After data processing, only genes with expression levels ≥5 TPM in Ribo-seq and ≥10 TPM in RNA-seq were retained, reducing the likelihood of off-target effects and refining functional hit calls.

To further dissect the functional relevance of ncORFs, we computed normalized phenotypic effects by subtracting the Chronos score of the associated canonical coding sequence (CDS) from that of the ncORF. This normalization was performed only for ncORFs shown to be co-transcribed with a canonical ORF, ensuring that shared transcript context was preserved in the interpretation of fitness effects

Meta-analysis of CRISPR data

Three independent CRISPR screens (refs. 7,28 and the current dataset generated for this study) were integrated by first mapping sgRNA sequences to the GRCh38 reference genome. Target annotations were derived using GENCODE v45 augmented with phase 1 ncORF annotations (https://www.gencodegenes.org/pages/riboseq_orfs/). sgRNAs mapping to more than five genomic loci or exhibiting poor alignment quality were excluded from further analysis. This was then used as the mapping design. Raw counts data were obtained from the original publications. Gene-level essentiality scores were calculated using Chronos[84]. CNV data for each cell line were retrieved from depmap (release 25Q2). And for missing lines CNV was set to 1. Genes with a Chronos score below −0.5 were considered to be significant hits, following the authors' recommended threshold. ORFs where 60% of the samples show essentiality were considered pan-essentiality.

Comparison of CRISPR–Cas9 with Cas13 data

The CRISPR–Cas13 dataset was retrieved from ref. 85. Among the available cell lines, THP-1 was the only line overlapping with the Cas9 screens used in this study and was therefore selected for comparative analysis. Only targets shared between the Cas9 and Cas13 datasets were considered. For both platforms, normalized gene-level depletion scores were used to assess concordance, focusing on overlapping protein-coding and non-coding transcripts. Comparisons were restricted to high-confidence targets based on consistent guide performance and coverage in both datasets.

Analysis of CRISPRa data

Human CRISPRa Calabrese (P65 HSF) Pooled Libraries data were retrieved from ref. 31. Screens from Meljuso and A375 cell lines were reprocessed by first removing low-representation targets. Using the raw counts from the original manuscript and a harmonized mapping dictionary, fitness scores were calculated using Chronos[84].

CRISPR tiling analysis and functional enrichment score for ncORF

Two CRISPR tiling screens[7,28] were reprocessed by first standardizing sgRNA nomenclature across libraries. $\log_2$[FC] values were calculated by comparing the late-timepoint to the input plasmid DNA or earliest time point. For each cell line, gRNA with low representation were excluded, then library depth was corrected using CPMs, using the geometric mean of guides targeting positive control genes (that is, pan-essential genes). Normalized $\log_2$[FC] = (gRNA $\log_2$[FC]/mean positive control $\log_2$[FC]) × −1

This scaling anchors positive control dropout to −1, allowing consistent interpretation across experiments. We then integrated genomic mapping data (coordinates and annotations) with normalized $\log_2$[FC] values to evaluate ncORF essentiality. ORF-level effects were denoised using empirical Bayes shrinkage in the ashr package (v.2.2.63), fitting a global prior distribution and shrinking individual ORF estimates accordingly. For ORFs with >6 sgRNAs, we applied median absolute deviation (MAD) filtering, removing guides with $\log_2$[FC] values > 3 MAD units from the median to reduce the impact of outliers. To assess significance, we used a gene-specific permutation-based null model accounting for gene-level background of loss of effect. For each ncORF with n guides, we generated null distributions by resampling n guides from all guides targeting the same parental gene. Summary statistics (mean or median $\log_2$[FC]) were calculated for each permutation (typically 1,000–4,000 iterations). One-sided P values were calculated as P = (1 + number of null statistics ≥ observed statistics)/(1 + B), where B is the number of permutations. These were then converted to two-sided P values using $P_{\text{two-sided}} = 2 \times \min(P_{\text{one-sided}}, 1 - P_{\text{one-sided}})$. This approach accounts for guide count variability and gene-specific effects, enabling robust detection of essential ncORFs within the hierarchical structure of overlapping gene models.

Pooled c10riboseqorf92 knockout

Pooled knockout screens in the PRISM cell line set were performed as previously described[28]. This approach uses 486 barcoded human cancer cell lines, which are pooled and grown together in RPM1640 medium supplemented with 10% FBS. gRNAs used were non-cutting LacZ control (AACGGCGGATTGACCGTAAT), cutting control Chr2-2 (GGTGTGCGTATGAAGCAGTGG), c10riboseqorf921 (ACAGGGCACTG GTCTCCCAA) and c10riboseqorf92 2 (CAAGGCTGTATATTTCACCT). For pooled screening, 400,000 pooled cells per well were plated in a 6-well plate with a cell pellet collected for a no infection control. Then, 24 h later, cells were subjected to lentiviral infection with gRNA and Cas9 using an all-in-one plasmid. A lentiviral MOI of 10 was used, and transduction was performed with 4 μg ml^{-1} polybrene. On day 4, the cell culture medium was changed to include 1 μg ml^{-1} puromycin for 72 h, after which antibiotic-free medium was used. Cells were then passed every 72 h and a cell pellet (2×10^6 cells) was collected for DNA on days 6, 10 and 15. Genomic DNA extraction of cell pellets was performed using the DNA Blood and Tissue Kit according to the manufacturer's instructions (Qiagen). Cell line representation was determined by amplifying barcodes with universal primers. PCR products were pooled and purified with AMPur beads (Beckman Coulter). The DNA concentration was measured using Qubit fluorometric quantification (Thermo Fisher Scientific). DNA was sequenced on the NovaSeq (Illumina) system at the Genomics Platform at the Broad Institute.

Analysis of pooled c10riboseqorf92 knockout data

At day 15/20, 471 out of 486 cell lines were detectable and were used for data analysis. Cell line abundance was determined by RNA expression of each cell line's barcode using RNA-seq as previously described[28]. $\log_2$-transformed fold changes in abundance were calculated by comparing each cell line's day 15 abundance to that of the input plasmid pool. Replicates from days 15 and 20 were averaged. To identify outliers, linear regression was performed in R (v.4.4.2) to model sgRNA2 viability as a function of sgRNA1 viability at day 15, and residuals were computed from the fitted model. Cell lines with absolute residuals exceeding two standard deviations from the mean were flagged as outliers and excluded, along with those containing missing values (NaN). To assess phenotypic similarity between c10riboseqORF92 knockout and genome-wide perturbations, we correlated its viability profile with CRISPR screening data from the DepMap project (release 25Q2). Gene-level dependency scores were retrieved for all genes and cell lines, and only those overlapping with the c10riboseqORF92-KO dataset were retained. Spearman correlation coefficients were then computed between the c10riboseqORF92 profile and each gene's profile across the matched cell lines.

siRNA and overexpression experiments

Four siRNAs targeting distinct regions of *OLMALINC* (siRNAs 10, 11, 12 and 13) were tested in A375 and A549 cell lines to evaluate knockdown efficiency. Cells were seeded into six-well plates and transfected when 80% of confluence was reached, using the Mirus TransIT-X2 transfection kit according to the manufacturer's instructions. At 48 h after transfection, total RNA was extracted using the Zymo Total RNA Isolation Kit. The knockdown efficiency was assessed by quantitative PCR (qPCR) and analysed using QuantStudio Design & Analysis Software (v.1.6.1), using *GAPDH* and *ACTB* as the housekeeping genes. Expression levels of *OLMALINC* and the associated c10riboseqORF92 were quantified. The same transfection and RNA isolation protocol was used for generating the samples used in bulk RNA-seq analysis. qPCR primers were as follows: *OLMALINC* forward, AGGAACATCTTGCCAATTTCA; *OLMALINC* reverse, TGTGGATCTTCAGTTGCTTCA; *GAPDH* forward, TGCACCACCAACTGCTTAGC; *GAPDH* reverse, GGCATGGACTGTGGT CATGAG; *ACTB* forward, AAGGCCAACCGCGAGAAG; *ACTB* reverse, ACA GCCTGGATAGCAACGTACA; P1 sense primer for plasmid expression, TCTTGTGGAAAGGACGA; P2 antisense primer for plasmid expression, TTAAAGCAGCGTATCCACATAGCGT.

To assess the impact of *OLMALINC* knockdown on cell proliferation, 4 replicates of 2,500 cells per well were seeded into 24-well plates and transfected with siRNAs as described above. Cell proliferation was monitored in real-time using the Incucyte Live-Cell Analysis System. Images were acquired at 8 h intervals, and the cell confluence (percentage surface area occupied) was measured over time. The culture medium was refreshed every 24 h throughout the experiment. Proliferation curves were generated using the confluence data exported from the Incucyte software. Growth curves were fitted in R using the growthcurver (v.0.3.1) package, which models logistic growth and computes the AUC as a summary metric of proliferative capacity. AUC values were used for statistical comparison of proliferation across treatment conditions.

Bulk RNA-seq and data analysis

Cell lines were cultured in six-well plates until around 80% confluence. Cells were collected by gentle scraping, washed with 1× PBS, and pelleted by centrifugation. Cell pellets were lysed in RIPA buffer (Pierce 8900) supplemented with 1× protease inhibitor cocktail (cOmplete Mini, EDTA free, Roche) and homogenized on ice for 15 min. Total RNA was extracted using the Zymo Research Direct-zol RNA Miniprep Kit (R2052) according to the manufacturer's protocol. The RNA concentration and purity were initially assessed using a NanoDrop

spectrophotometer, and preliminary quality control was performed to assess potential contamination.

RNA-seq libraries were prepared by the University of Michigan Advanced Genomics Core using a poly(A) enrichment protocol. The fragment size distribution was evaluated using the Agilent 2100 Bioanalyzer (Agilent Technologies). Libraries were sequenced on either the Illumina NovaSeqX platform or Element Biosciences AVITI24 system, generating 300-cycle paired-end reads. RNA-seq count tables were obtained using Nextflow (v.25.04.2; https://www.nextflow.io), based on nf-core/rnaseq v.3.1.19. Raw FASTQ files were quality-checked with FastQC (v.0.11.9) and adapter trimming was performed using Trim Galore (v.0.6.7). Reads were aligned to the human reference genome (GRCh38.p14) using STAR, with the --twopassMode Basic option enabled. Gene-level and isoform-level expression quantification were performed using RSEM (v.1.3.3), referencing GENCODE v45 gene annotations. Differential gene expression analysis was conducted using DESeq2 (v1.46.0) in R (v.4.4.2). A multifactorial design formula was used to model the main effects of treatment (siCtrl versus siOLMALINC knockdown), genetic background (c10riboseqorf92 wild-type versus overexpression) and their interaction: design = ~ treatment + background + treatment:background. Genes with an adjusted *P* value (Benjamini–Hochberg FDR) < 0.05 and an absolute $\log_2$[FC] > 0.5 were considered significantly differentially expressed. $\log_2$[FC] shrinkage was applied using the apeglm method. Functional enrichment analysis was performed using the clusterProfiler R package (v.4.14.6), querying the MSigDB Hallmark gene sets (v2025.1.Hs). Enrichment was considered significant at adjusted *P* < 0.05.

Multiplexed single-cell transcriptional response

In total, 21 human cell lines expressing SpCas9 were individually cultured in RPMI-1640 medium supplemented with 10% FBS. Before lentiviral transduction, cell lines were grouped into two pools based on doubling times to ensure balanced growth dynamics. Each pool was transduced with four sgRNAs: a non-targeting cutting control (Chr2-2: GGTGTGCGTATGAAGCAGTGG), two targeting c10riboseqorf92 (sg 1: ACAGGGCACTGGTCTCCCAA; sg 2: CAAGGCTGTATATTTCACCT), and a positive control targeting KIF11 (CAGTATAGACACCACAGTTGG). Each virus was applied at four concentrations to empirically approximate a multiplicity of infection (MOI) of 1. Then, 24 h after infection, cells were selected with puromycin (5 μg ml^{-1}), and the selection medium was refreshed every 48 h. On day 7 after infection, cells were trypsinized; both adherent and suspension fractions were combined, pelleted and resuspended in cell capture buffer. The two pools were then merged, preserving equal representation of each original cell line. Before single-cell library preparation, cell viability (>90%) and total counts were confirmed. The samples were processed by the University of Michigan Advanced Genomics Core using the 10x Genomics Chromium platform with GEM-X On-chip Multiplexing and 3′ Gene Expression v4 chemistry. Target recovery was 5,000 cells per sample (approximately 200 cells per condition), with a sequencing depth of 30,000 reads per cell. Libraries were sequenced on the Illumina NovaSeq system using 10B 300-cycle kits.

Raw sequencing data were processed using Cell Ranger v.9.0.1 with the multi function and aligned to the GRCh38 reference (refdata-gex-GRCh38-2024-A). Cells were filtered using MAD-based outlier detection (5 MADs for log-transformed total counts, log-transformed gene counts and percentage of counts in top 50 genes; 3 MADs for mitochondrial percentage), with a hard mitochondrial cutoff of 20%. Genes detected in fewer than 20 cells were removed. Ambient RNA contamination was corrected using SoupX. Raw counts were normalized to 10,000 counts per cell and log-transformed. Highly variable genes (*n* = 2,000) were selected using the cell_ranger flavor with batch-aware selection across sample IDs. Principal component analysis was performed using 50 components. Cell cycle scores (S and G2M) were computed using Regev laboratory gene sets. Cell line identity deconvolution was performed using a combination of packages: demuxalot and dropulation, leveraging SNP profiles curated from DepMap (release 25Q2) and CellLineProject and genotype-free approach using scSplit[86]. Cell lines with under 100 cells were filtered out from downstream analysis.

