Reviewed Preprint
v1 • December 3, 2025
Not revised

Reviewed Preprint
v2 • May 27, 2026
Revised by authors

Evolutionary remodeling of non-canonical ORF translation in mammals

Yue Chang*, Tianyu Lei*, Feng Zhou, Jiawen Jiang, Yu Huang, Ziyang Zhu, Hong Zhang
College of Ecology, Lanzhou University, Lanzhou, China

For correspondence: hong_zhang@lzu.edu.cn

* These authors contributed equally

Competing interests: No competing interests declared

Funding: See page 17

Reviewing editor: Xiaohua Shen, Tsinghua University, China

eLife Assessment

This study presents a large, systematically curated catalog of non-canonical open reading frames (ncORFs) in human and mouse through the reanalysis of nearly 400 Ribo-seq datasets using a standardized pipeline; the resulting atlas consolidates ncORF annotations across tissues and provides a **valuable** resource for investigating non-canonical translation and ORF emergence. The main conclusions are supported by consistent data processing and multiple computational measures of translation and conservation. While the pipeline is transparent and technically robust, some analytical criteria and dataset limitations could be described more explicitly, and several downstream conclusions would benefit from more cautious interpretation, some evolutionary inferences are primarily correlative; dataset heterogeneity, uneven tissue representation, and limited experimental validation also constrain the strength of a subset of the findings. Overall, the evidence is **solid**, and the resource is likely to be broadly beneficial to the community.

https://doi.org/10.7554/eLife.109128.2.sa3

Abstract

Non-canonical open reading frames (ncORFs) are pervasive within transcripts annotated as “non-coding” or “untranslated regions” of mRNAs, yet their landscape under normal physiological conditions remains to be fully resolved, particularly outside humans. Here we applied a stringent and standardized pipeline to hundreds of high-quality ribosome profiling libraries from normal mammalian tissues and cell types, identifying 11,623 human and 16,485 mouse ncORFs. Evolutionary analyses revealed that thousands of ncORFs are subject to coding constraint and exhibit lineage-specific conservation, underscoring their functional potential. Ancient ncORFs are preferentially highly translated, broadly expressed, and enriched for lineage-specific conservation. Co-expression patterns further indicate that many ncORFs, especially ancient ones, are cotranslated with canonical coding sequences, consistent with functions mediated through protein–protein interactions. Together, these findings establish a comprehensive atlas of mammalian ncORFs and provide fundamental insights into their evolutionary dynamics and functional integration within the proteome.

Introduction

The systematic identification of functional genomic elements represents a foundational objective in the field of genomics. Among these elements, non-canonical open reading frames (ncORFs)—which are open reading frames (ORFs) located within long non-coding RNAs (lncRNAs) or untranslated regions (UTRs) of mRNAs—have attracted considerable attention in recent years[1-5]. While certain ncORF subtypes, such as upstream ORFs (uORFs) located in 5’ UTRs and downstream ORFs (dORFs) in 3’ UTRs, have been extensively characterized for their ability to regulate mRNA translation and stability in *cis*, a growing body of evidence demonstrates that ncORFs can also give rise to functional proteins that engage in diverse cellular pathways[1,2,6,7]. As functional ncORFs

continue to be discovered across multiple eukaryotes[8-17], systematic elucidation of their composition and function promises to uncover an additional layer of regulatory complexity and to refine our conception of the so-called "non-coding" genome.

Although ncORFs may play important biological roles, the number of translated ncORFs in the human genome remains uncertain. Detection by mass spectrometry (MS)-based proteomics is challenging, likely due to their short lengths and low abundances[18]. Ribosome profiling (Ribo-Seq) offers greater sensitivity for identifying translation events[18,19], but downstream computational pipelines often produce highly divergent ncORF sets [20]. Consequently, published estimates of ncORF abundance span several orders of magnitude—ranging from thousands to millions [15-22]. To establish a standardized reference, the GENCODE consortium integrated data from seven distinct studies to produce a catalog of 7,264 human ncORFs (hereafter GENCODE ncORFs)[2]; however, the heterogeneity of sample sources and prediction algorithms likely introduced systematic bias. This catalog is also restricted to AUG-initiating ORFs, thereby overlooking those with near-cognate start codons. A later effort generated a high-resolution human translatome map and identified 7,767 ncORFs with a unified pipeline[22]. Nevertheless, this conservative annotation disproportionately under-represents coding sequence (CDS)-overlapping ncORFs. Indeed, later estimation suggests that between 10,500 to 22,500 ncORFs should exist in the human genome[18]. Furthermore, many ncORFs were reported from studies of cancer cells[1,13,23,24], where extensive transcriptome reprogramming alters the translational landscape[25,26]. While paired healthy tissues were also analyzed in these studies, the landscape of ncORFs translated under normal physiological conditions is not yet fully resolved.

While a limited subset of ncORFs has been shown to encode signaling molecules, cofactors, membrane proteins, or subunits of large protein complexes[1,27,28], functions of most ncORF-encoded proteins (ncEPs) are largely uncharacterized. Advances in high-throughput screen technologies have enabled the systematic interrogation of functional ncEPs[23,28-32]. Nevertheless, there is minimal concordance among hits from different studies[3]. Evolutionary analyses can offer insights into potential functionality of these translation products and mechanisms of origination[33-35], yet how ncORF translation has been remodeled to integrate into proteome since their evolutionary origins is still poorly understood. Consequently, the extent to which translated ncORFs encode functional proteins and the evolutionary dynamics of their integration into the existing proteome are largely unresolved. Moreover, because most ncORF studies have focused on humans, it is unclear whether ncORFs in other species share similar sequence features and expression patterns.

To address these gaps, we generated a standardized annotation of translated ncORFs by uniformly reprocessing ~400 high-quality ribosome profiling libraries from normal mammalian tissues and cell types. This stringent approach yielded 11,623 human and 16,485 mouse ncORFs, consistently recovered across diverse samples and prediction strategies. Comparative and functional genomic analyses revealed that thousands of these ncORFs are under evolutionary constraint and exhibit distinct translation profiles. The resulting dataset provides a robust, comprehensive resource that will facilitate future investigations into the biological roles of ncORFs across cellular and organismal contexts.

Results

Annotation of ncORFs translated in normal mammalian cells

We developed a standardized pipeline to systematically annotate ncORFs translated in normal human and mouse cells (Fig. 1A). We first curated publicly available Ribo-Seq libraries, excluding those from pathological or genetically modified samples. Since both library size and the proportion of in-frame ribosome-protected fragments (RPFs) affect the sensitivity and accuracy of ncORF prediction[20], we filtered out libraries containing fewer than 5 million RPFs or with in-frame rates below 60%. After this quality control step, we retained 174 human and 209 mouse high-quality Ribo-Seq libraries, spanning a range of tissues and cell types (Fig. S1A; Table S1). We predicted translated ORFs that begin with NUG (where N stands for any nucleotide) start

codons and encode at least five amino acids using three tools—PRICE[36], RiboCode[37] and RibOTISH[38]—which we previously benchmarked for optimal performance[20]. After excluding annotated CDSs and their extension or truncation variants arising from alternative initiation sites, we obtained a preliminary list of ncORFs in each library.

To generate reliable ncORF annotations, we applied stringent filtering and integration criteria. We removed potential false positives caused by non-ribosomal contamination or translational noise based on metrics including Fragment Length Organization Similarity Score (FLOSS)[39], Ribosomal Release Score (RRS)[40,41] and the number of RPFs within an ORF by P-site. We then aggregated the remaining ORFs to identify those consistently detected across multiple methods and libraries in human and mouse datasets, respectively (Fig. 1A). In many cases, two ORFs predicted from different libraries or methods may share the same start or stop codon due to the alternative translation initiation or splicing, resulting in groups of overlapping ncORFs. For example, 11.1% (805) of GENCODE ncORFs overlap with at least one other ORF, forming 329 overlapping groups (Table S2). This overlap complicates the identification of reproducible ORFs, as ORFs within a group may represent distinct proteoforms of a protein. To address this issue, we collapsed each group of ncORFs sharing the same start or stop codon into a single cluster, and retained only those clusters supported by at least two independent methods and observed in more than one library. For each retained cluster, we selected a representative ORF by prioritizing transcript annotation[42,43], start codon type, and ORF length to facilitate downstream analyses (Fig. 1A).

This approach yielded 12,469 translated ncORFs in humans and 16,960 in mice. However, manual inspection revealed a subset of spurious ncORFs that are not truncation or extension variants of annotated CDSs, but instead partially overlap with annotated CDSs from alternative transcript isoforms in frame (Figs. 1B & C). For instance, in gene *AGTPBP1*, the CDS of transcript *ENST00000357081* begins downstream of *ENST00000628899* CDS due to the inclusion of an alternative exon. As a result, the 5' UTR of *ENST00000357081* appears to contain a translated uORF (Fig. 1B), which overlaps in-frame with the CDS of *ENST00000628899* and may be mistakenly identified as a translated ncORF owing to active translation of the latter. A similar case is seen with the lncRNA ORF (lncORF) in *ENST00000462211*, which overlaps in-frame with the CDS of *ENST00000328767* from the *TMTC4* gene and may thus be erroneously identified as a translated ncORF (Fig. 1C). To avoid such artifacts, we implemented an additional filtering step to exclude these overlapping cases and left with 11,623 and 16,485 high-confidence ncORFs in humans and mice, respectively (Table S3). Notably, while more than half ncORFs were detected in two to five libraries, 22.3–26.3% were identified in more than 10 libraries, indicating a substantial subset with robust and recurrent translation signals suitable for future experimental validation (Fig. S1B).

Annotated ncORFs are supported by independent evidence

To assess whether the annotated ncORFs are supported by independent evidence, we compared them with a previously compiled catalog of ncORFs identified using mass spectrometry (MS)-based proteomic data[20]. For the 3,494 MS-supported ncORFs in humans and 873 in mice, 44.5% and 67.6%, respectively, were also predicted as translated ncORFs in this study (Fig. S2). We further evaluated our annotation against the GENCODE ncORFs[2], a widely used community resource. Of the 7,264 GENCODE ncORFs, 6,974 are located in fully annotated transcripts that were included in our analysis. Among them, 2,997 were reported by at least two independent studies (hereafter "phase 1"), while the remaining were only reported in a single study ("single")[2]. We found that 75.6% of phase 1 ncORFs were independently recovered in our annotation, significantly more than the single-study ncORFs (Fig. 1D; $P = 6.4\times10^{-307}$, Fisher's exact test).

A recent study conducted a comprehensive pan-proteome analysis to provide high-confidence peptide evidence for GENCODE ncORFs[44]. In this study, ncORFs were categorized into four tiers: tier 1 and 2 ncORFs had support from both MS and Ribo-Seq data; tier 3 had MS evidence only; and tier 4 had Ribo-Seq evidence only[45]. Notably, our annotation was significantly enriched for tier 1 or 2 ncORFs ($P = 2.0\times10^{-64}$, Fisher's exact test), with 446 and 661 of them detected, respectively (Fig. 1E). Combining MS-supported ncORFs from our previous compilation and those from GENCODE ncORFs, 2,088 (18.0%) human ncORFs identified in this study were supported by MS evidence.

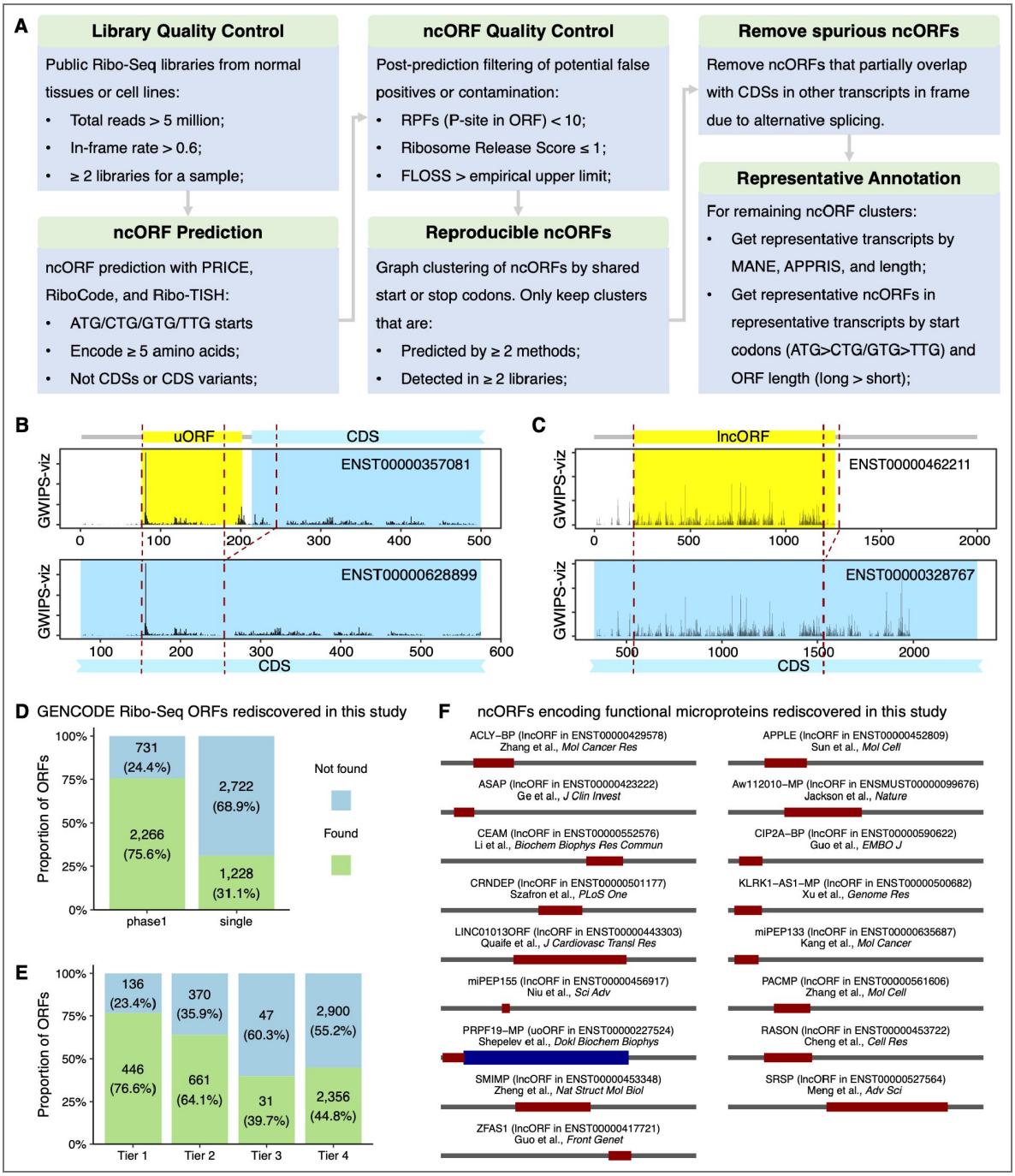

Figure 1. Annotation of ncORFs in humans and mice.