Differential gene expression analysis was performed using a pseudobulk approach. Raw counts were aggregated by cell line and treatment condition using Seurat's AggregateExpression function, yielding one pseudobulk profile per cell line–condition combination. Low-expression genes were removed using edgeR's filterByExpr, and library sizes were normalized with trimmed mean of *M*-values. Differential expression was tested using limma-voom with a design matrix including condition and cell line identity as covariates (~0 + condition + scsplit_assignment). The contrast c10riboseqorf92 KO versus control was tested using moderated *t*-statistics with empirical Bayes shrinkage (eBayes, trend = TRUE). *P* values were adjusted using the Benjamini–Hochberg method; genes with adjusted *P* < 0.05 and $|\log_2$[FC]| > 0.5 were considered significant. Gene set enrichment analysis was performed on log[FC]-ranked gene lists using gseGO from clusterProfiler against GO Biological Process terms (minGSSize = 5, maxGSSize = 500, Benjamini–Hochberg-adjusted *P* < 0.01).

Gene co-expression modules were identified using the hdWGCNA package in Seurat v5. Metacells were constructed by aggregating cells via *k*-nearest neighbours (*k* = 15, max shared = 12, minimum cells = 30), grouped by cell line and treatment condition in UMAP space, and subsequently normalized. Genes expressed in at least 5% of cells were retained for network construction. A signed weighted correlation network was built using a soft-thresholding power selected as the first value achieving a scale-free topology fit ($R^2 \geq 0.8$), with a minimum module size of 30 and a merge cut height of 0.25. Module eigengenes were computed per cell line and condition. For the consensus analysis, metacells from all cell lines were pooled into a single expression matrix and a shared network was constructed. Differential module eigengene (DME) analysis was performed per cell line using the Wilcoxon rank-sum test comparing control to c10riboseqorf92 KO metacells; modules with adjusted *P* < 0.05 and |average $\log_2$[FC]| > 0.5 were considered treatment responsive. Modules showing significant and directionally consistent responses in three or more cell lines were classified as conserved. Module–trait correlations were computed per cell line against binary trait indicators using ModuleTraitCorrelation. Over-representation analysis was performed on module gene sets using clusterProfiler against GO Biological Process, GO Molecular Function, KEGG, Reactome and MSigDB Hallmark gene sets (Benjamini–Hochberg-adjusted *P* < 0.05, *q* < 0.2).

Perturbation distances for each cell line were calculated relative to the unperturbed population (that is, sgRNA Ch2-2) using the definition for E-distance as described previously[87]. Cells were projected into PCA space (15 components), and pairwise energy distances (*e*-distances) were calculated between all cell line–condition groups, defined as 2× the mean between-group distance minus the mean within-group distance. Same-cell-line comparisons were extracted to quantify the transcriptional shift of each perturbation relative to the non-targeting control (sgRNA Chr2-2) within each genetic background. Hierarchical clustering of the full *e*-distance matrix was performed using Ward linkage. Finally, consensus non-negative matrix factorization was applied to the harmonized expression matrix to identify latent transcriptional programs. The number of factors (*K*) was evaluated from 5 to 29, using 5,000 highly variable genes and 50 iterations per *K* value.

Multiplexed PRM MS of ncORF targets

HEK293, HeLa S3 and K562 cells were obtained from the American Type Culture Collection (ATCC) and checked for mycoplasma prior to assays. HEK293 and K562 cells were cultured in RPMI-1640 medium supplemented with 10% FBS and HeLa cells were cultured in DMEM medium supplemented with 10% FBS. Cells were grown to

confluence and were washed in PBS, pelleted and frozen at −80 °C before use.

PRM sample preparation. From each cell line, two equal-sized cell pellets were processed in parallel using two different sample preparation protocols as follows:

Protocol 1: one frozen pellet per cell line was lysed in 8 M guanidine hydrochloride with Tris-HCl (pH 8.5). Cell lysates were incubated at 75 °C for 10 min and homogenized using three 1.4 mm ceramic beads in a Precellys 24 homogenizer (Bertin) at 8,500 rpm 9 times for 20 s with resting for 30 s between cycles. The lysates were centrifuged to remove debris and recover the supernatant. The protein content was determined using the bicinchoninic acid assay (Thermo Fisher Scientific). An aliquot corresponding to 100 µg of protein was reduced with 5 mM tris(2-carboxyethyl)phosphine (TCEP, 20 min, 37 °C) and alkylated with 15 mM iodoacetamide (25 min, room temperature, darkness), with shaking at 900 rpm on a Thermomixer (Eppendorf). The samples were diluted with 0.1 M tetraethylammonium bromide (TEAB, pH 8.5) to ensure a final guanidine hydrochloride concentration of ≤0.5 M. Proteins were digested with trypsin (Trypsin-Gold, Pierce Thermo Fisher Scientific, 4 µg, 37 °C, 900 rpm, 16 h), quenched with 5% formic acid (final concentration), and desalted by solid-phase extraction using Atlas columns (Tecan). In brief, the columns were equilibrated sequentially with 100% acetonitrile, 70% acetonitrile/0.1% trifluoroacetic acid (TFA) and 0.1% TFA in water. The samples were applied, then columns were washed with 0.1% TFA in water, and peptides eluted in 70% acetonitrile and 0.1% TFA. The eluates were dried under centrifugal evaporation (Savant, Thermo Fisher Scientific).

Protocol 2: a second frozen pellet per cell line was subjected to acetonitrile-based small protein extraction (modified protocol for acetonitrile depletion of large proteins[88,89]). Frozen cell pellets were resuspended in ice-cold 50 mM NaCl. An aliquot corresponding to 1 mg of protein was precipitated with acetonitrile and TFA for a final concentration of 76% acetonitrile, 0.1% TFA, and 50 mM NaCl, and the samples were vortexed and incubated for 1 h at room temperature with agitation (1,400 rpm). The samples were centrifuged (18,000*g*, 20 min at room temperature), the supernatants transferred to fresh microcentrifuge tubes and dried under centrifugal evaporation (Savant, Thermo Fisher Scientific). Next, the samples were resuspended in 0.1 M TEAB (pH 8.5), reduced with 5 mM TCEP (20 min, 37 °C) and alkylated with 15 mM iodoacetamide (25 min, room temperature, darkness), with shaking at 900 rpm. Proteins were digested with trypsin (Trypsin-Gold, Thermo Fisher Scientific, 1 µg, 37 °C, 900 rpm, 16 h) on a Thermomixer (Eppendorf). The samples were quenched, desalted with Atlas columns and dried under centrifugal evaporation as described above.

Each of the two samples per cell line was dissolved in 5% acetonitrile, 0.1% formic acid in Milli-Q H_2O. An aliquot of each sample was spiked with ncORF heavy-labelled synthetic peptide analogues and iRT peptides (Biognosys) for MS analysis. Specifically, the samples were spiked so that 1 µg digested cancer cell line peptide mixture, 10 fmol synthetic peptide and 25 fmol iRT peptide were injected onto the column for sample analysis using the Orbitrap Astral mass spectrometer (Thermo Fisher Scientific) and 250 ng digested peptide mixture, 6 fmol ncORF heavy-labelled synthetic peptide and 5 fmol iRT peptide for sample analysis with the ZenoTof 8600 (Sciex).

Synthetic heavy-isotope-labelled peptide standards. For each ncORF peptide, a synthetic peptide analogue was individually chemically synthesized as free amine at the N terminus and carboxylic acid at the C terminus (Synpeptide). Cysteine residues were incorporated as carboxyamidomethylated cysteine building blocks, and one heavy-isotope-labelled amino acid per peptide was incorporated as the terminal arginine as R[13C6, 15N4] or lysine as K[13C6, 15N2]. Each peptide was dissolved in 50% acetonitrile/1% formic acid/Milli-Q water to a concentration of 2 mg ml^{-1}. Peptides were further diluted and individually analysed in 5% acetonitrile/0.1% formic acid/Milli-Q water, and peptides were pooled and then further serially diluted for spiking into the prepared cell line samples.

PRM MS data collection and analysis. PRM data were analysed manually for each peptide and each cancer cell line processed using protocols 1 and 2 using both an Orbitrap Astral (Thermo Fisher Scientific) and ZenoTOF 8600 (Sciex) mass spectrometer with the following conditions:

Orbitrap Astral: peptide concentrations were measured using the Pierce Quantitative Fluorometric Peptide Assay kit (Thermo Fisher Scientific, 23290). The samples were analysed with an Orbitrap Astral mass spectrometer (Thermo Fisher Scientific) interfaced with a Vanquish Neo UHPLC (Thermo Fisher Scientific) equipped with a C18 column EV1109 (Evosep) over a 20 min acetonitrile gradient. The instrument was run in PRM mode with the following parameters: automatic gain controlled at 50%, quadrupole isolation window width (IsoWidth) at 2 Th, maximum injection time for MS2 acquisition (MaxIT) at 20 ms, radio frequency lens voltage (RFLens) at 45, and collision energy (CE) at 28%. Data acquisition was performed using a precursor list targeting both the endogenous and synthetic heavy labelled peptides. The Xcalibur 4.3 acquisition software was used.

ZenoTOF 8600: The samples were analysed using the ZenoTOF 8600 system (Sciex) equipped with an OptiFlow Pro ion source with nano interface (Sciex) and an ACQUITY UPLC M-Class system (Waters) in nanoflow mode. The system was operated with 99.9% water, 0.1% formic acid (v/v, buffer A), and 99.9% acetonitrile, 0.1% formic acid (v/v, buffer B). Peptides were loaded for 15 min at 2% B (loop load) and separated on a 15 cm × 75 µm inner diameter, 1.7 µm C18 Aurora Elite column (IonOptiks) using a stepwise gradient with 15–105 min from 2% to 35% B, 105–108 min from 35% to 45% B, 108–109 min from 45% to 80% B, 109–114 min held at 80% B, 114–115 min from 80% to 2% B, held to 135 min at 2% B at 250 nl min^{-1}. The source parameters were as follows: nano interface temperature, 240 °C; nano nebulizer gas, 15 psi; curtain gas, 35 psi; nano spray voltage, 2,300 V. Data were acquired with a TOF-MS scan in the *m/z* range of 200–2,000 Da with 200 ms accumulation time and Q Jet DP of 70, followed by MRMHR (PRM) with a TOF-MS/MS scan in the *m/z* range of 100–1,750 Da, Q1 resolution of Unit (0.7 ± 0.1 Da), Zeno pulsing on (on-demand), Zeno threshold of 1×10^6 cps using a MRMHR precursor list targeting the endogenous and synthetic heavy labelled peptides. Acquisition software SCIEX OS 4.0 was used.

For analysis and positive detection of the endogenous ncORF peptide above signal to noise Skyline (64-bit, v.25.1.0.142)[90] was used.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All MS data in this manuscript are publicly available through the PeptideAtlas database (https://peptideatlas.org/) and ProteomeXchange (https://proteomecentral.proteomexchange.org/). The Human HLA PeptideAtlas 2023-11 is freely accessible online (https://peptideatlas.org/builds/human/hla/index_2023-11.php) and the Human non-HLA PeptideAtlas 2023-06 is freely accessible at https://peptideatlas.org/builds/human/non-hla/. Specific dataset identifiers are listed in Supplementary Table 1. All ribosome profiling data manually inspected in this manuscript are publicly viewable at GWIPS-viz, as described in the Methods. MS PRM data are deposited to the ProteomeXchange Consortium via the PRIDE partner repository under dataset identifier PXD066599. Ribosome profiling, RNA-seq, CRISPR barcode sequencing data for eight cell lines screened in this Article as well as OLMALINC knockdown bulk RNA-seq and multiplexed scRNA-seq data are all submitted to the NCBI Short Read Archive under access code

PRJNA1294394. Primary gene and ncORF annotations were sourced from GENCODE (https://www.gencodegenes.org/), Ensembl Release 87 (https://www.ensembl.org/), UniProtKB/Swiss-Prot 2023 (https://www.uniprot.org/uniprotkb/) and the NCBI RefSeq database (https://www.ncbi.nlm.nih.gov/refseq/). Tissue-specific RNA-seq expression data were obtained from the Genotype-Tissue Expression (GTEx) portal (https://gtexportal.org/home/). Cancer dependency and CRISPR screening data were sourced from the DepMap portal (https://depmap.org/portal/). Ribosome profiling (Ribo-seq) data visualization and manual inspections were conducted using GWIPS-viz (https://riboseq.org/about.html). The Cancer Cell Line Encyclopedia (CCLE) was used for SNP extraction (https://sites.broadinstitute.org/ccle/datasets). The GSEA MSigDB was used to extract hallmark gene sets (https://www.gsea-msigdb.org/gsea/msigdb).

Code availability

Code generated for this Article has been deposited at GitHub (https://github.com/VanHeeschLab/deutsch_kok_et_al_2024) and Zenodo[91]. The code for the multilayer perceptron classifier model can be accessed at GitHub (https://git.embl.de/ivfimo/machine_learning_scripts) and Zenodo[92]. The code for ORBL is posted at GitHub (https://github.com/iljungr/ORBL_tools) and Zenodo[93]. The code for local enrichment scores in tiling screens has been deposited at GitHub (https://github.com/CFVALLS/tiling_screens_with_permutation) and Zenodo[94].

50. Vizcaíno, J. A. et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **32**, 223–226 (2014).
51. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nat. Methods* **14**, 513–520 (2017).
52. Deutsch, E. W. et al. Tiered human integrated sequence search databases for shotgun proteomics. *J. Proteome Res.* **15**, 4091–4100 (2016).
53. Keller, A., Eng, J., Zhang, N., Li, X. & Aebersold, R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.* **1**, MSB4100024 (2005).
54. Deutsch, E. W. et al. Trans-Proteomic Pipeline: robust mass spectrometry-based proteomics data analysis suite. *J. Proteome Res.* **22**, 615–624 (2023).
55. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383–5392 (2002).
56. Shteynberg, D. et al. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteom.* **10**, M111.007690 (2011).
57. Shteynberg, D. D. et al. PTMProphet: fast and accurate mass modification localization for the trans-proteomic pipeline. *J. Proteome Res.* **18**, 4262–4272 (2019).
58. Mendoza, L. et al. Flexible and fast mapping of peptides to a proteome with ProteoMapper. *J. Proteome Res.* **17**, 4337–4344 (2018).
59. Yates, A. et al. Ensembl 2016. *Nucleic Acids Res.* **44**, D710–D716 (2016).
60. Deutsch, E. W. et al. State of the Human Proteome in 2014/2015 As Viewed through PeptideAtlas: enhancing accuracy and coverage through the AtlasProphet. *J. Proteome Res.* **14**, 3461–3473 (2015).
61. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).
62. Feng, X. et al. Using the entrapment sequence method as a standard to evaluate key steps of proteomics data analysis process. *BMC Genom.* **18**, 143 (2017).
63. Frankenfield, A. M., Ni, J., Ahmed, M. & Hao, L. Protein contaminants matter: building universal protein contaminant libraries for DDA and DIA proteomics. *J. Proteome Res.* **21**, 2104–2113 (2022).
64. Zahn-Zabal, M. et al. The neXtProt knowledgebase in 2020: data, tools and usability improvements. *Nucleic Acids Res.* **48**, D328–D334 (2019).
65. van Wijk, K. J. et al. The *Arabidopsis* PeptideAtlas: harnessing worldwide proteomics data to create a comprehensive community proteomics resource. *Plant Cell* **33**, 3421–3453 (2021).
66. Omenn, G. S. et al. Progress identifying and analyzing the human proteome: 2021 metrics from the HUPO Human Proteome Project. *J. Proteome Res.* **20**, 5227–5240 (2021).
67. Omenn, G. S. et al. The 2022 Report on the Human Proteome from the HUPO Human Proteome Project. *J. Proteome Res.* **22**, 1024–1042 (2023).
68. Omenn, G. S. et al. The 2023 Report on the Proteome from the HUPO Human Proteome Project. *J. Proteome Res.* **23**, 532–549 (2024).
69. UniProt Consortium UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2022).
70. Sondka, Z. et al. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).
71. Ramsbottom, K. A. et al. Method for independent estimation of the false localization rate for phosphoproteomics. *J. Proteome Res.* **21**, 1603–1615 (2022).
72. Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* **48**, gkaa379 (2020).
73. Lybaert, L. et al. Challenges in neoantigen-directed therapeutics. *Cancer Cell* **41**, 15–40 (2023).
74. Lonsdale, J. et al. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
75. Michel, A. M., Kiniry, S. J., O'Connor, P. B. F., Mullan, J. P. & Baranov, P. V. GWIPS-viz: 2018 update. *Nucleic Acids Res.* **46**, gkx790 (2017).
76. Gaertner, B. et al. A human ESC-based screen identifies a role for the translated lncRNA LINC00261 in pancreatic endocrine differentiation. *eLife* **9**, e58659 (2020).
77. Kiniry, S. J., Judge, C. E., Michel, A. M. & Baranov, P. V. Trips-Viz: an environment for the analysis of public and user-generated ribosome profiling data. *Nucleic Acids Res.* **49**, W662–W670 (2021).
78. Tierney, J. A. S. et al. RiboSeq.Org: an integrated suite of resources for ribosome profiling data analysis and visualization. *Nucleic Acids Res.* **53**, D268–D274 (2025).
79. Ruiz-Orera, J. et al. Evolution of translational control and the emergence of genes and open reading frames in human and non-human primate hearts. *Nat. Cardiovasc. Res.* **3**, 1217–1235 (2024).
80. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2012).
81. Calviello, L., Sydow, D., Harnett, D. & Ohler, U. Ribo-seQC: comprehensive analysis of cytoplasmic and organellar ribosome profiling data. Preprint at *bioRxiv* https://doi.org/10.1101/601468 (2019).
82. Perez, G. et al. The UCSC Genome Browser database: 2025 update. *Nucleic Acids Res.* **53**, D1243–D1249 (2024).
83. Hecker, N. & Hiller, M. A genome alignment of 120 mammals highlights ultraconserved element variability and placenta-associated enhancers. *GigaScience* **9**, giz159 (2020).
84. Dempster, J. M. et al. Chronos: a cell population dynamics model of CRISPR experiments that improves inference of gene fitness effects. *Genome Biol.* **22**, 343 (2021).
85. Wessels, H.-H. et al. Massively parallel Cas13 screens reveal principles for guide RNA design. *Nat. Biotechnol.* **38**, 722–727 (2020).
86. Neavin, D. et al. Demuxafy: improvement in droplet assignment by integrating multiple single-cell demultiplexing and doublet detection methods. *Genome Biol.* **25**, 94 (2024).
87. Peidli, S. et al. scPerturb: harmonized single-cell perturbation data. *Nat. Methods* **21**, 531–540 (2024).
88. Kaulich, P. T., Jeong, K., Kohlbacher, O. & Tholey, A. Influence of different sample preparation approaches on proteoform identification by top-down proteomics. *Nat. Methods* **21**, 2397–2407 (2024).
89. Cassidy, L., Kaulich, P. T. & Tholey, A. Depletion of high-molecular-mass proteins for the identification of small proteins and short open reading frame encoded peptides in cellular proteomes. *J. Proteome Res.* **18**, 1725–1734 (2019).
90. MacLean, B. et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).
91. Kok, L. W. LeronKok/deutsch_kok_et_al_2024: analysis code v1.0.0. *Zenodo* https://doi.org/10.5281/zenodo.18878129 (2026).
92. Fierro-Monti, I. Binary classification models to predict mass spec (MS) detection of noncanonical(nc)ORF microproteins. *Zenodo* https://doi.org/10.5281/zenodo.18787106 (2026).
93. Jungreis, I. ORBL_tools: tools for measuring evolutionary conservation and constraint of "ORFness" of an open reading frame (v0.9-beta). *Zenodo* https://doi.org/10.5281/zenodo.18749292 (2026).
94. Valls, C. tiling_screens_with_permutation (v1.0.1). *Zenodo* https://doi.org/10.5281/zenodo.18865015 (2026).

Acknowledgements We acknowledge the authors of many excellent research publications in this field. This work was funded in part by the National Institutes of Health grants R01 GM087221 (E.W.D., R.L.M.), R24 GM148372 (E.W.D.), S10 OD026936 (R.L.M.) and by the National Science Foundation grants DBI-1933311 (E.W.D.) and MRI-1920268 (R.L.M.). J.R.P. acknowledges funding from the National Institutes of Health/National Cancer Institute (K08-CA263552-01A1); the V Foundation for Cancer Research (V2024-013); Tough2gether Foundation; Hyundai Hope on Wheels Foundation; the Yuvaan Tiwari Foundation; DIPG/DMG Research Funding Alliance; Book for Hope Foundation; CureSearch for Cancer Research; Morgan Adams Foundation; and the Andrew McDonough B+ Foundation (1548557); the Lindonlight Collective (GR-24-008 and GR-24-012); the Lung Cancer Research Foundation (1491799); the American Brain Tumor Association (DG2500077); the Cannonball Kids Cancer Foundation (45); the Chad Carr Pediatric Brain Tumor Center; and the University of Michigan Rogel Cancer Center. J.R.P. is the Ben and Catherine Ivy Foundation Clinical Investigator of the Damon Runyon Cancer Research Foundation (CI-127-24). S.v.H. acknowledges funding from Fonds Cancers (FOCA, Belgium), Stichting Reggeborgh and Villa Joep. This project was partially supported by the Fight Kids Cancer Funding Programme, supported by Imagine for Margo, Fondation KickCancer, Fondatioun Kriibskrank Kanner, Federazione Italiana Associazioni Genitori e Guariti Oncoematologia Pediatrica, Cris Cancer Foundation and Stichting Kinderen Kankervrij (KiKa). This publication is part of the project "Evolutionarily young microproteins in childhood brain cancer" with project number VI.Vidi.223.022 of the research programme NWO talent programme Vidi, which is partly financed by the Dutch Research Council (NWO), awarded to S.v.H. Research reported in this publication was supported by Oncode Accelerator, a Dutch National Growth Fund project under grant number NGFOP2201, awarded to S.v.H. This work is co-financed by Oncode Institute, which is partly funded by the Dutch Cancer Society. J.A.V. acknowledges funding from the Wellcome Trust (223745/Z/21/Z), BBSRC (BB/Y513829/1, BB/S01781X/1) and EPSRC (EP/Y035984/1). M.M., M.J.M. and S.O. acknowledge funding from the National Human Genome Research Institute (NHGRI), Office of Director (OD/DPCPSI/ODSS), National Institute of Allergy and Infectious Diseases (NIAID), National Institute on Aging (NIA), National Institute of General Medical Sciences (NIGMS), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Eye Institute (NEI), National Cancer Institute (NCI), National Heart, Lung and Blood Institute (NHLBI) of the National Institutes of Health under Award Number U24HG007822. I.F.-M. acknowledges funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 945405. P.V.B. acknowledges funding from Taighde Éireann—Research Ireland

(20/FFP-A/8929). T.F.M. acknowledges financial support from NIH grant R35GM157126. J.G.A and S.A.C. are supported in part by grants P01CA206978 from the NIH, and grants U24CA270823 and U01CA271402 from the NCI Clinical Proteomic Tumor Analysis Consortium program, as well as a grant from the Dr. Miriam and Sheldon G. Adelson Medical Research Foundation to S.A.C. J.G.A. acknowledges funding from the Broad Institute's Merkin Institute Fellowship. N.H. was supported by ERC Advanced Grant (EU Horizon 2020, AdG788970), Deutsche Forschungsgemeinschaft (DFG; CRC 1470 and CRC 1700) and the EU Horizon 2020 Pathfinder Program. P.F. was supported by a Victorian Cancer Agency Mid-Career Fellowship and the National Health and Medical Research Council of Australia (NHMRC). J.M.M., A.F., J.R.P., and I.J. are supported by the National Human Genome Research Institute (NHGRI) of the US National Institutes of Health (NIH) under award number U24HG007234. J.M.M. and A.F. are supported by the Wellcome Trust (grant number 222155/Z/20/Z). J.M.M., A.F., J.A.S.T., J.A.V., M.M., M.J.M. and S.O are supported by EMBL core funding. E.A.B. is funded by the NHGRI grant U24HG00334. M.A.B. acknowledges funding from the NSERC (Discovery grant RGPIN-2023-05203 and RTI grant RTI-2024-00556). M.A.B. is supported by a FRQS Junior 2 award (367740) and a research chair from the Centre de Recherche Medicale de l'Universite de Sherbrooke (CRMUS). F.-A.T. and F.B. are supported by FRQS scholarships (354090 and 352705, respectively). V.V. is supported by a FRQNT scholarship (354964). J.S.C. and S.B. acknowledge funding from the Wellcome Trust (223745/Z/21/Z) and Biomedical Research Centre. M.I.Ś. was supported by Foundation for Polish Science START scholarship, CRIDO Roots of the Future scholarship and the Poland National Science Centre (UMO-2021/41/B/NZ2/03036). We thank J. Causon, K. Tran and C. Feasley for providing access and for data collection on the Sciex ZenoTOF 8600 MS system. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Ensembl is a registered trademark of EMBL.

Author contributions Conceptualization: E.W.D., L.W.K., J.M.M., R.L.M., J.R.P. and S.v.H. Methodology: E.W.D., L.W.K., J.M.M., R.L.M., J.R.P., S.v.H., I.F.-M., J.R.-O., N.T., M.B.-S., J.S.C., J.C., J.A.V., C.F.V., S.W., U.K., I.J., M.K., J.A.R., M.G.R., L.W., K.B., F.B., S.B., A.G., D.E.R., T.G., V.V., F.-A.T., A.P., P.V.B., J.A.S.T., M.I.Ś. and M.A.B. Formal analysis: E.W.D., L.W.K., J.M.M., J.R.-O., I.F.-M., Z.S., S.C., I.J., C.F.V. and M.A.B. Investigation: E.W.D., L.W.K., J.M.M., J.R.-O., I.F.-M., Z.S. and I.J. Resources: R.L.M., J.R.P., S.v.H. and M.A.B. Data curation: E.W.D., L.W.K., J.M.M., P.V.B., J.A.S.T., M.I.Ś., R.L.M., J.R.P. and S.v.H. Writing—original draft: E.W.D., L.W.K., J.M.M., R.L.M., J.R.P., S.v.H., J.R.-O., I.F.-M., J.S.C., M.B.-S., J.A.V. and N.T. Writing—review and editing: E.W.D., L.W.K., J.M.M., R.L.M., J.R.P., S.v.H., J.G.A., M.M.A., J.L.A., M.A.B., S.C., A.A.B., E.A.B., L.C., S.A.C., J.C., A.-R.C., K.D., P.F., N.H., N.T.I., M.M., M.J.M., T.F.M., G.M., U.O., S.O., O.J.L.R., X.R., S.A.S., E.V., A.W., J.S.W., W.W., Z.X., J.R.-O., I.F.-M., Z.S., J.S.C., M.B.-S., J.A.V., N.T., C.F.V., S.W., U.K., I.J., M.K., J.A.R., F.J.S., M.G.R., L.W., K.B., F.B., S.B., A.G., D.E.R., T.G., V.V., F.-A.T., A.F., A.P. and P.V.B. Visualization: E.W.D., L.W.K., C.F.V., I.J., U.K., P.V.B., J.A.S.T. and M.I.Ś. Supervision: R.L.M., J.R.P. and S.v.H. Project administration: R.L.M., J.R.P. and S.v.H. Funding acquisition: R.L.M., J.R.P. and S.v.H.