(**A**) Overview of the integrative pipeline for ncORF annotation in human and mouse genomes. (**B-C**) RPF coverage plot for a representative uORF (B) and a lncORF (C), visualized using the Ribo-Seq signal track from GWIPS-viz[72]. ncORFs were shown in yellow and CDSs in cyan. Dashed lines indicate identical genomic regions shared across transcript isoforms. (**D-E**) Proportion of GENCODE ncORFs rediscovered in this study. GENCODE ncORFs were stratified either by reproducibility across independent studies (D) or by tiered translation evidence strength (E). (**F**) Known ncORFs that were experimentally characterized in earlier studies and independently rediscovered in this study from humans and mice. CDS, coding sequence; FLOSS, fragment length organization similarity score; MANE, matched annotation from NCBI and EBI; ncORF, non-canonical open reading frame; ORF, open reading frame; RPF, ribosome-protected fragment.

Despite this overlap, the majority (65.2%) of human ncORFs in our annotation are not included in either of the two catalogs. This highlights both the complexity of the non-canonical translatome and the critical role of tissue diversity in comprehensive ncORF annotation.

To further validate our findings, we manually curated ncORFs with known functions from the literature. We identified 17 ncORFs in our annotation that have been experimentally confirmed to be translated—either through reporter assays or antibody detection—and shown to produce functional proteins (Fig. 1F; Table S4). Notably, 16 of these validated ncORFs are from humans, and only 12 are included in the GENCODE catalog. This analysis further underscores the accuracy and functional relevance of our annotations. Together, these results indicate that our annotation pipeline generates a robust and reliable catalog of translated ncORFs in mammals.

ncORFs exhibit signatures of translation and encode putative proteins lacking known domains

We classified ORFs into categories based on community recommendations[2] and found that the distribution of ORF types closely aligns with previous estimates[18]. In both human and mouse, uORFs, upstream overlapping ORFs (uoORFs), and lncORFs are more prevalent than other types, whereas internal ORFs (iORFs), downstream overlapping ORFs (doORFs), and dORFs are relatively rare (Fig. 2A). Although most ncORFs are short, encoding fewer than 100 amino acids (Fig. 2B), their sequences show clear signatures of translation similar to CDSs. These include preferential usage of AUG start codons (Fig. S3A), enrichment toward 5' transcript ends (Fig. S3B), and reduced RNA secondary structure[46] (Fig. S3C) and optimized Kozak contexts[47,48] around start codons, especially in non-AUG ncORFs (Figs. S3D & S4). Despite these similarities, ncORFs differ markedly from CDSs in both codon and amino acid usage (Figs. S5A-B). This divergence is further supported by their lower tRNA adaptation index values and higher effective number of codons, indicating less optimized codon usage (Figs. S5C-D). Together, these findings show that ncORFs share core sequence and structural features with canonical CDSs, yet maintain distinct codon and amino acid usage profiles.

Although ncORFs can encode putative proteins, the functional significance of ncEPs remains poorly understood. As an initial step toward functional inference, we systematically searched for annotated protein domains within ncEP sequences, reasoning that the presence of conserved domains would suggest potential biochemical activity. Overall, only 1.29% (150) of human and 0.78% (129) of mouse ncEPs contained at least one known domain (Fig. 2C; Table S5), presumably due to their short lengths relative to canonical proteins. Consistently, domain-containing ncORFs are significantly longer than those without annotated domains (Fig. S6) and are predominantly classified as lncORFs (Fig. 2D). Unlike canonical proteins, which have relatively low intrinsically disordered region (IDR) content (median ± standard error: 14.6±0.28% in humans and 12.7±0.26% in mice), ncEPs exhibit significantly higher IDR proportions (Fig. 2E), with medians of 62.5±1.42% and 60.0±1.27%, respectively. These results indicate that most ncEPs lack structured domains and are instead dominated by IDRs.

Among the eight domains observed in at least five ncORFs, three are DNA-binding domains: zinc finger, Krüppel-associated box, and helix-turn-helix (Fig. 2F). Because transposable elements (TEs) commonly encode DNA-binding proteins required for transposition[49], we tested whether domain-containing ncORFs preferentially overlap TE-derived sequences. Indeed, domain-containing ncORFs were significantly enriched for TE overlap in both humans and mice (P = 5.8×10^{-4} and 3.4×10^{-17}, respectively; Fisher's exact tests; Fig. 2G). Although enrichment alone does not establish origin, it supports a model in which a subset of longer ncORFs may have acquired modular domains through TE-derived sequence contributions.

Putative ncEPs are selectively constrained

Given the scarcity of known domains in putative ncEPs, a key question is whether they are functional in cells. If at least some of these ncEPs are not merely translational noise or byproducts of translation-dependent functions, we would expect them to exhibit evolutionary signatures of

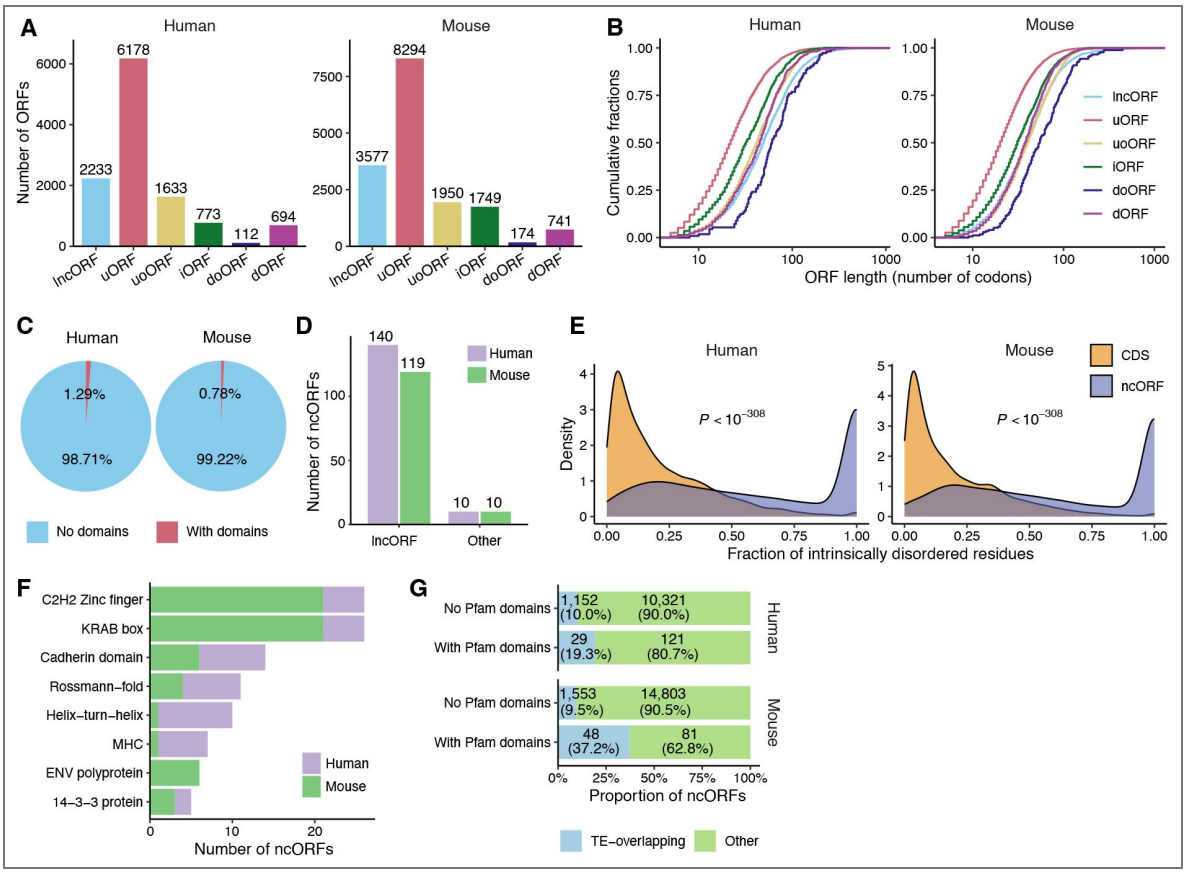

Figure 2. Sequence features of ncORFs and putative ncEPs in humans and mice.

(**A**) Number of ncORFs belonging to different categories. (**B**) Distribution of ncORF lengths. (**C**) Proportion of ncEPs that contain known Pfam domains. (**D**) Number of ncORFs with Pfam domains among lncORFs and other ncORF categories. (**E**) Distribution of the proportion of intrinsically disordered residues in CDS- and ncORF-encoded proteins. (**F**) Most frequently identified Pfam domains among human and mouse ncORFs. (**G**) Proportion of ncORFs overlapping with TEs, stratified by presence or absence of Pfam domains. Differences were assessed using Wilcoxon rank-sum tests. CDS, coding sequence; ncORF, non-canonical open reading frame; ncEP, ncORF-encoded protein; TE, transposable element.

selective constraints similar to canonical CDSs. To test this, we leveraged the recent high-resolution genomic constraint maps derived from both intra- and inter-species comparisons in humans [50,51]. First, we examined the genomic non-coding constraint of haploinsufficient variation (Gnocchi), which quantifies selective pressure by comparing observed versus expected variants across human genomes. A Gnocchi score above zero indicates purifying selection[50]. Although ncORFs exhibit lower Gnocchi scores than CDSs, their score remain significantly above zero and exceed those of lncRNAs (Fig. S7), suggesting that ncORF sequences are under evolutionary constraint. A hallmark of functional coding regions is codon position-specific constraints, but Gnocchi scores lack the resolution needed to assess such fine-scale features. Therefore, we next analyzed the base-resolution conservation metric PhyloP[52] calculated from multiple alignments of primate or mammalian genomes[51]. As expected, PhyloP scores in CDSs are substantially higher than those in flanking UTRs and display strong three-nucleotide periodicity, with the third codon position being less conserved than first and second (Fig. 3A). Although ncORFs (excluding those overlapping annotated CDSs) do not show higher overall PhyloP than flanking regions, they also exhibit significant three-nucleotide periodicity in both mammals (Fig. 3A) and primates (Fig. S8), suggesting that at least a subset of ncORFs are subject to similar evolutionary constraints as CDSs.

Motivated by the above results, we next sought to identify individual ncORFs under selective pressure. To this end, we determined their modes of origination[33,34] and evolutionary history through comparative genomic analyses. Our analysis revealed that most ncORFs are evolutionarily young, with 63.0% (7,319) and 82.8% (13,508) ncORFs in humans and mice originating within the mammalian lineage, respectively (Figs. 3B & S9). In line with previous studies[33,34], we estimate that at least 51.7% of human and 50.8% of mouse ncORFs arise *de novo* (Table S6). We also observed that the number of ncORFs originating at each ancestral node, when normalized by branch length, decreases with increasing node-to-tip distance (Fig. 3C), reflecting dynamic turnover and reduced long-term retention of ncORFs. To assess whether individual ncORFs are constrained for coding, we analyzed codon substitution patterns within ncORFs since origination[33] with PhyloCSF[53]. PhyloCSF estimates the log-likelihood ratio of a given alignment under coding versus non-coding models; positive values indicate stronger similarity to conserved coding sequences. We found 13.8% (1,608) of human and 32.0% (5,281) of mouse ncORFs have positive PhyloCSF scores (Fig. 3D), and 10.9% (1,264) of human and 22.6% (3,721) of mouse ncORFs even have PhyloCSF exceeding 10 (Table S6), providing strong evidence that these sequences may encode functional ncEPs.

Evolutionary dynamics can also provide insights into the function of ncORFs, as functionally constrained ncORFs are less likely to be lost over time. To quantify the conservation of each ncORF since its origination, we developed a local branch length score (BLS) (Fig. 4A), adapting the "global" BLS metric previously used to assess uORF conservation across full phylogenies[48,54]. Unlike global BLS, local BLS is designed to detect lineage-specific constraint, making it better suited for identifying functional ncORFs that may be restricted to certain evolutionary branches. Using a local BLS threshold of 0.9, we identified 16.3% (1,889) of human and 24.8% (4,087) of mouse ncORFs as exhibiting lineage-specific conservation (Figs. 4B & S9). Notably, these ncORFs are significantly more likely to show positive PhyloCSF scores (Fig. 4C), reflecting the effectiveness of local BLS. In total, we identified 681 human and 1,622 mouse ncORFs that are both lineage-specifically conserved and show evidence of coding potential, making them strong candidates for future functional validation. Together, these findings suggest that a subset of ncORFs may encode functional proteins rather than translational byproducts.