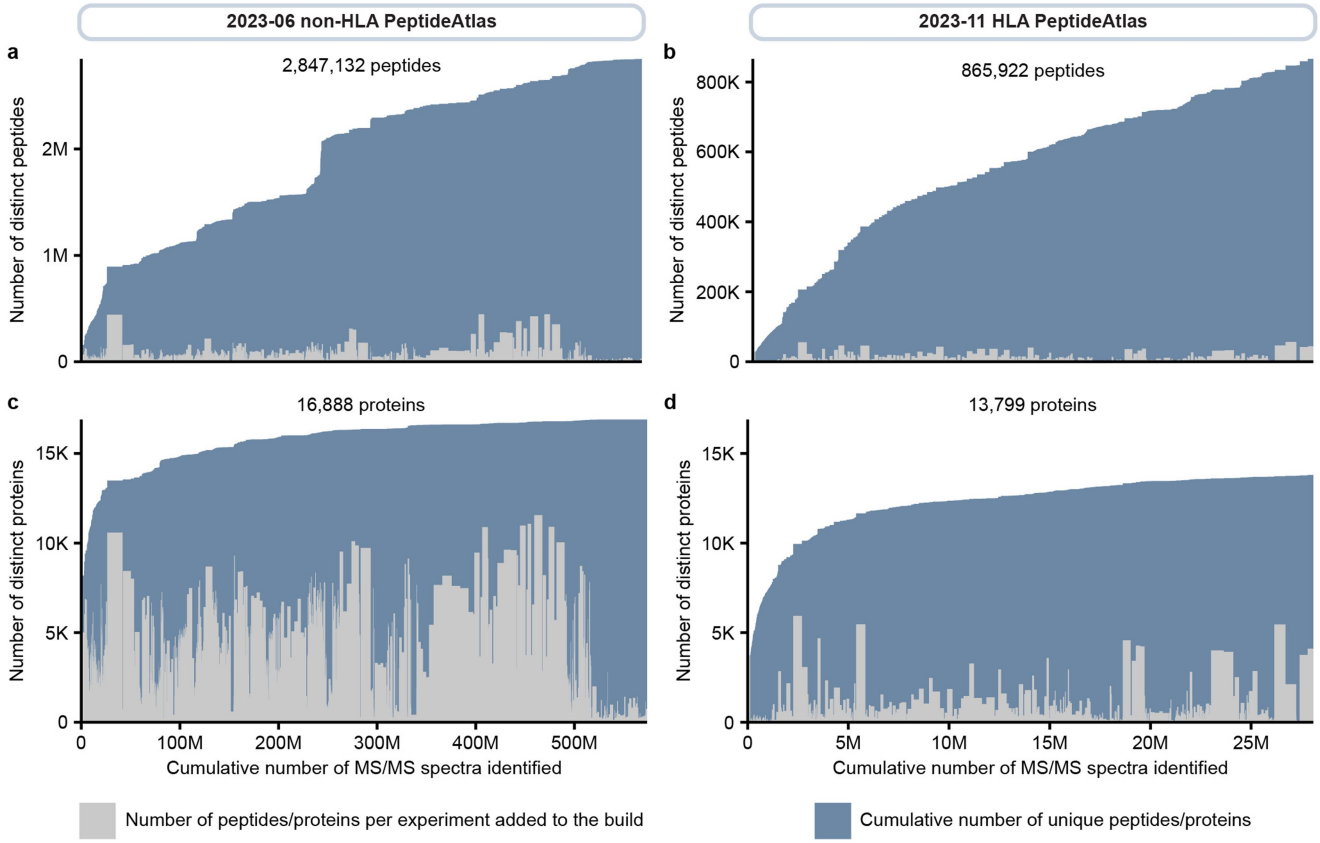
Competing interests J.R.P. has received research honoraria from Novartis Biosciences and Quantum-Si, and is on the scientific advisory board for, and receives research funding from, ProFound Therapeutics. J.G.A. is a paid consultant for Enara Bio and Moderna. J.L.A. is an advisor to Microneedle Solutions. G.M. is co-founder and CSO of OHMX.bio. S.A.C. is a member of the scientific advisory boards of Kymera, PTM BioLabs, MOBILion Systems and PrognomIQ. N.T.I. holds equity and serves as a scientific advisor to Tevard Biosciences. P.F. is a member of the scientific advisory board of Infinitopes. A.-R.C. is a member of the advisory board of ProFound Therapeutics. P.V.B. is a cofounder and shareholder of Eirnabio. D.E.R. receives research funding from members of the Functional Genomics Consortium (Abbvie, BMS, Jannsen, Merck) and is a director of Addgene. J.S.W. declares the following outside interests, which are unrelated to this work: 5 AM Venture, Amgen, nChroma Bio, KSQ Therapeutics, Maze Therapeutics, Tenaya Therapeutics, Tessera Therapeutics, Thermo Fisher Scientific, Third Rock Ventures and Xaira. The other authors declare no competing interests.

Additional information

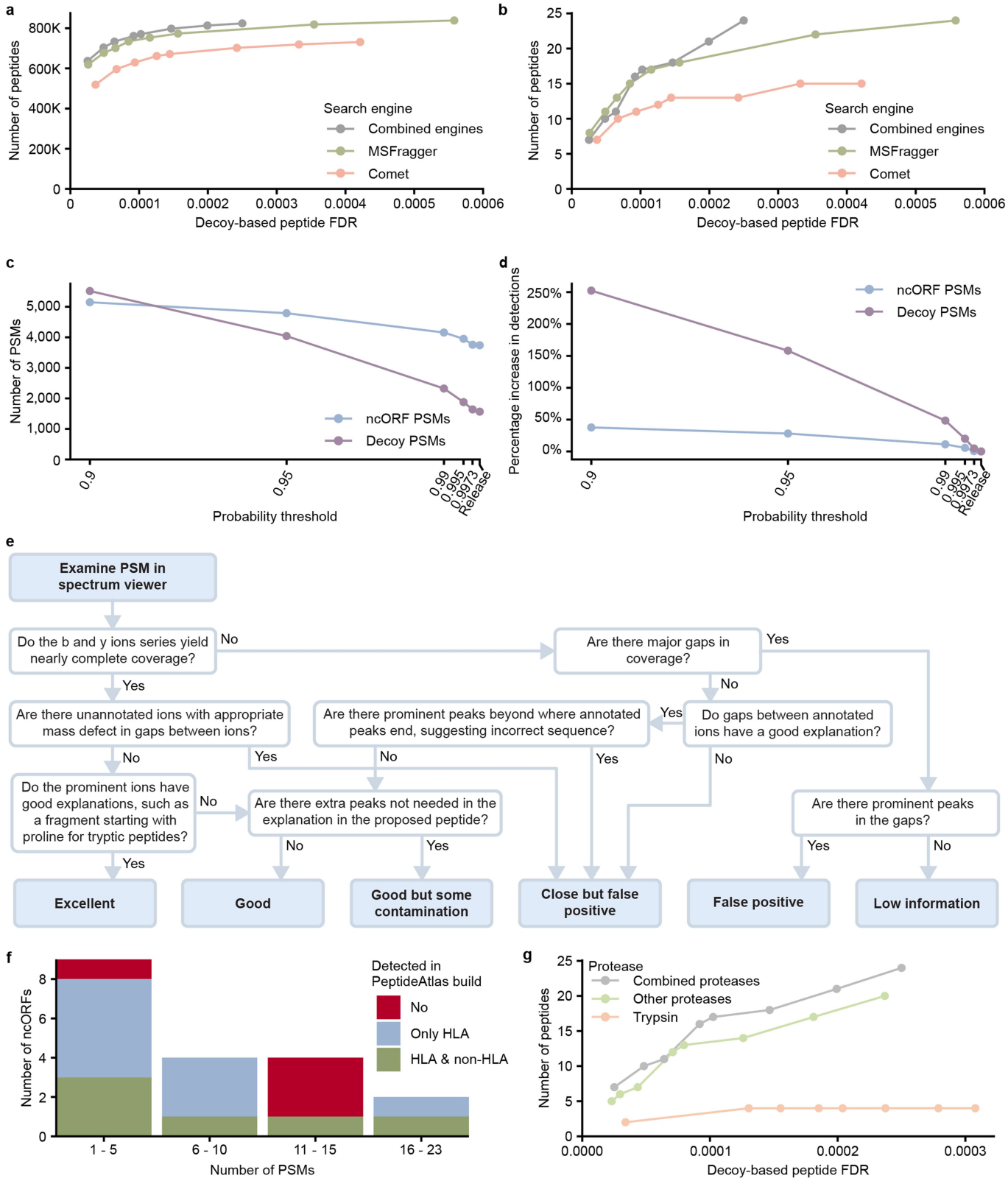
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-026-10459-x.

Correspondence and requests for materials should be addressed to Robert L. Moritz, John R. Prensner or Sebastiaan van Heesch.

Peer review information *Nature* thanks Christoph Dieterich, Akiyasu Yoshizawa and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

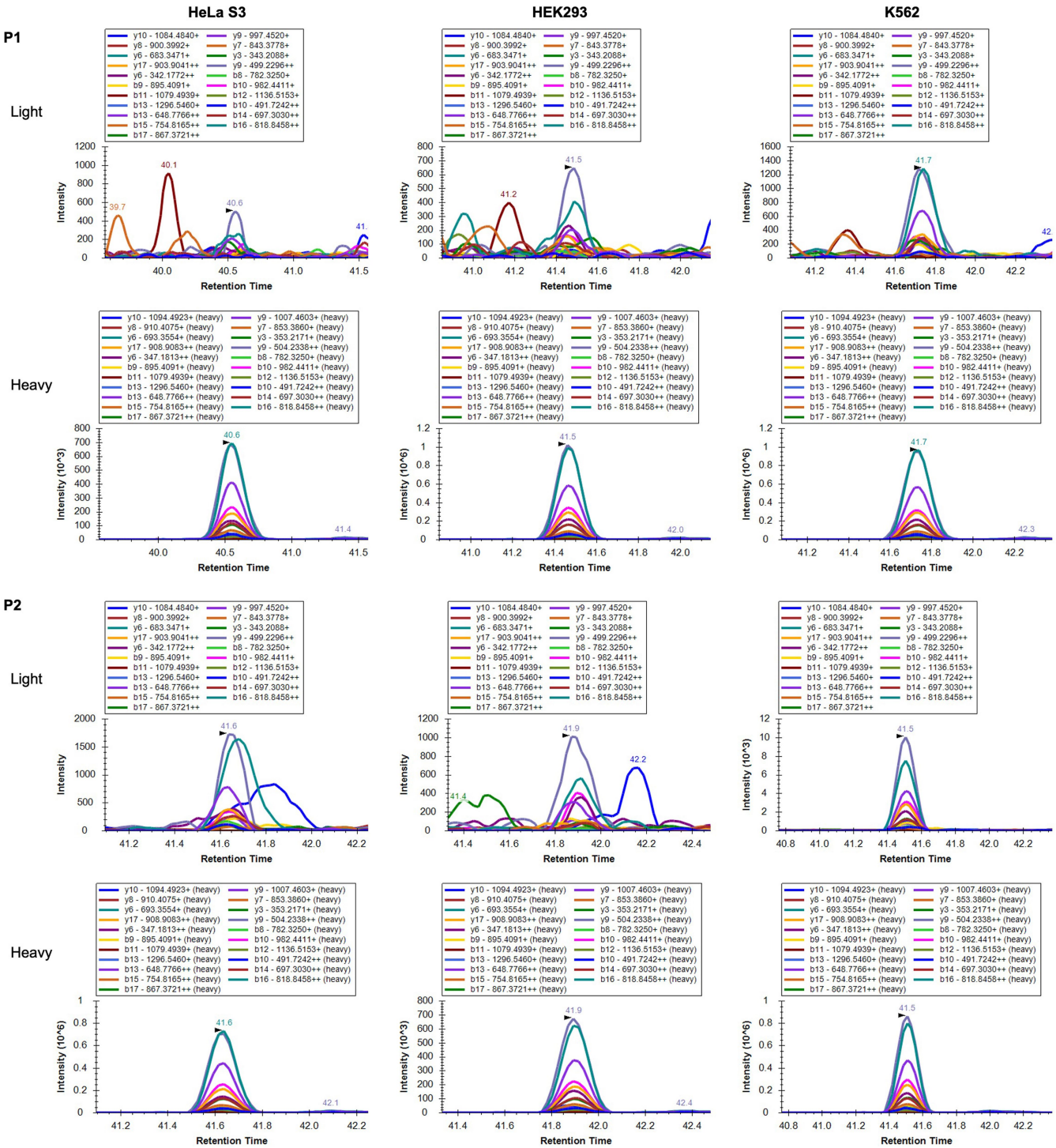


Extended Data Fig. 1 | The number of distinct peptides and proteins as datasets were added to the Human non-HLA (left) and HLA (right) PeptideAtlas. (**a**) Over 2.8 million distinct peptides have been observed in the 573 million PSMs in the non-HLA build. Each rectangle is one of the 1,172 experiments. Blue rectangles represent the cumulative number of distinct peptides in the build, while the grey rectangles depict the total number of distinct peptides within each experiment. (**b**) Over 0.86 million distinct peptides have been observed in the 28 million PSMs in the HLA build. Each rectangle is one of the 592 experiments. Blue rectangles represent the cumulative number of distinct peptides in the build, while the grey rectangles depict the total number of distinct peptides within each experiment. (**c**) The blue rectangles depict the cumulative 16,888 canonical proteins that have been catalogued in the 2023-06 Human non-HLA PeptideAtlas, whereas the grey rectangles show the total number of proteins present in each of the 1,172 experiments. (**d**) The blue rectangles depict the cumulative 13,799 canonical proteins that have been catalogued in the 2023-11 Human HLA PeptideAtlas, whereas the grey rectangles show the total number of proteins present in each of the 592 experiments. Although the total number of peptides continues to increase steadily, progress in the number of proteins is now very slow. Over the last 100 million PSMs, the cumulative counts are increasing by ~2,000 peptides per million PSMs and ~1 newly identified protein per million PSMs.