Expression of ncORFs is shaped by their evolutionary dynamics

Gene expression is a tightly regulated process, and non-functional products are unlikely to be robustly expressed. To assess the expression of ncORFs, we examined their translation across diverse tissues and cell lines (Table S7). On average, 71.9% of human and 48.5% of mouse ncORFs are translated with reads per kilobase per million mapped reads (RPKM) ≥ 1, which is comparable to or slightly exceeds the proportions observed for CDSs (Fig. S10). To investigate

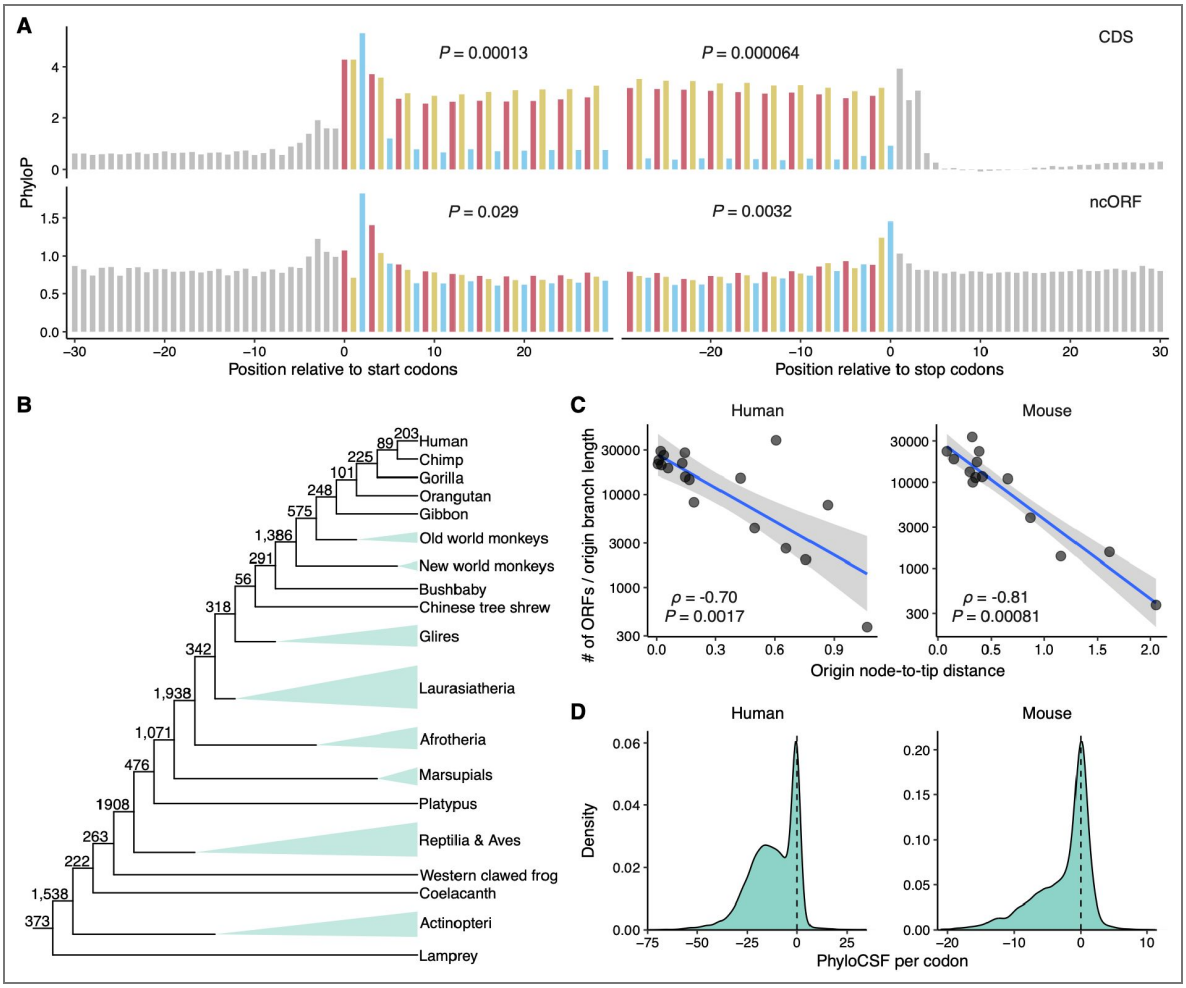

Figure 3. Evolutionary constraints of ncORFs.

(**A**) The mean PhyloP scores of 30 base pairs upstream and downstream of the ORF start or stop codons in mammals. Codons were delineated with bars of alternating colors by different frames, and nucleotides in untranslated regions were shown in grey. The significance of three-nucleotide periodicity was assessed by autocorrelation with a lag of three. (**B**) Cladograms of vertebrates with the number of ncORFs originating at each ancestral branch of humans. Triangles indicate species merged into larger clade for visual simplicity. (**C**) Relationship between ORF origination rates and node ages measured as node-to-tip distances. Origination rate was defined as the number of ncORFs that originated on a branch divided by the branch length. Blue lines show linear regression fits, and grey bands represent 95% prediction confidence intervals. Spearman's correlation is indicated. (**D**) Distribution of ncORF PhyloCSF scores normalized by the number of codons in per ncORF. The dashed line denotes zero. CDS, coding sequence; ncORF, non-canonical open reading frame; ORF, open reading frame.

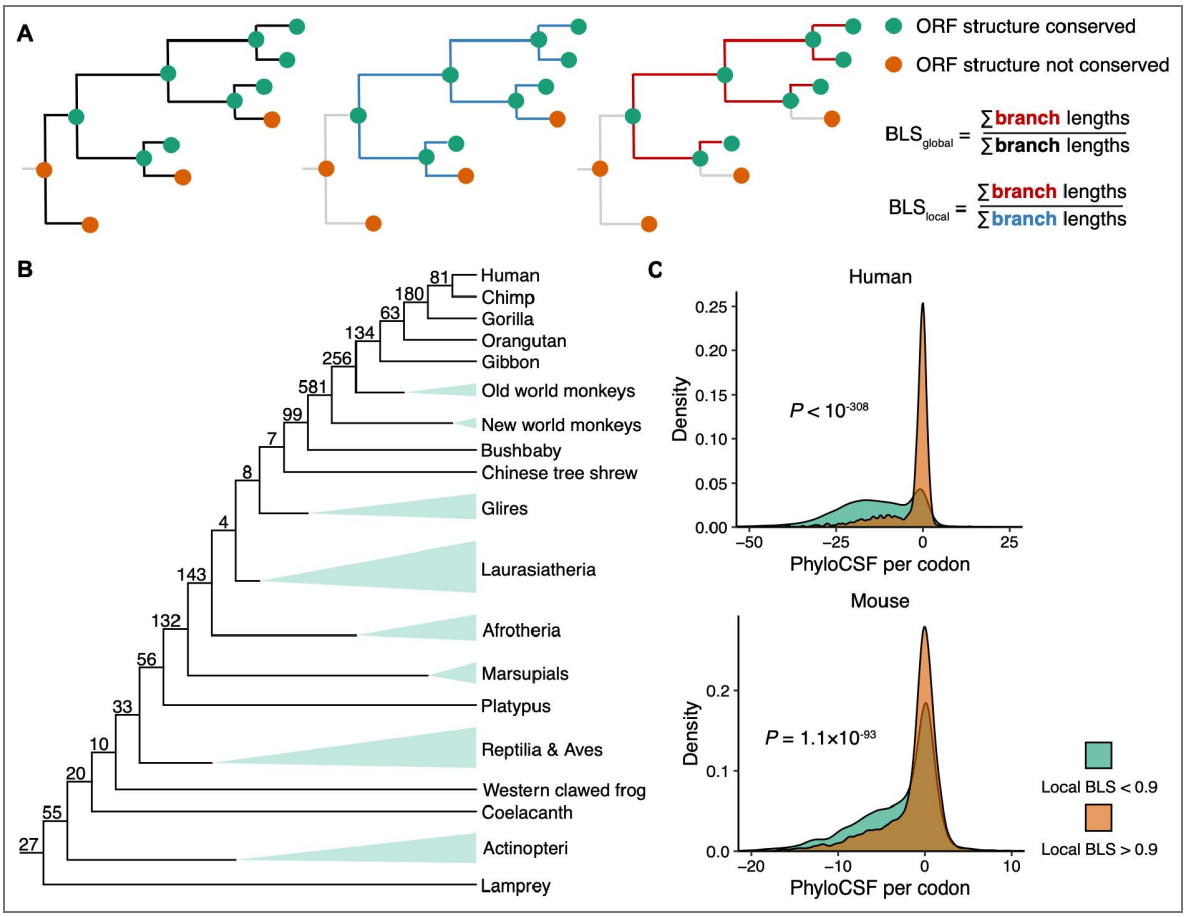

Figure 4. Lineage-specifically conserved ncORFs.

(**A**) Schematic illustration for BLS calculation. (**B**) Cladograms of vertebrates showing the number of lineage-specifically conserved ncORFs (local BLS > 0.9) at each ancestral node for humans and mice. (**C**) Distribution of PhyloCSF scores per codon for lineage-specifically conserved ncORFs compared to all other ncORFs. BLS, branch length score; ncORF, non-canonical open reading frame; ORF, open reading frame.

the influence of ncORF evolution on their expression, we analyzed mean translation levels of ncORFs across different samples. We found that ncORFs with higher local BLS tend to be more highly translated in both species (Fig. S11A), whereas coding potential showed only a minor influence (Fig. S11B). To disentangle the effects of evolutionary constraint and gene age, we stratified ncORFs by origination nodes and examined their translation levels in relation to local BLS and PhyloCSF. Although older ncORFs are generally more highly translated, those with higher local BLS scores consistently show elevated translation levels compared to their age-matched counterparts (Figs. 5A & S12A). This suggests that lineage-specific conservation is associated with enhanced and possibly optimized translation, potentially reflecting acquired functional roles. Conversely, the relationship between coding potential and translation is more nuanced: younger ncORFs with positive PhyloCSF scores tend to be more highly translated, while the opposite trend is observed in older ncORFs (Figs. 5B & S12B). These patterns point to complex interplays between evolutionary constraint, coding potential, and translational output, likely reflecting divergent selective pressures across different ncORF age groups.

While the above analyses revealed the overall trends in ncORF translation, they overlooked the heterogeneity across tissues and cell types. We found that ncORF translation exhibits strong tissue specificity (Fig. S13), which may stem from either transcriptional bias in ncORF-containing genes or post-transcriptional regulation specific to ncORFs. To disentangle these factors, we examined RNA levels of ncORF-containing genes across different tissues[55,56]. Although transcriptional enrichment in testis was evident in both humans and mice (Fig. S14), testis-biased translation was observed only in mice (Fig. S13). Moreover, ncORF translation displayed significantly higher tissue specificity than the transcription of their host genes in both species (Fig. 5C), indicating a substantial post-transcriptional component to their regulation. We next assessed how evolutionary characteristics shape tissue specificity. Older ncORFs tend to have lower tissue specificity than younger ones, suggesting they are more broadly translated (Figs. 5D & S12C). Intriguingly, coding potential exhibits opposite effects depending on evolutionary age: younger ncORFs with higher coding potential are more broadly expressed, whereas older ones show increased tissue specificity (Figs. 5E & S12D). These patterns may help to explain the lower average translation levels observed for older ncORFs across tissues. A similar trend was also evident for ncORFs with high local constraint scores in mice (Fig. S12C). Collectively, these results support a model in which ncORFs initially evolve widespread expression to establish interactions and avoid loss[57], but may later become functionally specialized, resulting in more restricted expression.

Extensive co-translation between ncORFs and canonical CDSs

Given our earlier observation that many ncEPs lack modular domains and may act via interactions with canonical proteins, we next explored the landscape of co-translation between ncORFs and CDSs across different tissues or cell types as we expect that interacting proteins tend to be translated simultaneously spatiotemporally. Using weighted gene co-expression network analysis (WGCNA)[58], we identified 61,513 human and 55,449 mouse ncORF-CDS co-translated pairs. As expected, ncORFs in mRNAs tend to co-translate with the main CDS of the same gene, even after excluding CDS-overlapping ones (Fig. S15), consistent with previous experimental findings[28]. Among non-overlapping co-translating ncORFs, only 10 in humans were dORFs; all others in both species were uORFs. Of note, co-translation does not contradict the canonical model in which uORFs repress downstream CDS translation. Co-translation reflects concurrent ribosome occupancy driven by overall initiation rates at the 5' end, whereas repression concerns the fraction of initiating ribosomes that reach the CDS. Thus, absolute translation of both ORFs may increase without altering the relative inhibitory effect of the uORF.

ncORF-CDS co-translation forms a highly modular bipartite network, in which one CDS may be linked to multiple ncORFs and vice versa (Fig. 6A-B). Network analysis revealed that hub CDS nodes—those with the most connections—in the largest modules are significantly enriched for functional gene categories (Fig. 6C-D). For instance, the largest module in humans is associated with chromatin-related proteins, while the second largest module in humans and the largest one

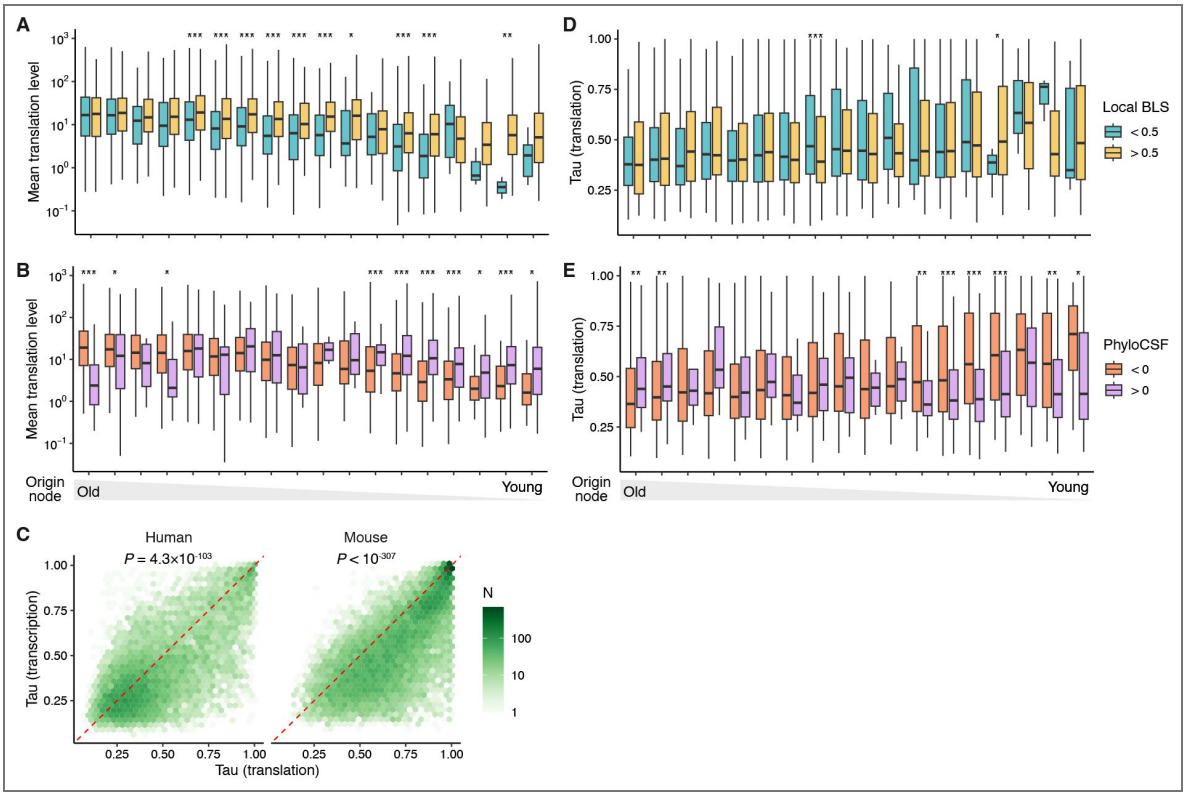

Figure 5. Evolutionary dynamics of ncORF expression.

(**A**) Distribution of mean translation levels of human ncORFs grouped by their origin nodes and further stratified by local BLS. Statistical significance was assessed using Wilcoxon rank-sum tests. ***, $P < 0.001$; **, $P < 0.01$; *, $P < 0.05$. (**B**) Similar to (A) but ncORFs are further stratified by local BLS. (**C**) Relationship between ncORF tissue specificity at translation and transcription levels. Differences were determined with Wilcoxon signed-rank tests. (**D-E**) Similar to (A) and (B) but showing the distribution of tissue specificity at the translation level. CDS, coding sequence; ncORF, non-canonical open reading frame; ORF, open reading frame.

in mice are linked to extracellular matrix organization. Notably, ancient ncORFs (those originating before the last common ancestor of mammals) are significantly more likely to co-translate with CDSs than younger ones (Fig. 6E), potentially reflecting a greater likelihood of acquiring functional interactions over evolutionary time. In summary, our analyses revealed that evolutionary dynamics of ncORFs have widespread impacts on their translation and potential function.

Discussion

Reliable annotation of translated ncORFs is critical for understanding their functional roles and evolutionary significance. In this study, we developed a standardized pipeline for the annotation of translated ncORFs and applied it to both humans and mice due to the extensive publicly available Ribo-Seq datasets. Compared to previous efforts, our annotation framework offers several key improvements. First, we selected only high-quality libraries with sufficient read coverage and acceptable in-frame RPF rates. Our earlier work demonstrated that both library size and in-frame rates impact the reproducibility and accuracy of ncORF predictions[20]. This step helps mitigate false annotations arising from suboptimal data quality. Second, although many previously characterized functional ncEPs are associated with cancer-related ncORFs[1], ncORFs may also have important functions under physiological conditions. Importantly, such ncORFs are more likely to be subject to evolutionary constraints. Therefore, we restricted our analysis to normal cells and tissues, allowing us to identify more biologically relevant and conserved ncORFs. Third, given the variable performance across ncORF detection methods and the low reproducibility of predictions between biological replicates[20], we implemented stringent reproducibility criteria. Specifically, we retained only ncORFs detected by multiple methods and across independent datasets, enhancing the robustness of our final annotation. Finally, we applied a suite of well-validated filtering metrics from prior studies[39-41] to minimize the inclusion of false positives, including those arising from translational noise[59,60] or non-ribosomal contamination. As a result, our annotated ncORFs show significant enrichment for GENCODE ncORFs, especially those with MS evidence or reproducible detection across independent studies. These features collectively underscore the reliability and utility of our annotation for future functional and evolutionary studies of non-canonical translation.

However, our catalog may remain incomplete due to the uneven distribution of publicly available Ribo-seq datasets across tissues and cell types. Broader and more balanced tissue coverage will be necessary to generate a more comprehensive and less biased ncORF annotation, which would also enhance the cotranslation analysis. Additionally, our proteomic validation relies on previously compiled sets of MS-supported ORFs rather than de novo reanalysis of raw mass spectrometry datasets. Incorporation of candidate ncORFs into reference proteome databases prior to large-scale, standardized searches of public MS runs would substantially improve detection sensitivity and provide stronger evidence for their translational products[45].

From the perspective of new gene evolution, ncORFs represent a transitional stage between non-coding DNA/RNA and protein-coding genes[61,62]. A major challenge in reclassifying ncORFs as bona fide coding genes lies in determining whether the putative proteins they encode are functionally relevant. Our analysis of ncEPs revealed that most lack recognizable protein domains and are enriched for IDRs. Therefore, we propose that many ncORFs may function in a non-autonomous manner by interacting with other proteins to exert their effects. This hypothesis aligns with prior studies showing physical interactions between ncEPs and known proteins[28,29,34]. Supporting this view, our WGCNA analysis uncovered widespread co-translation between ncORFs and canonical CDSs, suggesting potential interaction partners and functional contexts.

We also assessed whether ncORFs are subject to evolutionary constraints consistent with functional protein coding. Using both population-level and cross-species analyses, we found that a substantial fraction of ncORFs bear signatures of purifying selection, in line with previous studies[33,63-66]. Notably, their conservation profiles exhibit nucleotide-level features characteristic of canonical CDSs, supporting the interpretation that selection acts on the encoded peptide products rather than merely on translational activity per se. Furthermore, we reconstructed the

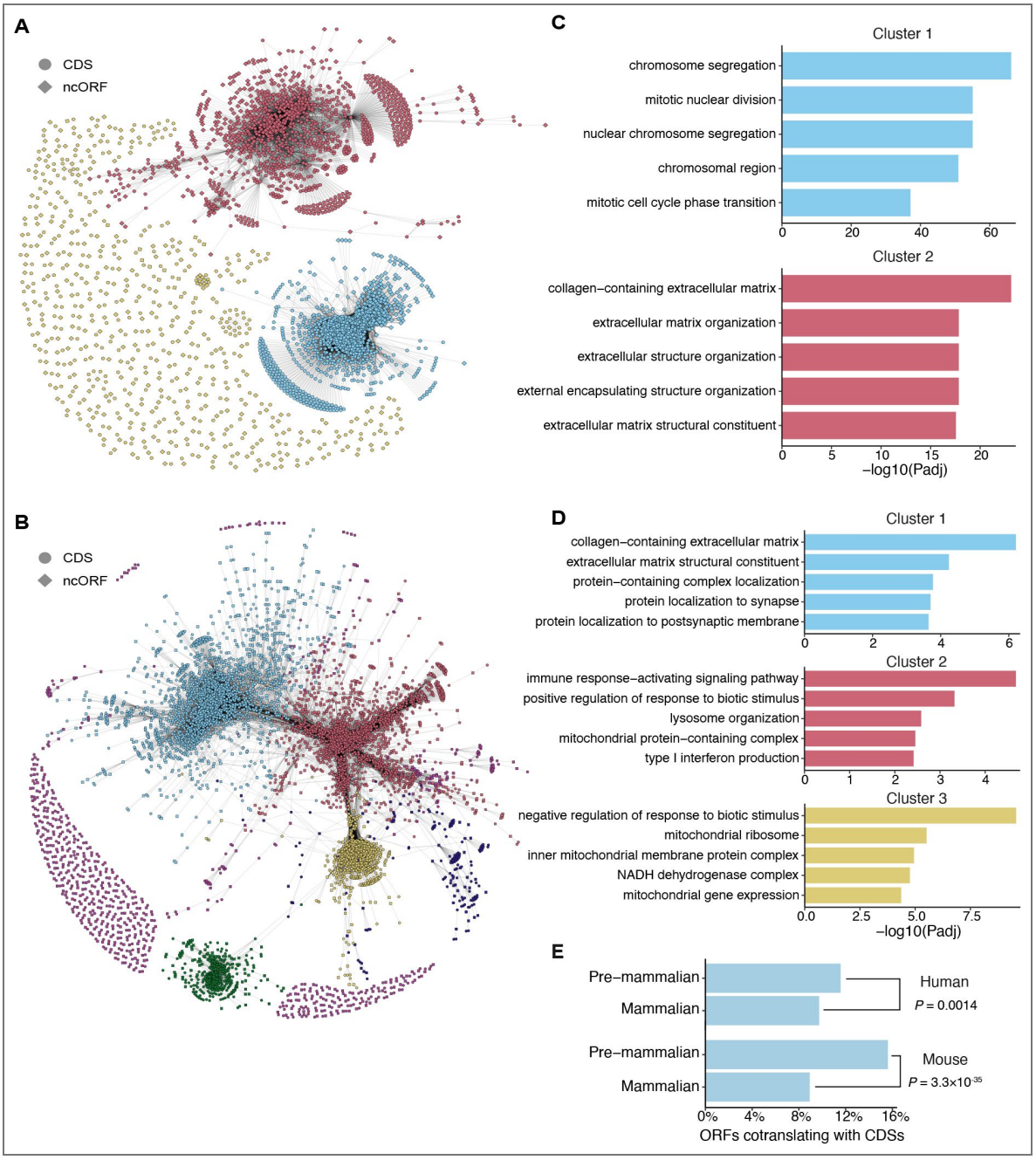

Figure 6. ncORF-CDS co-translation network.

(**A**) Bipartite network of ncORF-CDS co-translation in humans. (**B**) Bipartite network of ncORF-CDS co-translation in mice. ncORFs and CDSs are represented by different shapes and colored according to their cluster membership. Only the largest clusters were highlighted (top two in humans and top five in mice). (**C**) Top five enriched gene ontology terms for each of the two largest clusters in the human network shown in (A). (**D**) Top five enriched gene ontology terms for each of the three largest clusters in the mouse network shown in (B). (**E**) Proportion of ncORFs cotranslating with CDSs among ancient (pre-mammalian origin) versus younger (mammalian-specific) ncORFs. Differences were tested with Fisher's exact tests. CDS, coding sequence; ncORF, non-canonical open reading frame; ORF, open reading frame.

evolutionary trajectories of thousands of ncORFs in both humans and mice and observed sustained coding potential since their origination. Complementing these findings, we identified a subset of ncORFs that preserve intact ORF structures across most descendant species, suggesting ongoing purifying selection against disabling mutations. These lineage-specifically conserved and potentially coding ncORFs likely contribute to clade-specific functional innovation and represent compelling candidates for future functional characterization.

We further explored how evolutionary history shapes ncORF translation across tissues or cell types. Our results indicate that older ncORFs are more broadly and highly translated and have established more extensive co-expression with canonical CDSs. Moreover, when comparing ncORFs of similar age, those that are lineage-specifically constrained consistently exhibit higher translation levels, reinforcing their likely functional importance. Interestingly, we found that younger ncORFs with strong evolutionary constraints tend to have lower tissue specificity in their translation, potentially reflecting a phase of widespread expression that facilitates functional recruitment. In contrast, older ncORFs with positive coding potential tend to display increased tissue specificity, implying functional specialization over evolutionary time. Collectively, these findings support a model in which ncORFs evolve from broadly expressed, weakly functional elements into specialized, conserved effectors— though further empirical work is needed to validate this hypothesis.

In conclusion, we provide a comprehensive annotation of translated ncORFs in humans and mice, offering a valuable resource for functional and evolutionary genomics. Our integrated analysis of sequence evolution and translational landscape sheds light on the evolutionary trajectory and functional integration of ncORFs. The consistency of these patterns in both species suggests that our findings may generalize to other mammals. However, realizing the full biological significance of ncORFs, particularly lineage-specific ones, will require experimental validation. Recent advances in high-throughput screening technologies[19,29,32,67] offer a promising path forward. Lastly, although our study focused on humans and mice due to the availability of high-quality Ribo-Seq data, we anticipate that our annotation will serve as a training resource for machine learning models aimed at predicting translated ncORFs in non-model organisms, thus broadening our understanding of genome coding potential across species.

Methods

Quality control and preprocessing of Ribo-Seq libraries

Ribo-Seq libraries from normal human or mouse samples without genetic modifications or perturbations (Table S7) were downloaded from the European Nucleotide Archive (ENA) or the Short Read Archive (SRA). For ncORF annotation, libraries containing fewer than five million reads were excluded. Libraries derived from cancer tissues or cell lines were also removed based on sample metadata or the associated publications. Next, adapter sequences and any random nucleotide additions were trimmed using Cutadapt (v3.7)[68] and reads shorter than 18 nucleotides were discarded. Remaining reads were aligned to the human or mouse reference genome (GRCh38 or GRCm39, ENSEMBL[69] release 107) using the STAR aligner (v2.7.10a)[70]. To assess library quality, the proportion of in-frame ribosome-protected fragment (RPF) reads was calculated using the PSite package[71]. Libraries with ≥60% in-frame RPFs were classified as high quality. Only samples containing at least two high-quality libraries were retained for ncORF annotation (Table S1). For the expression analyses, all the downloaded Ribo-Seq libraries were used to cover more samples.

ncORF annotation

For each Ribo-Seq library, translated ORFs initiating with AUG, CUG, GUG, or UUG and encoding at least five amino acids were predicted using PRICE[36], RiboCode[37], and Ribo-TISH[38], as described previously[20]. To define novel ORFs, annotated CDSs and their truncation or extension variants (i.e., ORFs with the same start or stop position as annotated CDSs) were excluded. To eliminate likely false positives resulting from non-ribosomal contamination or stochastic molecular errors, we calculated RRS[40,41] and FLOSS[39] of each ncORF. RRS measures the drop in RPF coverage after

stop codon—a key signature of active translation beyond three-nucleotide periodicity[40,41], whereas FLOSS assesses whether the RPF length distribution resembles that of canonical CDSs, helping to distinguish authentic ORFs from non-ribosomal signals[39]. In each library, uniquely mapped RPFs were assigned to the corresponding P-site, and genome-wide P-site coverage was computed with the PSite package. Similar to previous studies[40,41], RRS was calculated as the sum of coverage across the five codons preceding the stop codon divided by the sum across the five trinucleotides following the stop codon. FLOSS was calculated by first deriving a reference RPF length distribution vector (P_{ref}) using all RPFs uniquely mapped to annotated CDSs. For each ORF (ncORF or CDS), a read length frequency vector (P_{ORF}) was computed, and FLOSS was calculated as $\sum|P_{ref} - P_{ORF}|/2$. The empirical upper bound for acceptable FLOSS was set using Tukey's method (i.e., the third quartile + 1.5 × inter-quartile range of FLOSS values across all CDSs)[39]. ncORFs with fewer than 10 P-site-mapped RPFs, RRS ≤ 1, or FLOSS values exceeding the empirical upper bound were filtered out.

Integration of ncORF prediction across different libraries of the same species was nontrivial due to extensive overlap among ncORFs in genomic coordinates. To address this issue, we used a graph-based clustering approach. We created a graph where two ORF vertices were connected if they share the same start or stop codon. Then, each set of connected ncORFs in the graph represents a cluster of overlapping ncORFs. To ensure reproducibility, only clusters identified by ≥2 methods in ≥2 libraries were retained, yielding 13,071 clusters in humans and 17,674 clusters in mice. To ensure compatibility with the latest reference gene annotations (ENSEMBL release 111), we removed any ncORFs that were newly classified as annotated CDSs. A manual inspection of RPF signals using the GWIPS-viz Ribo-Seq track[72] in the UCSC Genome Browser[73] revealed additional spurious ncORFs overlapping in-frame with CDSs from other transcripts. These were likely misidentified due to shared translation signals and were conservatively excluded using custom scripts, leaving 11,623 clusters in humans and 16,485 in mice.