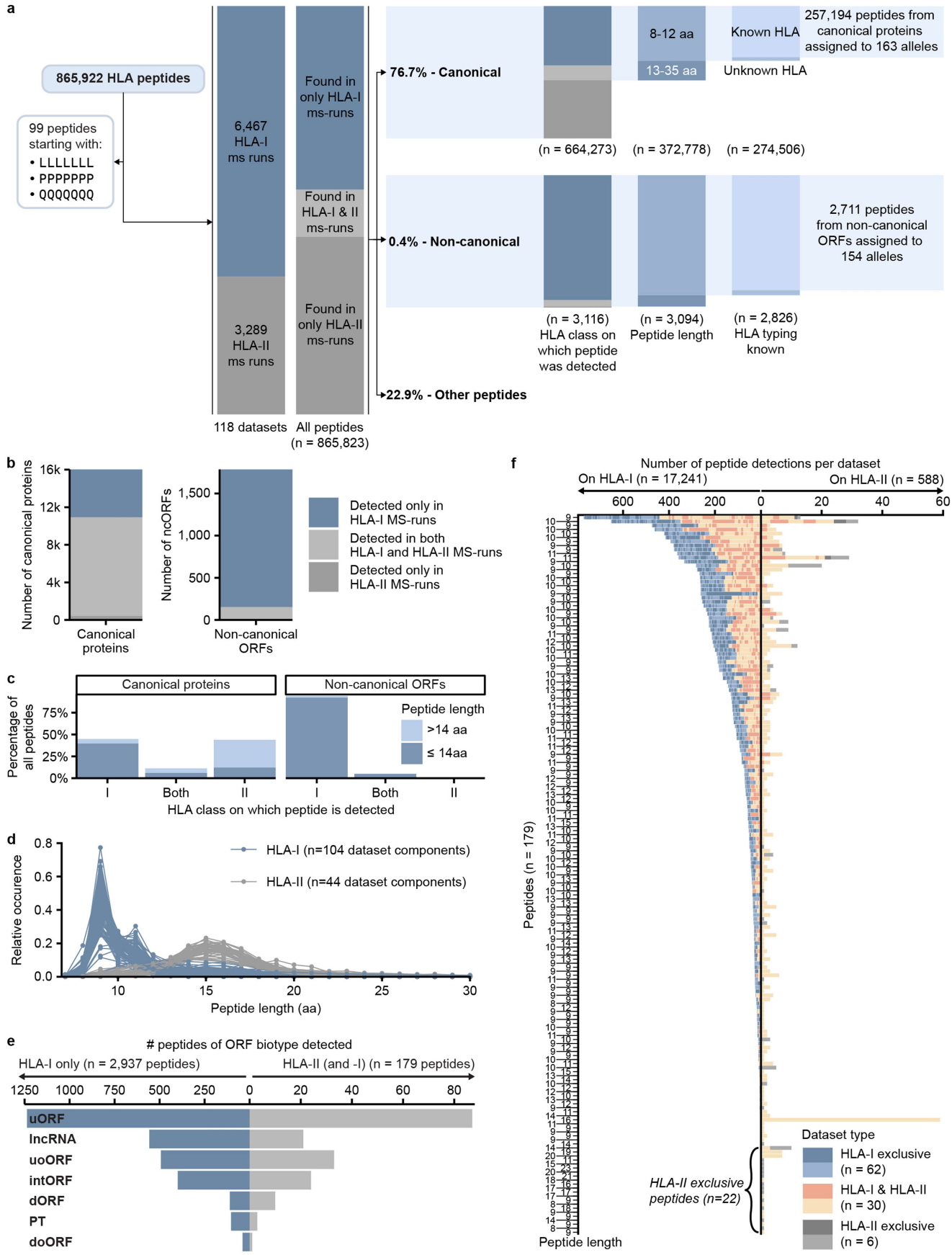


Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Additional validation information. (**a**) Number of total identified peptides in the analysis of the evaluation dataset PXD010154 at selected decoy-based FDR thresholds for Comet+TPP, MSFragger+TPP, and iProphet-combined search engine results. Combining search engines provides only a marginal improvement. (**b**) Number of identified ncORF peptides in the analysis of the evaluation dataset PXD010154 at selected decoy-based FDR thresholds for Comet+TPP, MSFragger+TPP, and iProphet-combined search engine results. Combining search engines provides only a marginal improvement. (**c**) Number of ncORF PSMs and decoy PSMs at selected posterior error probability thresholds lower than the PeptideAtlas build release threshold. The number of decoy PSM increases faster than ncORF PSMs with decreased threshold. (**d**) Percent increase in the number of ncORF PSMs and decoy PSMs at selected posterior error probability thresholds lower than the PeptideAtlas build release threshold. (**e**) Overview of the workflow used for manual validation of putative ncORF peptide spectrum matches. (**f**) Number of ncORF peptides identified upon analysis of 11 ubiquitinated datasets. ncORFs are binned by the number of PSMs by which they were detected. Colours indicate whether the ncORFs were already detected in the PeptideAtlas non-HLA or HLA build. (**g**) Number of identified ncORF peptides in evaluation dataset PXD010154 for tryptic digest runs as compared and combined with MS runs analysing digests with other proteases. The other proteases contribute a majority of ncORF PSMs in this dataset.

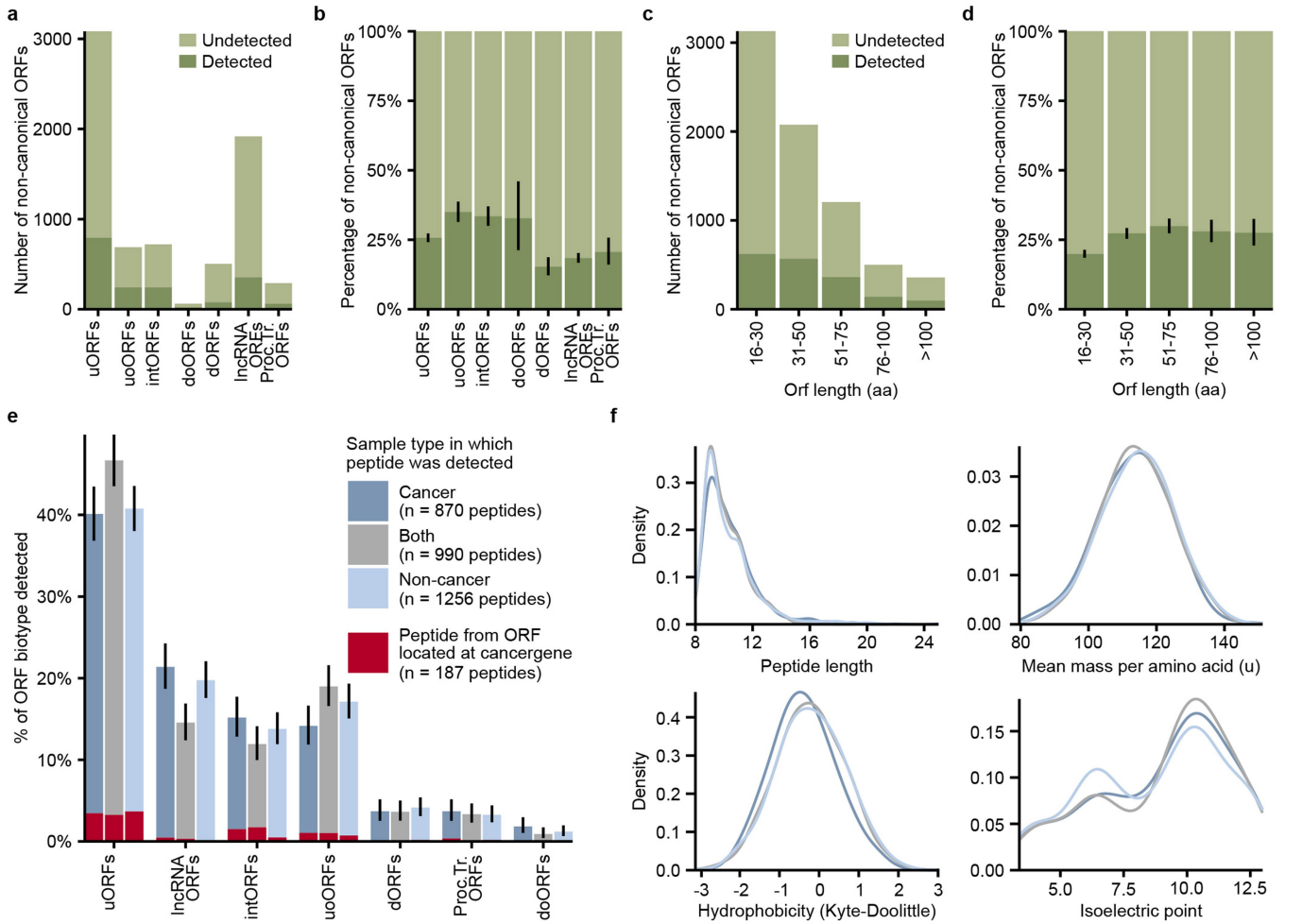


Extended Data Fig. 3 | Peptide verification of c11riboseqorf4 through targeted proteomics. Visualization of ion transitions from the endogenous (top and third row) and synthetic heavy labelled (second and bottom row) peptide ATPGHTGCLSPGCPDQPAR (c11riboseqorf4, Tier 1A) in cell lysates from HeLa S3, HEK 293 and K562 (left to right). The endogenous peptide was detected in all cell lysates from sample processing protocols P1 (top two rows) and P2 (bottom two rows). Data acquired on a Sciex ZenoTOF 8600 mass spectrometer. This peptide was previously seen in the PeptideAtlas non-HLA build.

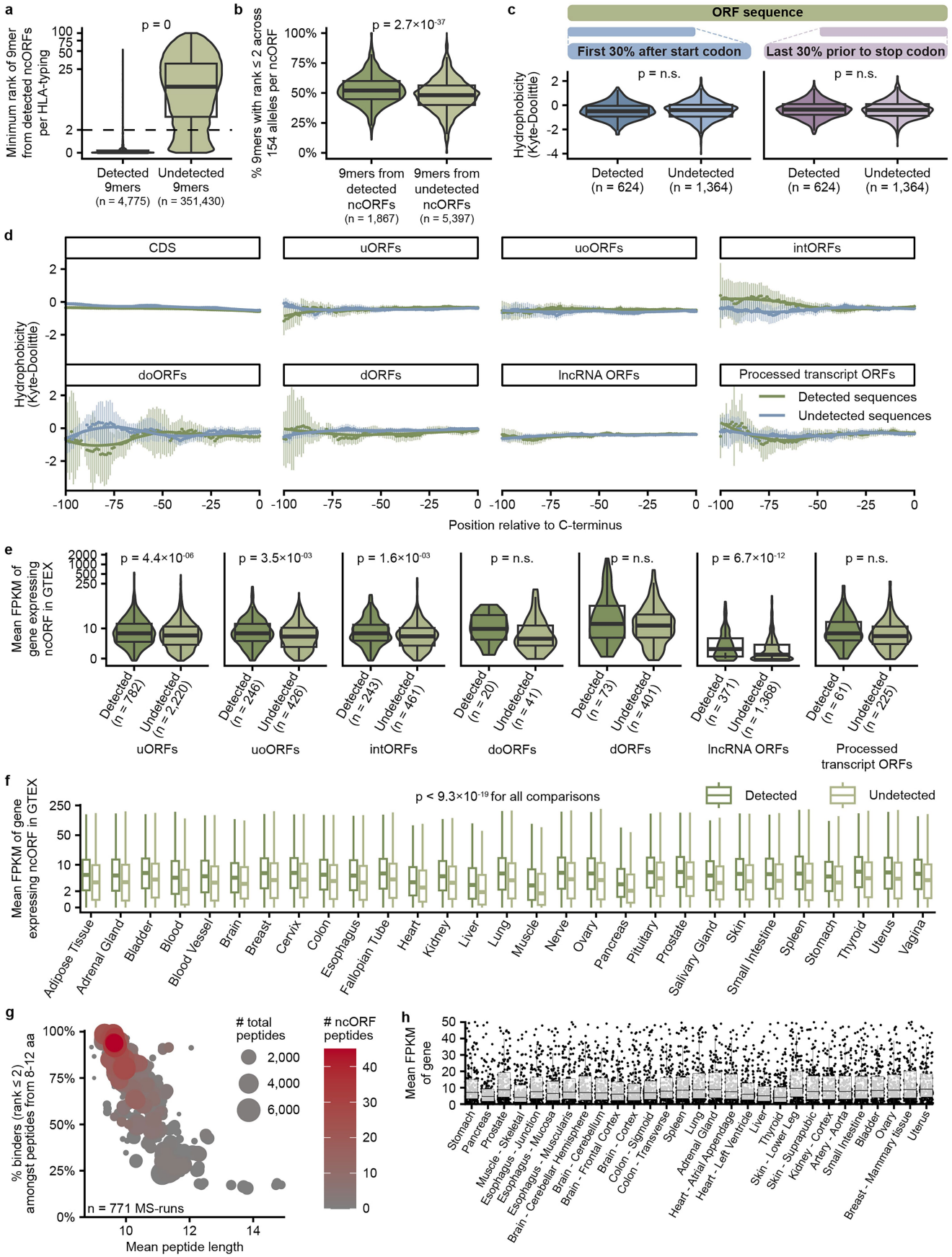


Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Detection of ncORF peptides in HLA-I and HLA-II, and in cancer and non-cancer samples. (**a**) Schematic illustrating the total numbers of peptides (from both normal proteins and ncORFs) extracted from the total set of peptides. Depending on the analysis, peptides were further selected for those that were detected on HLA-I, those that had a length from 8–12 amino acids, and those that originated from an MS-run with a known HLA-typing. The counts below each bar denote the number of distinct peptides. The distinction between "canonical", "non-canonical", and "other peptides" is defined in the methods. (**b**) Barplots showing for detected canonical proteins (left) and ncORFs (right) whether their detected peptides were exclusively detected in HLA-I or HLA-II MS-runs, or in both. (**c**) Barplots showing for canonical proteins and ncORFs the percentage of peptides found exclusively in HLA-I and HLA-II MS-runs, or in both. Bars are coloured by the peptide length being ≤14aa, or >14aa in length. (**d**) Line graph showing the peptide length distribution per dataset component split by HLA-class. (**e**) Bar plot showing the number of peptides detected per ORF exclusive to HLA-I samples (left) and those present in HLA-II samples, possibly in addition to HLA-I samples (right). Please note the x-axis scales differ by an order of magnitude between the left and right part of this panel. HLA-I and HLA-II peptide detection is not mutually exclusive as HLA-I peptides might be accidentally recovered from HLA-II pulldown experiments. (**f**) The frequency of peptide detection in HLA-I MS-runs (left) and HLA-II MS-runs (right) per peptide. Each alternating shade corresponds to a different dataset, with shades grouped by dataset type. The left axis denotes peptide lengths. Please note x-axis scales differ by an order of magnitude between the left and right part of this panel. Only peptides detected in at least one HLA-II sample are included. 22 of the 179 distinct peptides were exclusively detected in HLA-II samples. Fourteen of these peptides have a length 14 amino acids or greater, suggesting a potential presentation by HLA-II. This is still a minority in contrast to the total amount of 3,116 non-canonical ORF derived HLA peptides.