To facilitate downstream analyses, a representative ncORF was selected for each cluster using a two-step strategy. First, a representative transcript was selected by prioritizing those with matched annotation from NCBI and EMBL-EBI (MANE)[42] (for humans) or APPRIS principal[74]; otherwise, the longest transcript was chosen. Then, among ncORFs within the representative transcript, a single ncORF was selected based on start codon hierarchy (AUG > CUG/GUG > UUG) following empirical evidence[28,36]. If multiple ncORFs shared the same preferred start codon, the longer one was chosen.

For validation, our ncORFs were compared with previously annotated GENCODE ncORFs[2] and those with MS evidence compiled previously[20]. Tier annotations of GENCODE ncORFs were obtained from a recent study[45]. Known functional ncORFs and their genomic coordinates were manually curated from published literature (Table S4). To account for possible proteoforms arising from alternative initiation and splicing, we considered an ncORF to be recovered if either its start or stop codon was identical to any ncORF in our annotation.

ncORF sequence analyses

The secondary structure surrounding each ORF (ncORF or CDS) start codon was predicted with RNAfold[75] on a 150 nt window extracted from the transcript in which the ncORF resides starting from the 30th nucleotide upstream of the start codon. For each position within this window, the overall pairing probability was calculated by dividing the number of ORFs with a base-paired residue at that position by the total number of ORFs analyzed. The tRNA adaptation index and effective number of codons for ncORFs and CDSs were calculated using the "cubar" R package[76]. tRNA gene annotations in humans and mice were obtained from GtRNAdb (release 22)[77].

Putative ncEPs analyses

Known domains within ncEPs were identified using InterProScan (v5.64-96.0)[78] with default parameters. Repetitive sequence annotations were downloaded from the UCSC Table Browser. BEDTools (v2.31.1)[79] was used to identify TE-overlapping ncORFs. IDRs were predicted with

IUPred3[80] with analysis type set to "short disorders" and smoothing disabled. For each ORF, residues with disorder tendency score > 0.5 were considered disordered, and the IDR fraction was calculated as the proportion of disordered residues relative to the ORF length.

Gnocchi Score and PhyloP of ncORFs and CDSs

Genome-wide Gnocchi scores computed using 1 kb sliding windows (with 100bp step size) were downloaded from the gnomAD database[50]. For each base pair, the Gnocchi score was defined as the average across the ten overlapping 1kb windows that covered that position using BEDTools. The final Gnocchi score for each ORF (ncORF or CDS) was obtained by averaging the scores of all base pairs within the ORF.

Base-wise PhyloP scores computed across 441 mammals or 239 primates[51] were downloaded from the UCSC Genome Browser. For each ncORF or CDS, PhyloP values were extracted for a ±30 bp window surrounding the start and stop codons using custom scripts. These values were then averaged over different ORFs to generate metagene profiles for visualization. To evaluate three-nucleotide periodicity, the metagene profiles were subjected to autocorrelation analysis using the "ac.test" function from the "testcorr" R package[81]. Statistical significance was assessed using the robust $\tilde{t}$ statistics.

ncORF origination node and mode

Whole genome alignments and corresponding species trees were downloaded from the UCSC Genome Browser (GRCh38 100-way for humans and GRCm39 35-way for mice). The origination node and mode for each ncORF were inferred following previous approaches[33,34]. For each ORF, the sequence from the reference genome and its orthologous regions in other species were extracted. Orthologous sequences covering less than 50% of the ncORF length relative to the reference species were excluded. The remaining sequences were realigned with MAFFT (v7.520)[82] with default parameters. The corresponding gene tree was derived by pruning the species tree using Gotree (v0.4.3)[83] with default settings. Ancestral sequence reconstruction was performed using FastML (v3.11)[84] with parameters "--seqType nuc --SubMatrix HKY --OptimizeBL no --indelReconstruction ML". An ORF structure was considered intact in an orthologous or ancestral sequence if one of the first three codons was a start codon and at least 70% of the 5' region of the reference ncORF was free of in-frame stop codons. The origination node was defined as the most ancient ancestral node that retained an intact ORF structure. An ncORF was classified as *de novo* if the origination node had an ancestor covering at least 50% of the ncORF sequence but lacking an intact ncORF structure. Otherwise, the origination mode was classified as uncertain. The distance from an ancestral node to the reference species (tip node) was calculated by summing branch lengths along the connecting path using the ape R package (v5.8)[85].

ncORF BLS and PhyloCSF

For each ncORF, global BLS was calculated as the ratio of the total branch length of the subtree connecting all leaves with intact ORF structures to the total branch length of the full species tree as in our previous study[48] with custom scripts. Local BLS was calculated similarly, but the denominator was replaced by the total branch length of the subtree defined by the origination node and all of its descendant nodes (including leaves). To evaluate coding potential, the alignment of all species descending from the origination node was converted into Multiple Alignment Format after removing the stop codon. PhyloCSF scores were then computed using the "score-msa" module of PhyloCSF++[86].

ncORF expression analysis

To quantify translation levels of ncORFs and CDSs in Ribo-Seq libraries, we assigned uniquely mapped reads to corresponding P-site positions with the PSite package. The read count for each ORF was calculated as the total number of RPFs whose P-sites fell within the ORF. To reduce bias caused by RPF enrichment at ORF termini, reads mapping to start or stop codons were excluded. For CDSs, only the count of the representative transcript of each gene as defined above was used

for downstream analyses. For samples with multiple libraries, read counts were aggregated prior to normalization. Sample-specific size factors were estimated using CDS counts and applied to normalize both ncORF and CDS counts across samples using DESeq2[87]. RPKM for an ORF was calculated as normalized read counts / ORF length (nt) / library size × 10^9, where library size was the median of total read counts across samples to ensure cross-sample comparability. An ORF (ncORF or CDS) was considered translated if its RPKM was ≥ 1. To evaluate tissue specificity of ncORF translation, the normalized RPF count X_k of an ORF in sample k was log-transformed as $x_k = log(X_k + 1)$. Then, tissue specificity index[88] was calculated as $\tau = \sum_k^n (1 - \tilde{x}_k) / (n - 1)$, where $\tilde{x}_k = x_k / \max_{1 \le j \le n} x_j$ and n is the total number of samples.

To quantify RNA levels, raw reads of deep transcriptome profiling across different human or mouse tissues from previous studies[55,56] were downloaded from SRA. Gene expression levels (transcript per million, TPM) were estimated using salmon (v1.10.3)[89] with parameters "--seqBias --gcBias –posBias -l A".

WGCNA analysis[58] of ncORF and CDS translation was performed using the PyWGCNA package (v2.0.5)[90] with default settings. Only ORFs with normalized RPF counts ≥ 10 in more than half of the samples were included. The soft threshold was determined automatically by PyWGCNA. Due to differences in soft thresholds, two ORFs were considered cotranslated if their topological overlap > 0.1 (human) or > 0.01 (mouse).[58] To test whether co-translation is enriched between ncORFs and CDSs from the same transcript, gene identities of CDSs were randomly permuted among all co-translated ncORF–CDS pairs. This procedure was repeated 1,000 times, and the empirical P value was calculated as (m + 1) / 1,000, where m is the number of permutations in which the number of same-gene pairs was greater than or equal to the observed count. The bipartite network of ncORF-CDS co-translation was clustered using the Leiden algorithm[91] implemented in igraph (v2.1.3)[92] by optimizing modularity with a resolution of 0.2. The top 20% of nodes by betweenness centrality were designated as hub nodes. Gene ontology analysis of hub genes in the largest modules was conducted using ClusterProfiler (v4.12)[93].

Data availability

Ribo-Seq libraries for ncORF annotation were obtained from the ENA or SRA, with accession codes listed in Table S1. Libraries used for ncORF and CDS translation quantification were listed in Table S7. RNA-Seq libraries for gene expression profiling across different human or mouse tissues were from previous studies[55,56] and are available with ENA accession numbers E-MTAB-2836 and E-MTAB-10276, respectively. Whole genome alignment across vertebrates, annotation of genomic repetitive elements, and base-wise PhyloP in primates and mammals[51] were downloaded from the UCSC Genome Browser[73]. Gnocchi scores were downloaded from gnomAD[50]. All the other data were presented in the manuscript or supplementary tables and are available from the corresponding author upon reasonable request.

Acknowledgements

We thank Dr. Chuan Yan and all the Zhang lab members for helpful discussion. This study was supported by the National Natural Science Foundation of China (32200433) and the Fundamental Research Funds for the Central Universities (LZUJBKY-2022-2). We thank the Supercomputing Center of Lanzhou University for providing computational resources.

Additional information

Code availability

All analyses were primarily conducted using R software (v4.3). Unless otherwise stated, statistical tests are two-sided. The code is available at *https://github.com/gxelab/ncorf_mammals* and custom scripts can be found at *https://github.com/gxelab/scripts*.

Funding

Funder	Grant reference number	Author
MOST | National Natural Science Foundation of China (NSFC)	32200433	Hong Zhang
MOE | Fundamental Research Funds for the Central Universities (Fundamental Research Fund for the Central Universities)	LZUJBKY-2022-2	Hong Zhang

Author ORCID iDs

Hong Zhang: https://orcid.org/0000-0002-4064-9432

Additional files

Supplementary Information

Supplementary Tables

References

1. **Wright B.W.,** Yi Z., Weissman J.S., Chen J. (2021) The dark proteome: translation from noncanonical open reading frames. *Trends in Cell Biology* https://doi.org/10.1016/j.tcb.2021.10.010 | PubMed
2. **Mudge J.M.,** Ruiz-Orera J., Prensner J.R., Brunet M.A., Calvet F., Jungreis I., Gonzalez J.M., Magrane M., Martinez T.F., Schulz J.F., *et al.* (2022) Standardized annotation of translated open reading frames. *Nature Biotechnology* **40**:994-999 https://doi.org/10.1038/s41587-022-01369-0 | PubMed
3. **Azam S.,** Yang F., Wu X. (2025) Finding functional microproteins. *Trends Genet* **41**:107-118 https://doi.org/10.1016/j.tig.2024.12.001 | PubMed
4. **Baena-Angulo C.,** Platero A.I., Couso J.P. (2025) Cis to trans: small ORF functions emerging through evolution. *Trends Genet* **41**:119-131 https://doi.org/10.1016/j.tig.2024.10.012 | PubMed
5. **Hofman D.A.,** Prensner J.R., van Heesch S. (2025) Microproteins in cancer: identification, biological functions, and clinical implications. *Trends Genet* **41**:146-161 https://doi.org/10.1016/j.tig.2024.09.002 | PubMed
6. **Orr M.W.,** Mao Y., Storz G., Qian S.B. (2020) Alternative ORFs and small ORFs: shedding light on the dark proteome. *Nucleic Acids Research* **48**:1029-1042 https://doi.org/10.1093/nar/gkz734 | PubMed
7. **Lu S.,** Zhang J., Lian X., Sun L., Meng K., Chen Y., Sun Z., Yin X., Li Y., Zhao J., *et al.* (2019) A hidden human proteome encoded by 'non-coding' genes. *Nucleic Acids Res* **47**:8111-8125 https://doi.org/10.1093/nar/gkz646 | PubMed
8. **Ingolia N.T.,** Ghaemmaghami S., Newman J.R., Weissman J.S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**:218-223 https://doi.org/10.1126/science.1168978 | PubMed
9. **Mao Y.,** Jia L., Dong L., Shu X.E., Qian S.B. (2023) Start codon-associated ribosomal frameshifting mediates nutrient stress adaptation. *Nature Structural & Molecular Biology* **30**:1816-1825 https://doi.org/10.1038/s41594-023-01119-z | PubMed
10. **Ferguson L.,** Upton H.E., Pimentel S.C., Mok A., Lareau L.F., Collins K., Ingolia N.T. (2023) Streamlined and sensitive mono- and di-ribosome profiling in yeast and human cells. *Nature Methods* **20**:1704-1715 https://doi.org/10.1038/s41592-023-02028-1 | PubMed
11. **Wang Y.,** Zhang H., Lu J. (2020) Recent advances in ribosome profiling for deciphering translational regulation. *Methods* **176**:46-54 https://doi.org/10.1016/j.ymeth.2019.05.011 | PubMed
12. **Cao X.,** Sun S., Xing J. (2024) A Massive Proteogenomic Screen Identifies Thousands of Novel Peptides From the Human "Dark" Proteome. *Molecular & Cellular Proteomics* **23**:100719 https://doi.org/10.1016/j.mcpro.2024.100719 | PubMed

13. **Ouspenskaia T.,** Law T., Clauser K.R., Klaeger S., Sarkizova S., Aguet F., Li B., Christian E., Knisbacher B.A., Le P.M., *et al.* (2021) Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. *Nature Biotechnology* **40**:209-217 https://doi.org/10.1038/s41587-021-01021-3 | PubMed

14. **Wu Q.,** Wright M., Gogol M.M., Bradford W.D., Zhang N., Bazzini A.A. (2020) Translation of small downstream ORFs enhances translation of canonical main open reading frames. *EMBO Journal* **39**:e104763 https://doi.org/10.15252/embj.2020104763 | PubMed

15. **Li Y.,** Zhou H., Chen X., Zheng Y., Kang Q., Hao D., Zhang L., Song T., Luo H., Hao Y., *et al.* (2021) SmProt: A Reliable Repository with Comprehensive Annotation of Small Proteins Identified from Ribosome Profiling. *Genomics, Proteomics & Bioinformatics* **19**:602-610 https://doi.org/10.1016/j.gpb.2021.09.002 | PubMed

16. **Leblanc S.,** Yala F., Provencher N., Lucier J.F., Levesque M., Lapointe X., Jacques J.F., Fournier I., Salzet M., Ouangraoua A., *et al.* (2024) OpenProt 2.0 builds a path to the functional characterization of alternative proteins. *Nucleic Acids Research* **52**:D522-d528 https://doi.org/10.1093/nar/gkad1050 | PubMed

17. **Yang H.,** Li Q., Stroup E.K., Wang S., Ji Z. (2024) Widespread stable noncanonical peptides identified by integrated analyses of ribosome profiling and ORF features. *Nat Commun* **15**:1932 https://doi.org/10.1038/s41467-024-46240-9 | PubMed

18. **Prensner J.R.,** Abelin J.G., Kok L.W., Clauser K.R., Mudge J.M., Ruiz-Orera J., Bassani-Sternberg M., Moritz R.L., Deutsch E.W., van Heesch S. (2023) What Can Ribo-Seq, Immunopeptidomics, and Proteomics Tell Us About the Noncanonical Proteome?. *Molecular & Cellular Proteomics* **22**:100631 https://doi.org/10.1016/j.mcpro.2023.100631 | PubMed

19. **Valdivia-Francia F.,** Sendoel A. (2024) No country for old methods: New tools for studying microproteins. *iScience* **27**:108972 https://doi.org/10.1016/j.isci.2024.108972 | PubMed

20. **Lei T.,** Chang Y., Yao C., Zhang H. (2023) A systematic evaluation of computational methods for predicting translated non-canonical ORFs from ribosome profiling data. *Journal of Genetics and Genomics* **51**:105-108 https://doi.org/10.1016/j.jgg.2023.08.010 | PubMed

21. **Olexiouk V.,** Crappé J., Verbruggen S., Verhegen K., Martens L., Menschaert G. (2016) sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Research* **44**:D324-D329 https://doi.org/10.1093/nar/gkv1175 | PubMed

22. **Chothani S.P.,** Adami E., Widjaja A.A., Langley S.R., Viswanathan S., Pua C.J., Zhihao N.T., Harmston N., D'Agostino G., Whiffin N., *et al.* (2022) A high-resolution map of human RNA translation. *Molecular Cell* **82**:2885-2899.e2888 https://doi.org/10.1016/j.molcel.2022.06.023 | PubMed

23. **Prensner J.R.,** Enache O.M., Luria V., Krug K., Clauser K.R., Dempster J.M., Karger A., Wang L., Stumbraite K., Wang V.M., *et al.* (2021) Noncanonical open reading frames encode functional proteins essential for cancer cell survival. *Nature Biotechnology* **39**:697-704 https://doi.org/10.1038/s41587-020-00806-2 | PubMed

24. **Camarena M.E.,** Theunissen P., Ruiz M., Ruiz-Orera J., Calvo-Serra B., Castelo R., Castro C., Sarobe P., Fortes P., Perera-Bel J., *et al.* (2024) Microproteins encoded by noncanonical ORFs are a major source of tumor-specific antigens in a liver cancer patient meta-cohort. *Sci Adv* **10**:eadn3628 https://doi.org/10.1126/sciadv.adn3628 | PubMed

25. **Sendoel A.,** Dunn J.G., Rodriguez E.H., Naik S., Gomez N.C., Hurwitz B., Levorse J., Dill B.D., Schramek D., Molina H., *et al.* (2017) Translation from unconventional 5' start sites drives tumour initiation. *Nature* **541**:494-499 https://doi.org/10.1038/nature21036 | PubMed

26. **Robichaud N.,** Sonenberg N., Ruggero D., Schneider R.J. (2019) Translational Control in Cancer. *Cold Spring Harb Perspect Biol* **11** https://doi.org/10.1101/cshperspect.a032896 | PubMed

27. **Malekos E.,** Carpenter S. (2022) Short open reading frame genes in innate immunity: from discovery to characterization. *Trends Immunol* **43**:741-756 https://doi.org/10.1016/j.it.2022.07.005 | PubMed

28. **Chen J.,** Brunner A.D., Cogan J.Z., Nunez J.K., Fields A.P., Adamson B., Itzhak D.N., Li J.Y., Mann M., Leonetti M.D., *et al.* (2020) Pervasive functional translation of noncanonical human open reading frames. *Science* **367**:1140-1146 https://doi.org/10.1126/science.aay0262 | PubMed

29. **Zheng C.,** Wei Y., Zhang P., Lin K., He D., Teng H., Manyam G., Zhang Z., Liu W., Lee H.R.L., *et al.* (2023) CRISPR-Cas9-based functional interrogation of unconventional translatome reveals human cancer dependency on cryptic non-canonical open reading frames. *Nature Structural & Molecular Biology* **30**:1878-1892 https://doi.org/10.1038/s41594-023-01117-1 | PubMed

30. **Zheng C.,** Wei Y., Zhang P., Xu L., Zhang Z., Lin K., Hou J., Lv X., Ding Y., Chiu Y., *et al.* (2023) CRISPR/Cas9 screen uncovers functional translation of cryptic lncRNA-encoded open reading frames in human cancer. *J Clin Invest* **133** https://doi.org/10.1172/jci159940 | PubMed

31. **Schlesinger D.,** Dirks C., Luzon C.N., Lafranchi L., Eirich J., Elsässer S.J. (2023) A large-scale sORF screen identifies putative microproteins and provides insights into their interaction partners, localisation and function. *bioRxiv* https://doi.org/10.1101/2023.06.13.544808

32. **Hofman D.A.,** Ruiz-Orera J., Yannuzzi I., Murugesan R., Brown A., Clauser K.R., Condurat A.L., van Dinter J.T., Engels S.A.G., Goodale A., *et al.* (2024) Translation of non-canonical open reading frames as a cancer cell survival mechanism in childhood medulloblastoma. *Molecular Cell* **84**:261-276.e218 https://doi.org/10.1016/j.molcel.2023.12.003 | PubMed

33. **Vakirlis N.,** Vance Z., Duggan K.M., McLysaght A. (2022) De novo birth of functional microproteins in the human lineage. *Cell Rep* **41**:111808 https://doi.org/10.1016/j.celrep.2022.111808 | PubMed

34. **Sandmann C.L.,** Schulz J.F., Ruiz-Orera J., Kirchner M., Ziehm M., Adami E., Marczenke M., Christ A., Liebe N., Greiner J., *et al.* (2023) Evolutionary origins and interactomes of human, young microproteins and small peptides translated from short open reading frames. *Mol Cell* **83**:994-1011.e1018 https://doi.org/10.1016/j.molcel.2023.01.023 | PubMed

35. **Wacholder A.,** Parikh S.B., Coelho N.C., Acar O., Houghton C., Chou L., Carvunis A.R. (2023) A vast evolutionarily transient translatome contributes to phenotype and fitness. *Cell Syst* https://doi.org/10.1016/j.cels.2023.04.002 | PubMed

36. **Erhard F.,** Halenius A., Zimmermann C., L'Hernault A., Kowalewski D.J., Weekes M.P., Stevanovic S., Zimmer R., Dolken L. (2018) Improved Ribo-seq enables identification of cryptic translation events. *Nat Methods* **15**:363-366 https://doi.org/10.1038/nmeth.4631 | PubMed

37. **Xiao Z.,** Huang R., Xing X., Chen Y., Deng H., Yang X. (2018) De novo annotation and characterization of the translatome with ribosome profiling data. *Nucleic Acids Res* **46**:e61 https://doi.org/10.1093/nar/gky179 | PubMed

38. **Zhang P.,** He D., Xu Y., Hou J., Pan B.F., Wang Y., Liu T., Davis C.M., Ehli E.A., Tan L., *et al.* (2017) Genome-wide identification and differential analysis of translational initiation. *Nat Commun* **8**:1749 https://doi.org/10.1038/s41467-017-01981-8 | PubMed

39. **Ingolia N.T.,** Brar G.A., Stern-Ginossar N., Harris M.S., Talhouarne G.J., Jackson S.E., Wills M.R., Weissman J.S. (2014) Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep* **8**:1365-1379 https://doi.org/10.1016/j.celrep.2014.07.045 | PubMed

40. **Chew G.L.,** Pauli A., Rinn J.L., Regev A., Schier A.F., Valen E. (2013) Ribosome profiling reveals resemblance between long non-coding RNAs and 5′ leaders of coding RNAs. *Development* **140**:2828-2834 https://doi.org/10.1242/dev.098343 | PubMed

41. **Guttman M.,** Russell P., Ingolia N.T., Weissman J.S., Lander E.S. (2013) Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* **154**:240-251 https://doi.org/10.1016/j.cell.2013.06.009 | PubMed

42. **Morales J.,** Pujar S., Loveland J.E., Astashyn A., Bennett R., Berry A., Cox E., Davidson C., Ermolaeva O., Farrell C.M., *et al.* (2022) A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* **604**:310-315 https://doi.org/10.1038/s41586-022-04558-8 | PubMed

43. **Pozo F.,** Rodriguez J.M., Martínez Gómez L., Vázquez J., Tress M.L. (2022) APPRIS principal isoforms and MANE Select transcripts define reference splice variants. *Bioinformatics* **38**:ii89-ii94 https://doi.org/10.1093/bioinformatics/btac473 | PubMed

44. **Chothani S.,** Ruiz-Orera J., Tierney J.A.S., Clauwaert J., Deutsch E.W., Alba M.M., Aspden J.L., Baranov P.V., Bazzini A.A., Bruford E.A., *et al.* (2025) An expanded reference catalog of translated open reading frames for biomedical research. *bioRxiv* https://doi.org/10.1101/2025.07.03.662928 | PubMed

45. **Deutsch E.W.,** Kok L.W., Mudge J.M., Ruiz-Orera J., Fierro-Monti I., Sun Z., Abelin J.G., Alba M.M., Aspden J.L., Bazzini A.A., *et al.* (2024) High-quality peptide evidence for annotating non-canonical open reading frames as human proteins. *bioRxiv* https://doi.org/10.1101/2024.09.09.612016 | PubMed

46. **Shabalina S.A.,** Spiridonov N.A., Kashina A. (2013) Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. *Nucleic Acids Res* **41**:2073-2094 https://doi.org/10.1093/nar/gks1205 | PubMed

47. **Hernández G.,** Osnaya V.G., Pérez-Martínez X. (2019) Conservation and Variability of the AUG Initiation Codon Context in Eukaryotes. *Trends Biochem Sci* **44**:1009-1021 https://doi.org/10.1016/j.tibs.2019.07.001 | PubMed

48. **Zhang H.,** Wang Y., Wu X., Tang X., Wu C., Lu J. (2021) Determinants of genome-wide distribution and evolution of uORFs in eukaryotes. *Nature Communications* **12**:1076 https://doi.org/10.1038/s41467-021-21394-y | PubMed

49. **Wells J.N.,** Feschotte C. (2020) A Field Guide to Eukaryotic Transposable Elements. *Annu Rev Genet* **54**:539-561 https://doi.org/10.1146/annurev-genet-040620-022145 | PubMed

50. **Chen S.,** Francioli L.C., Goodrich J.K., Collins R.L., Kanai M., Wang Q., Alföldi J., Watts N.A., Vittal C., Gauthier L.D., *et al.* (2024) A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**:92-100 https://doi.org/10.1038/s41586-023-06045-0 | PubMed

51. **Kuderna L.F.K.,** Ulirsch J.C., Rashid S., Ameen M., Sundaram L., Hickey G., Cox A.J., Gao H., Kumar A., Aguet F., *et al.* (2024) Identification of constrained sequence elements across 239 primate genomes. *Nature* **625**:735-742 https://doi.org/10.1038/s41586-023-06798-8 | PubMed

52. **Pollard K.S.,** Hubisz M.J., Rosenbloom K.R., Siepel A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**:110-121 https://doi.org/10.1101/gr.097857.109 | PubMed

53. **Lin M.F.,** Jungreis I., Kellis M. (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**:i275-282 https://doi.org/10.1093/bioinformatics/btr209 | PubMed

54. **Stark A.,** Lin M.F., Kheradpour P., Pedersen J.S., Parts L., Carlson J.W., Crosby M.A., Rasmussen M.D., Roy S., Deoras A.N., *et al.* (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature* **450**:219-232 https://doi.org/10.1038/nature06340 | PubMed

55. **Wang D.,** Eraslan B., Wieland T., Hallstrom B., Hopf T., Zolg D.P., Zecha J., Asplund A., Li L.H., Meng C., *et al.* (2019) A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol Syst Biol* **15**:e8503 https://doi.org/10.15252/msb.20188503 | PubMed

56. **Giansanti P.,** Samaras P., Bian Y., Meng C., Coluccio A., Frejno M., Jakubowsky H., Dobiasch S., Hazarika R.R., Rechenberger J., *et al.* (2022) Mass spectrometry-based draft of the mouse proteome. *Nat Methods* **19**:803-811 https://doi.org/10.1038/s41592-022-01526-y | PubMed

57. **Phillips P.C.** (2008) Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet* **9**:855-867 https://doi.org/10.1038/nrg2452 | PubMed

58. **Langfelder P.,** Horvath S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**:559 https://doi.org/10.1186/1471-2105-9-559 | PubMed

59. **Zhang J.,** Xu C. (2022) Gene product diversity: adaptive or not?. *Trends Genet* **38**:1112-1122 https://doi.org/10.1016/j.tig.2022.05.002 | PubMed

60. **Mao Y.,** Qian S.B. (2024) Making sense of mRNA translational “noise”. *Semin Cell Dev Biol* **154**:114-122 https://doi.org/10.1016/j.semcdb.2023.03.004 | PubMed

61. **Broeils L.A.,** Ruiz-Orera J., Snel B., Hubner N., van Heesch S. (2023) Evolution and implications of de novo genes in humans. *Nat Ecol Evol* https://doi.org/10.1038/s41559-023-02014-y | PubMed

62. **Liu X.,** Xiao C., Xu X., Zhang J., Mo F., Chen J.Y., Delihas N., Zhang L., An N.A., Li C.Y. (2024) Origin of functional de novo genes in humans from “hopeful monsters”. *Wiley Interdiscip Rev RNA* **15**:e1845 https://doi.org/10.1002/wrna.1845 | PubMed

63. **Mackowiak S.D.,** Zauber H., Bielow C., Thiel D., Kutz K., Calviello L., Mastrobuoni G., Rajewsky N., Kempa S., Selbach M., *et al.* (2015) Extensive identification and analysis of conserved small ORFs in animals. *Genome Biology* **16**:179 https://doi.org/10.1186/s13059-015-0742-x | PubMed

64. **Chew G.L.,** Pauli A., Schier A.F. (2016) Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish. *Nat Commun* **7**:11663 https://doi.org/10.1038/ncomms11663 | PubMed

65. **Zhang H.,** Dou S., He F., Luo J., Wei L., Lu J. (2018) Genome-wide maps of ribosomal occupancy provide insights into adaptive evolution and regulatory roles of uORFs during Drosophila development. *PLoS Biology* **16**:e2003903 https://doi.org/10.1371/journal.pbio.2003903 | PubMed