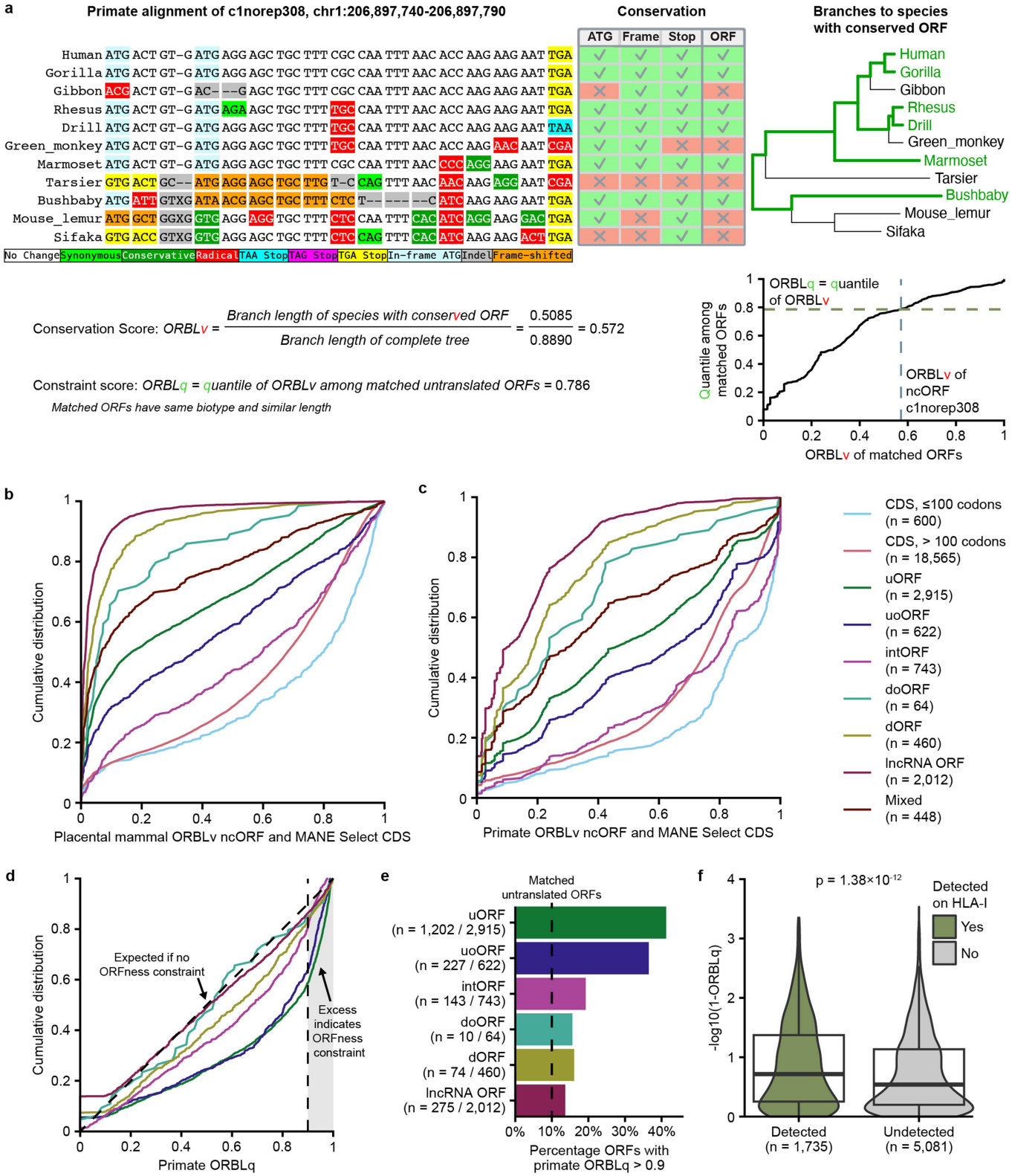


Extended Data Fig. 5 | Analysis of ncORFs detected by immunopeptidomics. (**a**–**d**) Comparisons of the detected and undetected ncORFs (n = 1,785 vs. 5,479). (**a**) The total number of ncORFs per ORF biotype and the number of ncORFs for which a peptide was observed. (**b**) As in (a), but now shown in percentages. (**c**) The total number of ncORFs grouped by length and the number of ncORFs for which a peptide was observed. (**d**) As in (c), but now shown in percentages. (**e**) Bar plot showing the proportion of ncORF-derived HLA peptides detected per biotype, categorized by whether the peptide was exclusively identified in immunopeptidomics analyses of cancer tissues or cell lines, non-cancer samples, or both. No significant changes in ORF biotype recovery are observed between these sample types. Peptides originating from ncORFs located on a known cancer gene are coloured red. (**f**) Density plots comparing the ncORF-derived HLA peptides differentiated on sample type (as depicted in (e)): cancer, non-cancer, or both. The plots compare peptides by their length, mass, hydrophobicity (Kyte-Doolittle scale), and isoelectric point. No significant changes between the distributions of these density plots can be observed. Black lines on bar graphs in (b, d, e) indicate 95% confidence intervals computed using a binomial test.



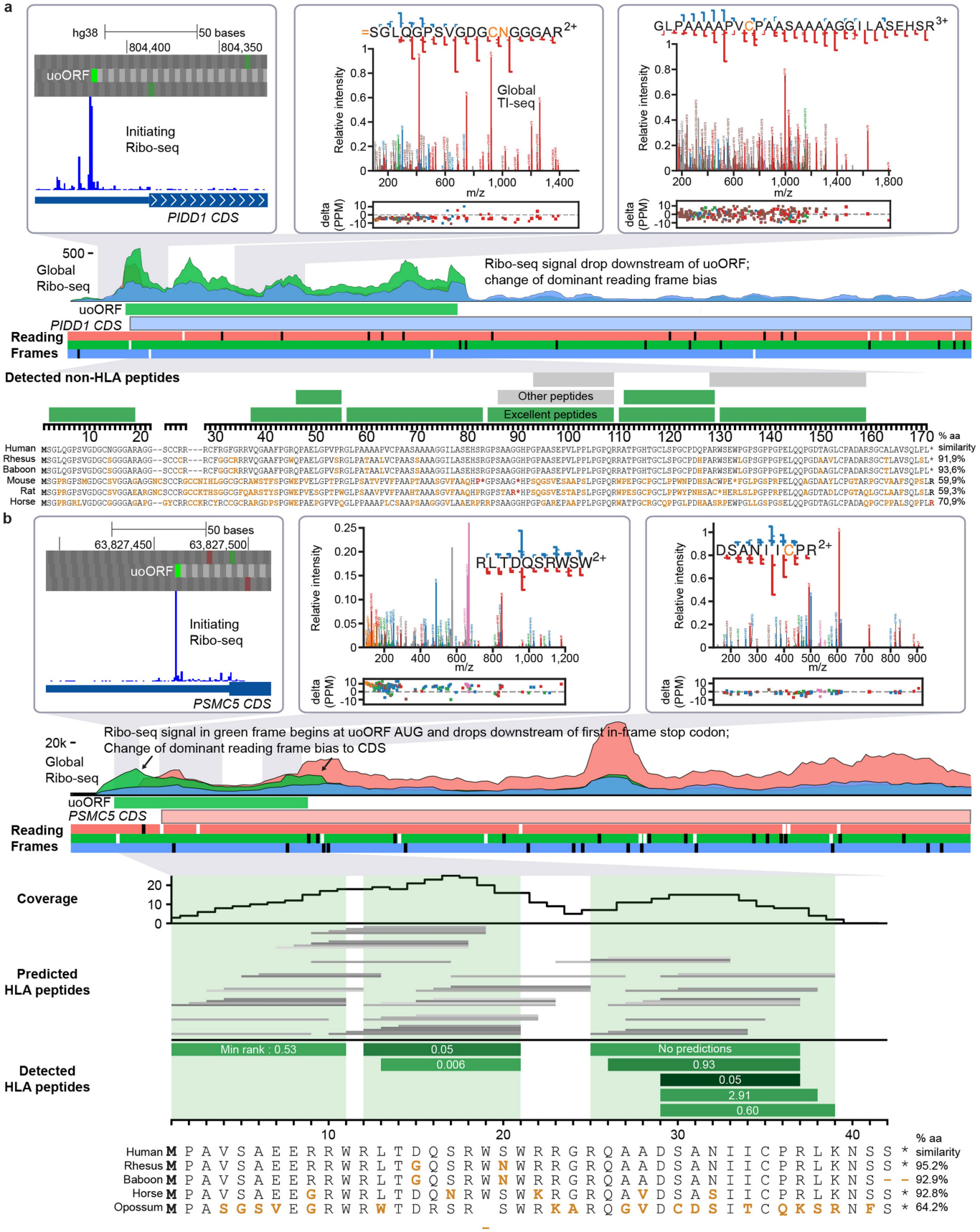
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Potential determinants of ncORF detection. (**a**) Violin plot comparing for all MS-runs grouped by HLA-typing the minimum binding prediction rank for detected and undetected (non-unique) 9mers. (**b**) Violin plot comparing all 9mers from detected and undetected non-canonical ORFs. The y-axis shows per ORF the percentage of 9mers with a NetMHCpan rank ≤ 2 across all 154 alleles associated with ncORF peptides. (**c**) Violin plots similar to (3a) comparing the hydrophobicity by the Kyte-Doolittle scale between detected and undetected ncORFs for the first 30% of the ncORF sequence after the start codon, or the last 30% of the ncORF sequence. Statistical tests for (a–c) were performed with a two-sided Wilcoxon rank-sum test, reported *P* values were adjusted for multiple testing with Holm-Bonferroni correction. (**d**) Comparison of the hydrophobicity similar to (3b) between detected and undetected ncORFs/CDS per ncORF biotype. Each dot represents the average hydrophobicity of the amino acids at that position and the 14 amino acids before that position per ncORF biotype or CDS grouped by whether these were detected or not in the immunopeptidomics data. The lines were fitted using Local Polynomial Regression Fitting. Vertical bars represent 95% confidence intervals. Note that because ncORFs are mostly smaller than 100 aa, confidence intervals get larger with increasing C-terminus offset. (**e**) Comparison of the expression levels of detected and undetected ncORFs similar to (3d), but split per biotype. On the *y*-axis, the mean FPKM in GTEX of genes expressing an ncORF is shown on a pseudo-log scale. 326 ncORFs for which the gene-id was not present in GTEX are not shown. Significance was determined using two-sided Wilcoxon rank-sum tests, reported *P* values were adjusted for multiple testing with Holm-Bonferroni correction. (**f**) Comparison of the expression levels of detected (n = 1,796) and undetected (n = 5,142) ncORFs similar to (e), but split per tissue. Outliers are not shown in the graph. Significance was determined using two-sided Wilcoxon rank-sum tests, and *P* values were adjusted for multiple testing with Holm-Bonferroni correction. All comparisons were found to be significant. (**g**) Dot plot similar to (2i), for MS-runs originating from the HLA-ligand-atlas. The plot visualizes the correlation between mean peptide length and the percentage of predicted binders amongst peptides with a length between 8 and 12 amino acids (NetMHCpan rank ≤ 2) per MS run. Dot size corresponds to the total number of peptides per MS-run. Dot colour corresponds with the percentage of non-canonical ORF-derived peptides per MS-run. (**h**) Comparison of the GTEx expression of 224/277 genes from which ncORFs in the HLA ligand atlas originate (53 genes with ncORFs in the HLA ligand atlas were not present in GTEx). GTEx tissues comparable to those from the HLA ligand atlas were selected, and sorted in the same way as in (3e). Each represents the mean FPKM of a gene across these tissue samples in GTEx. Only genes with a mean FPKM lower than 50 are plotted for clarity, but all 224 genes were included for the boxplots. Boxplots in (a-c, e, f, h) show the median and the 25th (Q1) and 75th (Q3) percentiles; whiskers extend to the most extreme values within 1.5x the interquartile range from Q1 and Q3.