66. **Sun Y.,** Duan Y., Gao P., Liu C., Jin K., Dou S., Tang W., Zhang H., Lu J. (2025) Upstream open reading frames buffer translational variability during Drosophila evolution and development. *eLife* **14** https://doi.org/10.7554/eLife.104074 | PubMed

67. **Treichel A.J.,** Bazzini A.A. (2022) Casting CRISPR-Cas13d to fish for microprotein functions in animal development. *iScience* **25**:105547 https://doi.org/10.1016/j.isci.2022.105547 | PubMed

68. **Martin M.** (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal* **17**:10-12 https://doi.org/10.14806/ej.17.1.200

69. **Cunningham F.,** Allen J.E., Allen J., Alvarez-Jarreta J., Amode M.R., Armean I.M., Austine-Orimoloye O., Azov A.G., Barnes I., Bennett R., *et al.* (2022) Ensembl 2022. *Nucleic Acids Res* **50**:D988-d995 https://doi.org/10.1093/nar/gkab1049 | PubMed

70. **Dobin A.,** Davis C.A., Schlesinger F., Drenkow J., Zaleski C., Jha S., Batut P., Chaisson M., Gingeras T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**:15-21 https://doi.org/10.1093/bioinformatics/bts635 | PubMed

71. **Chang Y.,** Lei T., Zhang H. (2023) PSite: inference of read-specific P-site offsets for ribosomal footprints. *bioRxiv* https://doi.org/10.1101/2023.06.27.546788

72. **Tierney J.A.S.,** Świrski M.I., Tjeldnes H., Kiran A.M., Carancini G., Kiniry S.J., Michel A.M., Kufel J., Valen E., Baranov P.V. (2025) RiboSeq.Org: an integrated suite of resources for ribosome profiling data analysis and visualization. *Nucleic Acids Res* **53**:D268-d274 https://doi.org/10.1093/nar/gkae1020 | PubMed

73. **Nassar L.R.,** Barber G.P., Benet-Pagès A., Casper J., Clawson H., Diekhans M., Fischer C., Gonzalez J.N., Hinrichs A.S., Lee B.T., *et al.* (2023) The UCSC Genome Browser database: 2023 update. *Nucleic Acids Res* **51**:D1188-d1195 https://doi.org/10.1093/nar/gkac1072 | PubMed

74. **Rodriguez J.M.,** Pozo F., Cerdán-Vélez D., Di Domenico T., Vázquez J., Tress M.L. (2022) APPRIS: selecting functionally important isoforms. *Nucleic Acids Res* **50**:D54-d59 https://doi.org/10.1093/nar/gkab1058 | PubMed

75. **Lorenz R.,** Bernhart S.H., Höner Zu Siederdissen C., Tafer H., Flamm C., Stadler P.F., Hofacker I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**:26 https://doi.org/10.1186/1748-7188-6-26 | PubMed

76. **Liu M.,** Zi B., Zhang H., Zhang H. (2024) cubar: a versatile package for codon usage bias analysis in R. R package. version: version 1.1 https://doi.org/10.5281/zenodo.10155990

77. **Chan P.P.,** Lowe T.M. (2016) GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res* **44**:D184-189 https://doi.org/10.1093/nar/gkv1309 | PubMed

78. **Jones P.,** Binns D., Chang H.Y., Fraser M., Li W., McAnulla C., McWilliam H., Maslen J., Mitchell A., Nuka G., *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**:1236-1240 https://doi.org/10.1093/bioinformatics/btu031 | PubMed

79. **Quinlan A.R.,** Hall I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**:841-842 https://doi.org/10.1093/bioinformatics/btq033 | PubMed

80. **Erdős G.,** Pajkos M., Dosztányi Z. (2021) IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Res* **49**:W297-w303 https://doi.org/10.1093/nar/gkab408 | PubMed

81. **Dalla V.,** Giraitis L., Phillips P.C.B. (2022) obust Tests for White Noise and Cross-Correlation. *Econometric Theory* **38**:913-941 https://doi.org/10.1017/S0266466620000341

82. **Katoh K.,** Standley D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**:772-780 https://doi.org/10.1093/molbev/mst010 | PubMed

83. **Lemoine F.,** Gascuel O. (2021) Gotree/Goalign: toolkit and Go API to facilitate the development of phylogenetic workflows. *NAR Genom Bioinform* **3**:lqab075 https://doi.org/10.1093/nargab/lqab075 | PubMed

84. **Ashkenazy H.,** Penn O., Doron-Faigenboim A., Cohen O., Cannarozzi G., Zomer O., Pupko T. (2012) FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Res* **40**:W580-584 https://doi.org/10.1093/nar/gks498 | PubMed

85. **Paradis E.,** Schliep K. (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**:526-528 https://doi.org/10.1093/bioinformatics/bty633 | PubMed

86. **Pockrandt C.,** Steinegger M., Salzberg S.L. (2022) PhyloCSF++: a fast and user-friendly implementation of PhyloCSF with annotation tools. *Bioinformatics* **38**:1440-1442 https://doi.org/10.1093/bioinformatics/btab756 | PubMed

87. **Love M.I.,** Huber W., Anders S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**:550 https://doi.org/10.1186/s13059-014-0550-8 | PubMed

88. **Kryuchkova-Mostacci N.,** Robinson-Rechavi M. (2017) A benchmark of gene expression tissue-specificity metrics. *Brief Bioinform* **18**:205-214 https://doi.org/10.1093/bib/bbw008 | PubMed

89. **Patro R.,** Duggal G., Love M.I., Irizarry R.A., Kingsford C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**:417-419 https://doi.org/10.1038/nmeth.4197 | PubMed

90. **Rezaie N.,** Reese F., Mortazavi A. (2023) PyWGCNA: a Python package for weighted gene co-expression network analysis. *Bioinformatics* **39** https://doi.org/10.1093/bioinformatics/btad415 | PubMed

91. **Traag V.A.,** Waltman L., van Eck N.J. (2019) From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* **9**:5233 https://doi.org/10.1038/s41598-019-41695-z | PubMed

92. **Csárdi G.,** Nepusz T., Traag V., Horvát S., Zanini F., Noom D., Müller K. (2024) igraph: Network Analysis and Visualization in R. R package. version: version 2.1.3 https://doi.org/10.5281/zenodo.7682609

93. **Wu T.,** Hu E., Xu S., Chen M., Guo P., Dai Z., Feng T., Zhou L., Tang W., Zhan L., *et al.* (2021) clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation* **2**:100141 https://doi.org/10.1016/j.xinn.2021.100141 | PubMed

Wang D, Eraslan B, Wieland T, Hallström B, Hopf T, Zolg DP, Zecha J, Asplund A, Li LH, Meng C, *et al.* (2019) A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. ArrayExpress. ID E-MTAB-2836 https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-2836

Giansanti P, Samaras P, Bian Y, Meng C, Coluccio A, Frejno M, Jakubowsky H, Dobiasch S, Hazarika RR, Rechenberger J, *et al.* (2022) Mass spectrometry-based draft of the mouse proteome. ArrayExpress. ID E-MTAB-10276 https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-10276

Kuderna LFK, Ulirsch JC, Rashid S, Ameen M, Sundaram L, Hickey G, Cox AJ, Gao H, Kumar A, Aguet F, *et al.* (2024) Identification of constrained sequence elements across 239 primate genomes. UCSC Genome Browser. ID phyloP447way https://hgdownload.soe.ucsc.edu/goldenPath/hg38/phyloP447way/

Peer reviews

Reviewer #1 (Public review):

This work compiles a comprehensive atlas of ncORFs across mammalian tissues and cell types, derived from reanalysis of ~400 public ribosome profiling datasets. The authors then evaluate cross-species conservation and functional signatures, proposing that evolutionarily ancient ncORFs tend to have higher translation potential, stronger expression, and closer relationships with canonical coding sequences.

Strengths:

In general, the study provides a large-scale and timely resource of annotated ncORFs, which could be broadly useful for the community. The authors collected ~400 public ribosome profiling datasets for annotations of ncORFs, which, to my best knowledge, is the largest collection of data for such purpose. The catalog could facilitate future investigations into ncORF biology and broaden understanding of the coding potential of the "non-coding" genome.

Weaknesses:

Based on the ncORF catalog, some of the analyses were not properly done. Some of the results are descriptive.

(1) Bias and representations of data source. Public ribo-seq datasets are unevenly distributed across tissues and cell lines, raising concerns about heterogeneity and underrepresentation of certain contexts. This may limit the generalizability of the catalog.

(2) The discussion on modular domains of ncORFs is unclear, and the claim that they may originate via TE-related mechanisms is not well supported. Stronger evidence or clearer reasoning is needed.

(3) The conservation comparisons are not fully convincing. Figure S7 shows only mild differences between ncORFs and CDS, and statistical significance is not clearly demonstrated. Comparisons with other non-coding RNAs should be added, and overlapping sequences between ncORFs and CDS should be excluded to avoid bias.

(4) Figure 3 indicates that some ncORFs are subject to evolutionary constraints. This is not surprising. The authors should provide further analyses on more detailed features of these "conserved" ncORFs vs. the "non-conserved" ones. Some pretty informative works have been done in drosophila, worms, mouse, and human. Figure 3 suggests some ncORFs are under evolutionary constraint, but this is not unexpected. More granular analyses contrasting "conserved" versus "non-conserved" ncORFs would be informative. In fact, small ORFs, especially uORFs, have been extensively studied, for their functions and corss-species conservations. The authors should explicitly show what is new here in their analyses.

(5) Translation levels are reported using RPF counts. However, translation efficiency (normalized by RNA expression) is a more appropriate measure to account for expression heterogeneity.

(6) The correlation analyses between ncORF translation levels and PhyloCSF are confusing and largely descriptive. These sections need sharper framing and clearer conclusions.

(7) Public ribo-seq datasets, generated by different research labs, are known for their strong batch effects. Representations of tissues and cells are also very unbalanced. Therefore, the co-translation analysis between ncORFs and canonical CDS is not well controlled. This should be

done by referring to a recent large-scale ribo-seq meta-analysis (Nat Biotechnol. 2025. doi: 10.1038/s41587-025-02718-5).

Comments on revisions:

The authors have made efforts to address most of the previous concerns, and several points have been clarified or improved in the revision. However, in a number of cases, the responses rely more on acknowledgment and reframing rather than substantive analytical strengthening. Overall, the manuscript is improved, particularly in terms of clarity, transparency, and positioning of claims. I support its publication and look forward to seeing how the field engages with and discusses these claims.

https://doi.org/10.7554/eLife.109128.2.sa2

Reviewer #2 (Public review):

Summary:

Chang et al. attempted to analyze a large number of ribo-seq datasets through a standardized pipeline, identifying novel non-canonical ORFs and elucidating their evolutionary and expression characteristics.

Strengths:

(1) The datasets analyzed by the authors are sufficiently comprehensive, and the use of standardized pipelines ensures excellent analytical consistency.

(2) Their analyses of ORF evolution and co-expression further deepen our understanding of these ORFs.

Weaknesses:

(1) The authors primarily conducted analyses through bioinformatics, lacking sufficient wet-lab experimental evidence.

(2) Some analytical methods and standards were not clearly presented in the manuscript.

https://doi.org/10.7554/eLife.109128.2.sa1

Author response:

The following is the authors' response to the original reviews.

> ***Public Reviews:***
>
> ***Reviewer #1 (Public review):***
>
> *This work compiles a comprehensive atlas of ncORFs across mammalian tissues and cell types, derived from reanalysis of ~400 public ribosome profiling datasets. The authors then evaluate cross-species conservation and functional signatures, proposing that evolutionarily ancient ncORFs tend to have higher translation potential, stronger expression, and closer relationships with canonical coding sequences.*
>
> *Strengths:*
>
> *In general, the study provides a large-scale and timely resource of annotated ncORFs, which could be broadly useful for the community. The authors collected ~400 public ribosome profiling datasets for annotations of ncORFs, which, to my best knowledge, is the largest collection of data for such a purpose. The catalog could facilitate future*

> *investigations into ncORF biology and broaden understanding of the coding potential of the "non-coding" genome.*

We thank the reviewer for the positive evaluation of our manuscript and for recognizing the significance of our contribution.

> *Weaknesses:*
>
> *Based on the ncORF catalog, some of the analyses were not properly done. Some of the results are descriptive.*
>
> *(1) Bias and representations of the data source. Public ribo-seq datasets are unevenly distributed across tissues and cell lines, raising concerns about heterogeneity and underrepresentation of certain contexts. This may limit the generalizability of the catalog.*

We agree with the reviewer that the uneven distribution of public Ribo-seq datasets across tissues can inevitably introduce bias in the ncORF composition of our catalog. This bias is likely more pronounced in humans due to the narrower tissue coverage. We have addressed this point in the Discussion section of the revised manuscript.

> *(2) The discussion on modular domains of ncORFs is unclear, and the claim that they may originate via TErelated mechanisms is not well supported. Stronger evidence or clearer reasoning is needed.*

We thank the reviewer for highlighting this point. We have revised the manuscript to more clearly explain the rationale behind our analysis of ncORF modular domains and have adopted more cautious language regarding their potential transposable element–related origins, limiting interpretations to what is directly supported by the data.

> *(3) The conservation comparisons are not fully convincing. Figure S7 shows only mild differences between ncORFs and CDS, and statistical significance is not clearly demonstrated.*
>
> *Comparisons with other non-coding RNAs should be added, and overlapping sequences between ncORFs and CDS should be excluded to avoid bias.*

We thank the reviewer for this comment and apologize for the lack of clarity in the original figure. Both CDSs and ncORFs show significant deviation from zero Gnocchi scores (two-sided Wilcoxon signed-rank tests), which is now stated explicitly in the revised legend and text. CDS-overlapping ncORFs were already excluded in the original analysis; this has been clarified to avoid confusion.

As suggested, we have added lncRNAs for comparison. ncORFs display modestly higher Gnocchi scores than lncRNAs, and this difference persists when restricting the analysis to lncRNA-derived ncORFs and their corresponding full-length lncRNAs (see revised Fig. S7). These additions strengthen the conservation comparison while controlling for transcript context.

> *(4) Figure 3 indicates that some ncORFs are subject to evolutionary constraints. This is not surprising. The authors should provide further analyses on more detailed features of these "conserved" ncORFs vs. the "non-conserved" ones. Some pretty informative works have been done in Drosophila, worms, mice, and humans. Figure 3 suggests some ncORFs are under evolutionary constraint, but this is not unexpected. More granular analyses contrasting "conserved" versus "non-conserved" ncORFs would be informative. In fact, small ORFs, especially uORFs, have been extensively studied for their functions and cross-species conservation. The authors should explicitly show what is new here in their analyses.*

We thank the reviewer for this insightful comment. We agree that cross-species conservation of ncORFs (particularly uORFs) has been extensively investigated in prior studies, including our own.

However, most prior analyses have focused on conservation of start codons or overall ORF integrity, which does not distinguish selection acting on translational activity from selection acting on the encoded peptide sequence itself. In contrast, our analysis leverages codon-level periodic PhyloP signals across the full ORF. The observed three-nucleotide periodicity is consistent with selective constraint at the amino acid level, rather than merely preservation of initiation sites or translational potential. Furthermore, our newly developed branch-length statistic uncovers lineage-restricted conservation patterns among ncORFs, enabling resolution of evolutionary dynamics not captured by conventional conservation metrics.

Thus, while the existence of conserved ncORFs is not unexpected, the conceptual advance of our study lies in demonstrating that a subset exhibits coding-like evolutionary constraint consistent with selection on their peptide products, as well as revealing lineage-specific conservation patterns. We have clarified this distinction in the revised Discussion.

> *(5) Translation levels are reported using RPF counts. However, translation efficiency (normalized by RNA expression) is a more appropriate measure to account for expression heterogeneity.*

We agree that translation efficiency (TE), which normalizes ribosome footprint counts by RNA abundance, is in principle an appropriate metric. We initially calculated TE and compared ncORFs with CDSs. However, we found that TE estimates for short ncORFs were substantially inflated by RPF enrichment near start and stop codons, leading to unstable and potentially misleading values.

For CDSs, this bias is commonly addressed by excluding the first and last 10 to 20 codons when quantifying RPF density. This strategy is not feasible for ncORFs because of their short length. We therefore used RPF counts in the final analysis, applying stringent positional filtering. Only RPFs whose P sites fall within the ORF body, excluding start and stop codons, were counted. RPFs overlapping the ORF but with P sites outside the annotated frame, likely derived from adjacent ORFs or initiation or termination pausing, were excluded.

TE and RPF counts both measure translation but capture different aspects. TE reflects ribosome density relative to transcript abundance, whereas RPF counts quantify overall ribosome engagement. Given the short lengths of ncORFs, count-based quantification provides a more robust and conservative estimate of their translational activity.

> *(6) The correlation analyses between ncORF translation levels and PhyloCSF are confusing and largely descriptive. These sections need sharper framing and clearer conclusions.*

We thank the reviewer for this comment. We agree that the original presentation lacked clear framing. The relationship between PhyloCSF scores and mean ncORF translation levels across

tissues is influenced by both evolutionary age and tissue specificity. Older ncORFs with higher coding potential tend to exhibit stronger tissue-restricted expression. As a result, their mean translation levels across all tissues appear lower, not because they are weakly translated, but because their translation is concentrated in specific tissues. This point is addressed in the revised manuscript.

> *(7) Public ribo-seq datasets, generated by different research labs, are known for their strong batch effects. Representations of tissues and cells are also very unbalanced. Therefore, the co-translation analysis between ncORFs and canonical CDS is not well controlled. This should be done by referring to a recent large-scale ribo-seq meta-analysis (Nat Biotechnol. 2025. doi: 10.1038/s41587-025-02718-5).*

We thank the reviewer for highlighting this important study and for raising concerns regarding batch effects and tissue imbalance in public Ribo-seq datasets. We are aware that public Ribo-seq data generated by different laboratories are subject to substantial batch effects. During the ncORF annotation phase, we applied stringent quality-control criteria to minimize technical variability. For the co-translation analysis, inclusion criteria were relaxed to increase tissue and cell-type coverage. To partially mitigate representation bias, libraries derived from the same tissue or cell type were merged when quantifying ORF translation levels, thereby reducing overrepresentation from heavily sampled contexts.

Nevertheless, we acknowledge that these measures cannot completely eliminate batch effects or imbalance inherent to public datasets. We agree that co-translation analysis would benefit from uniformly processed, high-quality datasets generated under standardized protocols with balanced tissue representation, representing a valuable direction for future research.

> ***Reviewer #2 (Public review):***
>
> *Summary:*
>
> *Chang et al. attempted to analyze a large number of ribo-seq datasets through a standardized pipeline, identifying novel non-canonical ORFs and elucidating their evolutionary and expression characteristics.*
>
> *Strengths:*
>
> *(1) The datasets analyzed by the authors are sufficiently comprehensive, and the use of standardized pipelines ensures excellent analytical consistency.*
>
> *(2) Their analyses of ORF evolution and co-expression further deepen our understanding of these ORFs.*

We thank the reviewer for the positive evaluation of our manuscript. It is encouraging to know that the analytical framework was found to be sound and appropriate.

> *Weaknesses:*
>
> *(1) The authors primarily conducted analyses through bioinformatics, lacking sufficient wet-lab experimental evidence.*

We thank the reviewer for this comment and acknowledge this limitation. We agree that functional validation through wet-lab experiments would provide important mechanistic insight into individual ncORFs. However, this study was designed as a systematic, genome-wide computational analysis to characterize translated ncORFs across species and tissues. Our objective was to define global patterns of translation, conservation, and structural features using large-scale datasets. Given the breadth and scale of these analyses, experimental validation of specific ncORFs falls beyond the scope of the current study. We

have clarified this point in the dicussion and noted that our results provide a framework for future targeted experimental investigation.

> *(2) Regarding the evolution of non-canonical ORFs, a considerable amount of prior work already exists. The authors need to further clarify what new insights and discoveries they have made based on the analysis of such a large dataset.*

We thank the reviewer for this suggestion. Similar concerns were also raised by Reviewer #1. In response, we have revised the Discussion to more clearly delineate the conceptual advances enabled by our large-scale dataset.

> ***Recommendations for the authors:***
>
> ***Reviewing Editor Comments:***
>
> *Several aspects of the downstream analyses would benefit from additional refinement. The heterogeneity and tissue imbalance inherent in public Ribo-seq datasets introduce potential biases in ncORF detection and inferences about co-translation. Given the breadth of the dataset, it would also be informative to quantify how consistently the newly identified ncORFs are detected across samples-distinguishing those observed broadly across tissues, those enriched in specific contexts, and those detected in only a few datasets. Such stratification would help differentiate reproducibly translated ORFs from candidates requiring further validation.*

We thank the editor for the helpful comments. We agree that heterogeneity and tissue imbalance in public Ribo-seq datasets can influence ncORF detection and downstream interpretations. We have added discussion of this limitation in the revised manuscript.

Detection of ncORF translation depends not only on biological activity but also on sequencing depth and data quality. Although all ncORFs reported here were reproducibly identified by multiple methods across independent libraries, we agree that those detected in a larger number of datasets represent stronger candidates for functional validation. Accordingly, we now report the number of methods and libraries in which each ncORF was detected in the final catalog (Supplementary Table 3). Overall, 22.3–26.3% of ncORFs were detected in more than 10 libraries, whereas more than half were observed in only two to five libraries (Fig. S1B), enabling clearer stratification of broadly translated versus more context-specific candidates.

> *Some evolutionary and functional interpretations are largely descriptive or consistent with established findings for small ORFs, and the authors should more clearly articulate what is novel in their analyses. The criteria separating "young," "old," and "ancient" ORFs require clearer definition, and conservation analyses would be strengthened by improved statistical rigor and explicit exclusion of regions overlapping annotated coding sequences. Evidence for modular domain features or transposable element-related origins is limited and warrants either stronger support or more cautious framing. Proteomics validation is currently minimal and could be substantially reinforced using existing public MS resources.*

We thank the reviewer for these constructive comments. In the revised manuscript, we more clearly delineate the novel insights derived from our evolutionary analyses of ncORFs, distinguishing them from established findings on small ORFs.

We have clarified the criteria used to classify ORFs by evolutionary age in figure 6E and refined the terminology describing “young,” “old,” and “ancient” categories to ensure precise definition. The conservation analyses have been strengthened through more rigorous statistical treatment and by explicitly excluding regions overlapping annotated coding sequences.

With respect to modular domain features and potential transposable element–related origins, we have adopted more cautious language and limited our interpretations to what is directly supported by the data. Finally, we acknowledge that current proteomic validation remains limited and have clarified this point in the manuscript while outlining the potential for future integration of large-scale public mass spectrometry datasets in Discussion.

> *The authors additionally report an interesting observation that many ncORFs on mRNA co-translate with the main CDS of the same gene. Because canonical models often posit that uORF translation suppresses downstream CDS translation, further analysis would be valuable. In particular, it would be useful to determine whether patterns of co-translation differ among ORF types or evolutionary categories and to discuss possible regulatory mechanisms underlying these relationships.*

We thank the editor for this thoughtful comment. As noted in our response to Reviewer #2, uORF–CDS co-translation does not contradict the canonical model in which uORFs repress downstream CDS translation. Co-translation reflects concurrent ribosome occupancy, whereas repression concerns the fraction of initiating ribosomes that ultimately reach and translate the CDS. Following the editor's suggestion, we further examined whether co-translation patterns differ across ORF types or evolutionary categories. We found that ncORFs co-translating with their corresponding main CDSs are predominantly uORFs. However, these uORFs do not show statistically significant differences in conservation metrics or evolutionary age compared with other non-overlapping uORFs. Thus, we did not detect clear subtype- or age-specific distinctions among co-translating ncORFs. We have clarified these analyses in the revised manuscript.

> *Addressing these points would enhance the precision, interpretability, and robustness of the study's conclusions.*
>
> ***Reviewer #2 (Recommendations for the authors):***
>
> *(1) The authors developed and refined a standardized pipeline to analyze nearly 400 ribo-seq datasets, identifying over 10,000 novel non-canonical ORFs in both human and mouse samples. Given the scale of this analysis, it is intriguing to consider how many of the newly identified non-canonical ORFs are consistently detected across multiple sample types (conservatively expressed ORFs), how many are restricted to specific tissues/ or tissue-specific ORFs), and how many were detected in only a single or very few samples (ORFs requiring further validation). Providing these data could offer new insights into understanding ORF translation.*

Thanks for this constructive suggestion. This information has been presented in the revised Supplementary Table 3 and in a newly added supplementary figure (Fig. S1B), which together provide a clearer overview of ncORF detection consistency and context specificity.

> *(2) The authors' validation of MS data lacks specific details in the paper. Regarding the MS-supported ORF mentioned in Lane 117, which dataset's MS data is being referenced? Or does it refer to the content in Reference 20? At present, substantial research exists in both public general proteomics studies (e.g., CPTAC) and MS investigations targeting non-canonical ORFs. We recommend the authors incorporate additional MS data or public MS-based databases to strengthen validation in this area (PMID: 34129944, 39794466, 37823596，39413795).*

We thank the reviewer for this comment and for the helpful suggestions. The MS-supported ORFs mentioned in line 117 refer to the compilation reported in Reference 20, which integrates evidence from multiple independent proteomics studies. In addition, we examined MS-supported ORFs curated by GENCODE and PeptideAtlas, which are shown in Fig. 1E.

We agree that incorporating additional MS datasets would further strengthen validation of ncORFs. Studies cited by the reviewer and recent community efforts such as the GENCODE and PeptideAtlas analyses (PMID: 39314370) provide valuable examples in this direction. However, performing a comprehensive reanalysis of more than 95,000 public human MS runs is computationally demanding and currently infeasible for our group given resource and funding constraints.

To our knowledge, ongoing community-wide initiatives are working toward more comprehensive catalogs of translated human ncORFs. Large-scale, exhaustive MS searches will be particularly effective once a community consensus annotation framework for ncORFs is established. We have added discussion of these limitations and future directions in the revised manuscript.

> *(3) The authors classified ncORFs into three groups-"Ancient," "Young," and "Old"-based on their origin nodes. However, both the "Young" and 'Old' groups appear to be "mammalian-specific," yet the specific criteria for their division remain unclear. It is recommended to more clearly define in the figure legend or main text how "Young" and "Old" are categorized (e.g., based on specific evolutionary nodes or distance thresholds from nodes to the end) to avoid reader confusion.*

In Fig. 5, “old” and “young” were intended as qualitative descriptors of relative evolutionary age based on the position of ncORF origination nodes along the phylogeny, as indicated on the x-axis. They were not meant to represent discrete categories. To avoid confusion, we have revised the manuscript to use “older” and “younger” throughout when referring to relative age differences. A binary classification is used only in Fig. 6E, where ncORFs are grouped into ancient (pre-mammalian) and younger (mammalian-specific) categories. This distinction is clearly defined in both the main text and the corresponding figure legend.

> *(4) The authors observed an intriguing phenomenon: ncORFs on mRNA tend to co-translate with the main CDS of the same gene. However, the conventional view holds that uORF translation often inhibits the translation of the main CDS. I suggest the authors could refine their analysis in this section further. For instance, do different types of ORFs or ORFs at different evolutionary levels exhibit distinct levels of cotranslation with the main CDS? Additionally, while observing this phenomenon, the authors should also propose hypotheses regarding the regulatory mechanisms involved in these processes.*

We thank the reviewer for these constructive suggestions. After excluding CDS-overlapping ORFs, we identified 258 human and 128 mouse ncORFs that co-translate with their corresponding main CDSs. With the exception of 10 human dORFs, all remaining cases were uORFs. We compared these cotranslating ncORFs with other non-overlapping uORFs and dORFs but did not detect statistically significant differences in evolutionary age and conservation metrics. Because no clear distinguishing features emerged, we did not include these results in the manuscript.

Importantly, the observation of uORF–CDS co-translation does not contradict the established repressive role of uORFs. Co-translation reflects concurrent ribosome occupancy, whereas repression concerns the proportion of initiating ribosomes that ultimately translate the CDS. For example, if two ribosomes initiate within a given interval and one translates the uORF while one translates the CDS, CDS output is reduced by 50% relative to a uORF-free transcript. If four ribosomes initiate under the same repressive regime, two may translate the uORF and two the CDS. In this case, absolute translation of both ORFs increases, while the fractional repression remains unchanged. Thus, co-translation is compatible with a regulatory model in which uORFs reduce CDS translation efficiency without abolishing it. This has been clarified in the revised manuscript.

https://doi.org/10.7554/eLife.109128.2.sa0