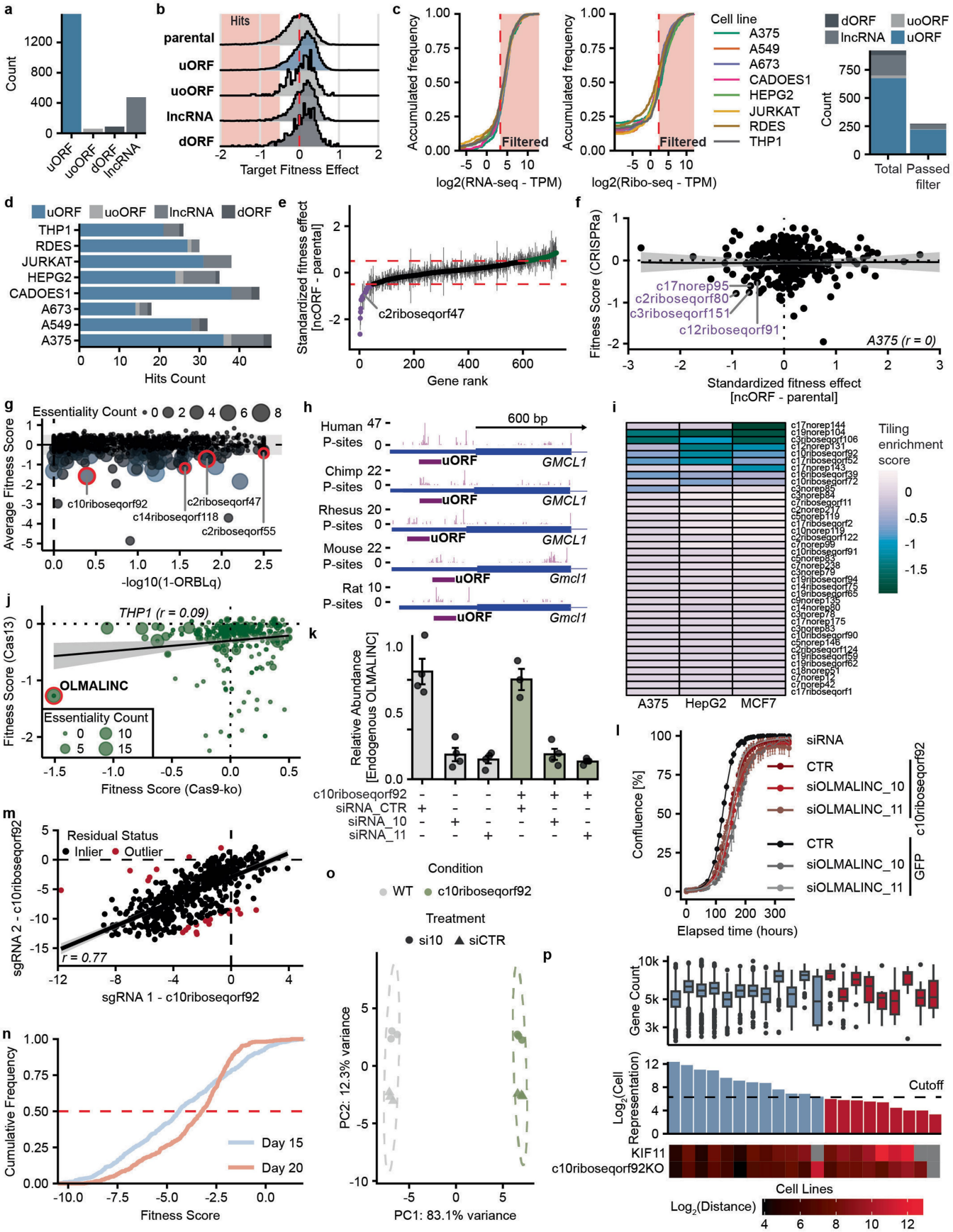
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | ORBL identifies a subset of ncORFs. (**a**) Illustration of how ORF Relative Branch Length (ORBL) scores are computed, using 17-codon uORF c1norep308 as an example. Local alignment of primate genomes to this human ncORF (upper left panel) colour-coded using CodAlignView, shows codons aligned to human start and stop codon, and frame-shifting insertions and deletions. Some species have been excluded for clarity. Table indicates in which species ATG, reading frame, and stop codon are conserved, with ORF considered conserved if all three are. Note that TAA stop codon in Drill is considered conserved even though it is not the same stop codon as in human. Note also that frame is considered conserved in Bushbaby because overall length is a multiple of 3 even though it includes a frame-shifting insertion and frame-shifting deletion. The Phylogenetic tree (upper right panel) shows branches leading to species having a conserved ORF (green). ORF Relative Branch Length conservation score (ORBLv) is calculated as the branch length of species having a conserved ORF divided by the branch length of all species in whole-genome alignment. To adjust for conservation due to chance or constraint on an overlapping CDS, ORF Relative Branch Length quantile score (ORBLq) measures constraint specifically on ORFness by computing the quantile of the ORBLv of the ncORF among the ORBLv's of matched untranslated ORFs of the same biotype and similar length (cumulative distribution plot in lower right panel). (**b**–**c**) Cumulative distribution of ORBLv scores quantifying conservation of ORFness for MANE Select CDS, grouped into two bins by length, and GENCODE ncORFs, grouped by biotype, for placental mammals (**b**) and primates (**c**). As expected, short ORFs tend to have more ORFness conservation than long ones, annotated CDS more than ncORFs, biotypes overlapping CDS more than other biotypes (presumably due to "free" conservation from the CDS), and in the primate clade more than in the placental mammal clade. (**d**) Cumulative distributions of ORBLq scores for ncORFs grouped by biotype similar to (Fig. 4b) but in primates. (**e**) Barplot showing the percentage of ncORFs of each biotype having an ORBLq score > 0.9, similar to (Fig. 4c) but in primates. (**f**) Violin plot comparing evolutionary constraint on ORFness of ncORFs detected and undetected by HLA-I peptides similar to (Fig. 4g), but for all biotypes combined. Detected ncORFs (n = 1,735) have significantly more constraint than undetected (n = 5,081) ($P = 1.38 \times 10^{-12}$, two-sided Wilcoxon rank-sum test). The 448 ncORFs for which ORBLq is not defined due to mixed biotype were not included. Boxplots show the median and the 25th (Q1) and 75th (Q3) percentiles; whiskers extend to the most extreme values within 1.5x the interquartile range from Q1 to Q3.



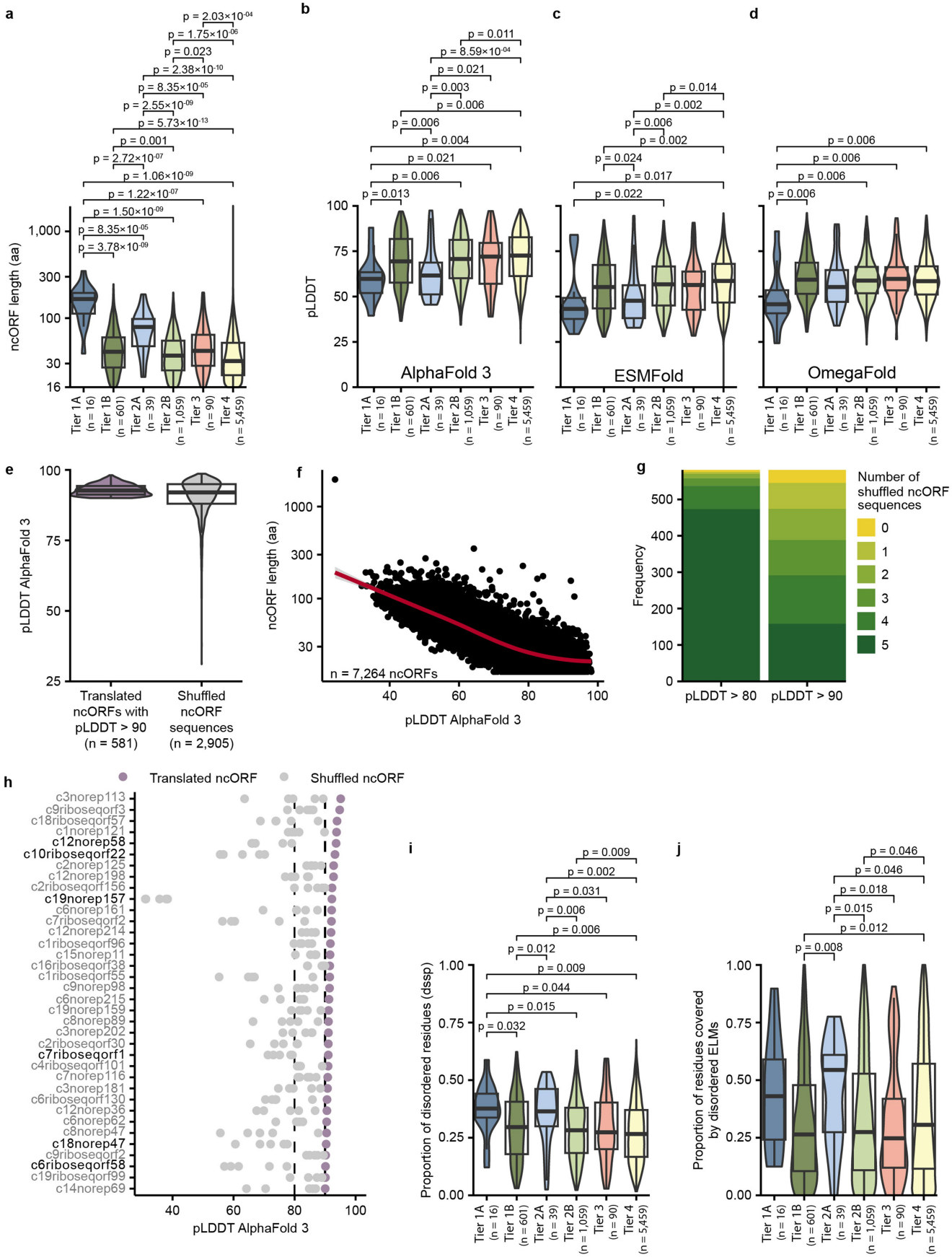
Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Examples of two ncORFs detected by either non-HLA or HLA data. (**a**) Ribo-seq, mass spectrometry, and evolutionary information for c11riboseqorf4, one of the best detected ncORFs in tryptic digests. This ncORF has 11 distinct peptides across 94 different experiments, 8 of which we classified as excellent evidence (green). The spectra for peptides SGLQGPSVG DGCNGGGAR and GLPAAAAPVCPAASAAAAGGILASEHSR are depicted with nearly complete y ion coverage and substantial b ion coverage, providing highly compelling evidence. We also note that SGLQGPSVGDGCNGGGAR begins as position 2 of the ORF and has peptide N-terminal acetylation, indicating ORF N-terminal acetylation after removal of the initiator methionine. (**b**) Overview of data available for c17norep146, an uoORF in the *PSMC5* gene. Ribo-seq data shows the initiation of translation at the methionine translation initiation codon (green). Two peptide spectral matches for HLA-I peptides RLTDQSRWSW and DSANIICPR are shown (USIs are mzspec: PXD004894:20141214_QEp7_MiBa_SA_HLA-I-p_MMf_4_2:scan:31976:RLTDQSR WSW/2, mzspec:PXD029567:UPN20_class_I_Rep3:scan:6685:DSANIIC[Cysteinyl]PR/2, respectively). The panel below the ribosome profiling data shows the position of all 8 peptides that were observed in the immunopeptidomics data. The colour 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. Detected peptides occurred in the regions with the highest numbers of predicted binders. While the initiation codon is found in most mammalian genomes (relative branch length of species with an aligned ATG is 0.819), overall ORF conservation is somewhat lower (ORBLv = 0.639, ORBLq = 0.937) due to frame-shifting insertions or deletions in some lineages.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Functional genomics for CRISPR screening and c10riboseqorf92. (**a**) Distribution of gRNA counts per biotype in 8 cell line screens. (**b**) Distribution of target fitness effects for each biotype, aggregated across all screened cell lines. Red shading marks the region containing hits (effect size below the selected threshold). The vertical dashed line at 0 indicates no effect. (**c**) Filtering strategy for ncORF hits based on expression evidence. Left: Cumulative $\log^2$(RNA-seq TPM) for targeted ncORFs across 8 cell lines, excluding parental genes. Middle: Cumulative $\log^2$(Ribo-seq TPM) for the same targets. Vertical dashed lines indicate minimum thresholds. Right: Total and filtered hit counts by biotype after applying both thresholds. (**d**) Bar plot showing the number of CRISPR hits per biotype in each cell line after filtering. (**e**) Ranked standardized fitness effects [ncORF − CDS]. Error bars show variability across 8 cell lines. Purple points indicate effects < −0.5 (stronger than CDS), green points > 0.5. (**f**) Scatter plot comparing standardized fitness effects in A375 cells with CRISPRa-derived fitness scores. Each dot represents an ncORF; select ncORFs of interest are labelled. (**g**) Relationship between the average fitness effect of ncORF loss-of-function and evolutionary constraint in placental mammals as measured by $-\log^{10}$ (1-ORBLq score). Each point represents a targeted ncORF; dot size corresponds to the number of cell lines in which the ncORF is classified as essential in the CRISPR screen. (**h**) Ribosome profiling (P-site density) for *GMCL1*/c2riboseqorf47 across five species, showing conserved translation signatures. (**i**) Heatmap of tiling enrichment scores for A375, HepG2, and MCF7 from Prensner et al.[28]. (**j**) Scatter plot comparing gene fitness effects derived from Cas9-based CRISPR knockout and Cas13-based RNA-targeting screens for lncRNAs. Each point represents a single lncRNA, with point size indicating the number of independent screens in which it was classified as essential. The ncORF-containing lncRNA c10riboseqorf92/*OLMALINC* is highlighted. (**k**) *OLMALINC* silencing validation using siRNA#10 and siRNA#11 in A375 cells expressing GFP or c10riboseqorf92. Relative abundance is measured by qPCR using *GAPDH* as a housekeeping gene ($n = 4$). (**l**) Proliferation curves for A375 cells expressing GFP or c10riboseqorf92 following transfection with siRNA#10/siRNA#11. Cell confluence was monitored over time and normalized to initial measurements. Data represent mean ± s.d. of technical replicates ($n = 4$). (**m**) Correlation of c10riboseqorf92 loss-of-function effects across 485 cell lines targeted with two independent sgRNAs. Fitness effects were measured at day 15 post-perturbation. A linear regression model ($r = 0.77$) was fitted, with the shaded area indicating the 95% confidence interval. Outlier points with large residuals are shown in red. (**n**) Cumulative frequency distributions of fitness effects from pooled CRISPR screens performed at days 15 and 20 post-perturbation. The horizontal dashed red line denotes the median of each distribution. (**o**) PCA of RNA-seq recovery assay. PCA was computed using the top 10% most variable genes across all samples. WT A375 cells are shown in red, and A375 cells with exogenous c10riboseqORF92 expression are shown in blue. Each point represents an individual sample. (**p**) Pooled single-cell RNAseq filtering. **Top:** Number of unique genes detected per cell in each cell line, shown on a log scale. **Middle:** $\log^2$-transformed representation (cell counts) of each cell line after preprocessing. The horizontal dashed red line indicates the cutoff applied for downstream analyses. **Bottom:** Heatmap showing the Euclidean distances between transcriptional profiles of c10riboseqORF92-perturbed cells and the positive control (*KIF11*) relative to unperturbed cells. Distances are $\log^2$-transformed.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Structure predictions of ncORFs separated by evidence-based Tier. (**a**) Box plots showing the distribution of amino acid sizes in ncORFs, categorized by Tier. (**b**–**d**) Box plots showing the distribution of average pLDDT scores per ncORF, categorized by Tier, for (b) AlphaFold3, (c) ESMFold, and (d) OmegaFold. Pairwise comparisons were performed using a two-sided Wilcoxon rank-sum test, with *P* values adjusted for multiple testing using false discovery rate (FDR). Only significant adjusted *P* values (<0.05) are displayed. (**e**) Box plot showing the distribution of average AlphaFold3 pLDDT scores for 2,905 shuffled amino acid sequences, generated from 581 original sequences with pLDDT > 90. Each original sequence was randomly shuffled five times. The dashed line marks the pLDDT threshold of 90. (**f**) Scatter plot showing the relationship between ncORF length and predicted pLDDT scores. A clear inverse linear trend is observed, with shorter ncORFs exhibiting higher predicted pLDDT values. Each point represents an individual ncORF; the regression line and shaded area indicate the best-fit linear model and its standard error, respectively. (**g**) Bar plots showing the number of shuffled ncORF sequences per original sequence that achieved predicted pLDDT scores >80 (left) or >90 (right). (**h**) Dot plot listing 36 ncORFs with AlphaFold3 pLDDT > 90 in which none of their five shuffled versions reached pLDDT > 90. In bold are six cases where none of the shuffled sequences exceeded pLDDT > 80. (**i**,**j**) Box plots showing the proportion of disordered residues (i) and the proportion of residues in disordered ELMs (j) per ncORF, categorized by Tier. Pairwise comparisons were performed using a two-sided Wilcoxon rank-sum test, with *P* values adjusted for multiple testing using false discovery rate (FDR). Only significant adjusted *P* values (<0.05) are displayed. Boxplots in (a–d, i, j) show the median and the 25th (Q1) and 75th (Q3) percentiles; whiskers extend to the most extreme values within 1.5x the interquartile range from Q1 and Q3.

nature portfolio

Corresponding author(s): Sebastiaan van Heesch
John Prensner
Robert Moritz

Last updated by author(s): 03/06/2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The PRM data collection on the Orbitrap Astral instrument was performed with Xcalibur 4.3. The PRM data collection on the ZenoTOF 8600 instrument was performed with SCIEX OS 4.0.
Data analysis	General statistical analysis, data processing, and visualization were performed using R (v4.4.0/4.4.2) and Python (v3.9+). Data manipulation and tidy data principles were implemented using the tidyverse (v2.0.0) suite, including dplyr (v1.1.4), tidyr (v1.3.1), readr (v2.1.5), stringr (v1.5.1), tibble (v3.2.1), purrr (v1.0.2), forcats (v1.0.0), lubridate (v1.9.3), stringi (v1.8.7), here (v1.0.2), and httr (v1.4.7). Data visualization was generated using ggplot2 (v3.5.1), ggpubr (v0.6.0), patchwork (v1.3.0), ComplexHeatmap (v2.20.0), circlize (v0.4.16), ggbeeswarm (v0.7.2), gghalves (v0.1.4), ggpattern (v1.1.1), scales (v1.3.9), RColorBrewer (v1.1-3), and rcartocolor (v2.1.1). Proteomics & Mass Spectrometry: MS data were processed using the Trans-Proteomic Pipeline (TPP v6.3.3/7.1), including the search engines MSFragger (v3.7) and Comet (2019.01.5), with probability modeling provided by PeptideProphet (v6.3.3/7.1). Targeted proteomics analysis, including parallel reaction monitoring (PRM) and data visualization, were performed with Skyline (v25.1.0.142), SCIEX OS (v4), and Xcalibur (v4.3). Bioinformatic protein property analysis utilized the Peptides (v2.4.6) R package. Genomics & Transcriptomics: Sequencing data were processed using the nf-core/rnaseq (v3.1.19) pipeline, utilizing STAR (v2.7.3a), Bowtie2 (v2.5.4), SAMtools (v1.2), and BedTools (v2.31.1) for alignment and genomic arithmetic. Differential expression was assessed with DESeq2 (v1.46.0) using apeglm (v1.28.0) and ashr (v2.2.6) for shrinkage. Ribo-seq data quality control was evaluated using RiboseQC (v0.99.0). Single-Cell Analysis: Single-cell processing and demultiplexing utilized Cell Ranger (v9.0.1), Demuxafy (v1.0.2), and scSplit (v1.0.0). Downstream analysis and integration were performed using Seurat (v5) and scanpy (v1.11.4). Perturbation analysis and E-distance calculations were conducted using pertpy (v1.0.3).

Functional Genomics & CRISPR: CRISPR screen data were normalized and analyzed using Chronos (v2.0.8), with guide RNAs designed via the CRISPick web portal. Real-time proliferation and live-cell imaging were analyzed using Incucyte Software (2023A Rev2 GUI) and growthcurver (v0.3.1). RT-PCR data were processed with QuantStudio™ Design & Analysis (v1.6.1). Functional enrichment was performed with clusterProfiler (v4.14.6).

Structural Biology & Modeling: Putative ncORF-derived protein structures and sequences were modeled and analyzed using AlphaFold3 (v3.0.1), ESM3, and OmegaFold. HLA-binding affinities were predicted using NetMHCpan (v4.1). Coding potential and evolutionary conservation were assessed via PhyloCSF, CodAlignView, and the ORBL tool.
Machine Learning: The Multi-Layer Perceptron Classifier Model and associated statistical tasks utilized TensorFlow (v2.18.0), scikit-learn (v1.5.x-1.6.x), and custom scripts.

Code Availability: Code generated for this manuscript is available at https://github.com/VanHeeschLab/deutsch_kok_et_al_2024 and Zenodo (https://doi.org/10.5281/zenodo.18878129). The Multi-Layer Perceptron Classifier Model code is accessible via https://git.embl.de/ivfimo/machine_learning_scripts and Zenodo (DOI: 10.5281/zenodo.18787106). The code for ORBL is available at https://github.com/iljungr/ORBL_tools and Zenodo (DOI: 10.5281/zenodo.18749292). Tiling screen local enrichment score scripts are available at https://github.com/CFVALLS/tiling_screens_with_permutation and Zenodo (DOI: 10.5281/zenodo.18865015).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All mass spectrometry data in this manuscript is publicly available through the PeptideAtlas database at https://peptideatlas.org/ and ProteomeXchange (https://proteomecentral.proteomexchange.org/). The Human HLA PeptideAtlas 2023-11 is freely accessible at https://peptideatlas.org/builds/human/hla/index_2023-11.php and the Human non-HLA PeptideAtlas 2023-06 is freely accessible at https://peptideatlas.org/builds/human/non-hla/. Specific dataset identifiers are listed in Extended Data Table 1. All ribosome profiling data manually inspected in this manuscript are publicly viewable at GWIPS-viz, as detailed in the methods. Mass spectrometry parallel reaction monitoring (PRM) data are deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD066599. Ribosome profiling, RNA sequencing, CRISPR barcode sequencing data for eight cell lines screened in this manuscript as well as OLMALINC knockdown bulk RNAseq and multiplexed single-cell RNAseq are all submitted to the NCBI Short Read Archive as PRJNA1294394. Primary gene and ncORF annotations were sourced from GENCODE (https://www.gencodegenes.org/), Ensembl Release 87 (https://www.ensembl.org/), UniProtKB/Swiss-Prot 2023 (https://www.uniprot.org/uniprotkb/), and the NCBI RefSeq database (https://www.ncbi.nlm.nih.gov/refseq/). Tissue-specific RNA-seq expression data were obtained from the Genotype-Tissue Expression (GTEx) portal (https://gtexportal.org/home/). Cancer dependency and CRISPR screening data were sourced from the DepMap portal (https://depmap.org/portal/). Ribosome profiling (Ribo-seq) data visualization and manual inspections were conducted using GWIPS-viz (https://riboseq.org/about.html). The Cancer Cell Line Encyclopedia (CCLE) was used for SNP extraction (https://sites.broadinstitute.org/ccle/datasets). The GSEA MSigDB was used to extract hallmark gene sets (https://www.gsea-msigdb.org/gsea/msigdb).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Our study does not involve human participants, clinical data, or biological material
Reporting on race, ethnicity, or other socially relevant groupings	Our study does not involve human participants, clinical data, or biological material
Population characteristics	Our study does not involve human participants, clinical data, or biological material
Recruitment	Our study does not involve human participants, clinical data, or biological material
Ethics oversight	Our study does not involve human participants, clinical data, or biological material

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size. The value of "n" in our manuscript is described in every figure legend and in the methods. For newly performed experiments, "n" refers to biological triplicates for CRISPR screens and quadruplicates for siRNA knockdown experiments. For analyzed datasets and studies, the value of "n" refers to the number of datasets, studies, datapoints, ncORFs, or identified peptides or proteins in these respective datasets. For proteomics database searches and evolutionary analyses, we used all ncORFs satisfying the specified condition rather than a sample chosen from them. The statistical significance of any conclusions are reported as p-values, indicating whether the number of data points was sufficient for any conclusion.
Data exclusions	No data points or samples were excluded from the analyses unless they failed predefined quality control criteria, which were applied uniformly across all datasets. For CRISPR screens, sgRNAs were excluded if they exhibited low counts falling more than three standard deviations below the mean of the total counts. In the pooled OLMALINC loss-of-function assays, cell lines were excluded if they contained missing values (NaNs) across replicates. For single-cell RNA-sequencing (scRNA-seq), cell lines with low representation—specifically those with fewer than 30–50 cells per identity—were discarded to prevent the inflation of statistical effects and ensure robust downstream analysis.
Replication	For CRISPR assays, lentiviral infections were performed in biological triplicate. OLMALINC siRNA knockdown experiments were performed with 4 technical replicates. All numbers of replicates are listed in the Methods and the respective Figure Legends.
Randomization	The MLP Classifier model was initialized with a maximum of 8000 iterations and a random state of 42 to ensure reproducibility. For protein structure predictions, ncORF sequences were randomly shuffled five times to assess the contribution of microprotein length and sequence composition to the AlphaFold structure prediction. For ORBLq calculations, randomly selected sets of 1,000 untranslated control ncORF sequences were selected as size, position, and ORF-type matched controls.
Blinding	Blinding was not applicable because the study did not involve experimental interventions or outcome assessment subject to investigator bias.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	The OLMALINC loss-of-function assays across 485 cell lines were performed as part of the PRISM project at the Broad Institute. Additional cell lines used for CRISPR tiling screens were purchased directly from the American Type Culture Collection (ATCC)
Authentication	Cell line identity was authenticated using a combination of Short Tandem Repeat (STR) profiling and Single Nucleotide Polymorphism (SNP) identification to ensure genomic consistency with reference standards
Mycoplasma contamination	All cell lines were routinely tested for mycoplasma contamination using the Lonza MycoAlert™. All results were confirmed negative prior to experimentation.
Commonly misidentified lines (See ICLAC register)	In accordance with ICLAC guidelines, all cell lines used in this study were cross-referenced against the Register of Misidentified Cell Lines. No commonly misidentified or contaminated cell lines were utilized in these experiments.

Palaeontology and Archaeology

Specimen provenance	*Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.*
Specimen deposition	*Indicate where the specimens have been deposited to permit free access by other researchers.*
Dating methods	*If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	*For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.*
Wild animals	*Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*
Reporting on sex	*Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☐	☐	Public health
☐	☐	National security
☐	☐	Crops and/or livestock
☐	☐	Ecosystems
☐	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☐	☐	Demonstrate how to render a vaccine ineffective
☐	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☐	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☐	☐	Increase transmissibility of a pathogen
☐	☐	Alter the host range of a pathogen
☐	☐	Enable evasion of diagnostic/detection modalities
☐	☐	Enable the weaponization of a biological agent or toxin
☐	☐	Any other potentially harmful combination of experiments and agents

Plants

Seed stocks	NA
Novel plant genotypes	NA
Authentication	NA

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	*For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*
Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. UCSC)	*Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates	*Describe the experimental replicates, specifying number, type and replicate agreement.*
Sequencing depth	*Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*
Antibodies	*Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Peak calling parameters	*Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*

Data quality	*Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*
Software	*Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*
Software	*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	*Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
Behavioral performance measures	*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s)	*Specify: functional, structural, diffusion, perfusion.*
Field strength	*Specify in Tesla*
Sequence & imaging parameters	*Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*
Area of acquisition	*State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software	*Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*
Normalization	*If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*

Normalization template

Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.

Noise and artifact removal

Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).

Volume censoring

Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

Statistical modeling & inference

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

(See Eklund et al. 2016)

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.