Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A unified framework for the systematic detection of microproteins in standard proteomics workflows using a Ribo-seq informed transcriptomic language model

Nicolas Provencher
Université de Sherbrooke
Sébastien Leblanc
Université de Sherbrooke
Jean-François Jacques
Université de Sherbrooke
Xavier Roucou
Xavier.Roucou@USherbrooke.ca

Université de Sherbrooke

Article

Keywords: microproteins, Ribo-seq, non-canonical ORFs, mass spectrometry, TIS Transformer

Posted Date: June 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9982577/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on September 25th, 2026. See the published version at https://doi.org/10.1038/s41598-026-72489-9.

Abstract

Although ribosome profiling (Ribo-seq) has revealed widespread translation of non-canonical open reading frames (ncORFs), integrating this information into routine mass spectrometry (MS) workflows remains challenging due to biochemical constraints and large search databases. We retrained TIS Transformer, a transcriptomic language model, by incorporating 40,832 Ribo-seq-derived non-canonical translation initiation sites, enabling the prediction of 48,265 ncORFs, including non-AUG starts, while preserving canonical protein detection. We combined these predictions with Swiss-Prot sequences to create Swiss-Prot/MicroProt, a size-controlled database suitable for standard proteomic analysis. Reanalysis of a comprehensive HeLa MS dataset using this database maintained over 98% concordance with Swiss-Prot searches, recovering 13,128 canonical proteins and identifying 348 microproteins, 270 of which had orthogonal evidence of their existence. This unified framework facilitates systematic microprotein detection in conventional proteomics workflows, bridging the gap between canonical and non-canonical proteome analyses without requiring specialized experimental protocols.

INTRODUCTION

Typical genomic annotations rely heavily on computational rules, including a minimum open reading frame (ORF) length threshold of approximately 100 codons. This threshold was imposed to avoid numerous false positives resulting from the random distribution of start and stop codons throughout the genome[1–4]. However, this approach inevitably excludes thousands of short or non-canonical open reading frames (ncORFs), despite the absence of a biological rule requiring a minimal length for coding sequences for functional proteins. Very small proteins play critical roles in fundamental cellular processes, including calcium homeostasis, mitochondrial function, and developmental signalling[5,6]. Additionally, traditional annotation guidelines have generally restricted eukaryotic genes or mRNAs to a single annotated coding sequence, further limiting the recognition of additional coding regions within the same transcripts. These methodological constraints have led to a systematic underestimation of the prevalence and importance of ncORFs and the proteins they encode[7–9].

Ribo-seq has now transformed our understanding of the coding landscape, revealing that translation occurs beyond the boundaries of conventionally defined coding sequences (CDSs) in transcripts annotated as mRNAs[6,10–12]. At the mechanistic level, Ribo-seq identifies the translatome by capturing ribosome-protected RNA fragments exhibiting the distinctive characteristics of active translation, such as triplet periodicity, ORF framing, and initiation site enrichment[10]. The ability to directly measure ribosome occupancy, rather than inferring coding potential from sequence characteristics alone, has been revolutionary. Studies using Ribo-seq, including the seminal study by Ingolia et al.[13], have revealed that numerous ncORFs—classified into 6 main biotypes (upstream ORFs [uORFs], upstream-overlapping ORFs [uoORFs], internal ORFs [intORFs], downstream-overlapping ORFs [doORFs], downstream ORFs [dORFs], and lncRNA ORFs; see Fig. 1a)—have robust translation signals, despite being previously overlooked[14,15]. An initial estimate from the international consortium TransCODE suggested that at least

7,264 human ncORFs show evidence of translation[15]. Reanalysis of 295 proteomics datasets from the ProteomeXchange data repository by this consortium revealed peptide evidence for 2.5% (183 out of 7,264) of these ncORFs[16]. As the majority of proteins predicted to be encoded by ncORFs are less than 100 amino acids in size, they are referred to as microproteins[17–19].

Furthermore, the application of specialized Ribo-seq variants, such as those utilizing harringtonine or lactimidomycin to stall ribosomes specifically at initiation sites, has facilitated the mapping of non-canonical translation initiation sites (TISs). This has revealed a vast landscape of translation initiation at near-cognate codons—codons that differ from the canonical AUG start codon by a single nucleotide (e.g., CUG, GUG)—which frequently initiate the translation of ncORFs[13,20–22]. Hence, Ribo-seq is the gold standard for discovering translated ncORFs[12,23,24] and serves as the first reliable, large-scale indicator of microprotein synthesis, laying the foundation for experimental validation.

Although proteomics has become a common and indispensable technique, microproteins are intrinsically challenging to detect with standard shotgun mass spectrometry due to combined biochemical and computational limitations. Biochemically, these proteins are typically present at low abundance, and their short length results in the generation of only a few tryptic peptides, many of which are either too short to be uniquely identifiable or too long to be efficiently fragmented and detected by MS/MS[25,26]. Computationally, a custom protein database, including both canonical proteins and microproteins must be generated. Databases such as OpenProt contain all predicted microproteins larger than 30 amino acids, in addition to canonical proteins[27]. However, such large databases reduce identification sensitivity and yield fewer peptide/protein identifications, requiring specialized mass spectrometry workflows and significant computing resources to address these issues[28–31]. Although these methods are useful for exploring the microproteome, they are not suitable for routine MS-based proteomic analyses. An alternative approach would be to combine Ribo-seq and proteomic analyses of the same sample to refine a database that includes only ncORFs with evidence of translation[32,33]. However, combining Ribo-seq and proteomics on the same biological samples is challenging because these methods differ significantly in sample preparation requirements, sensitivity, data integration workflows, and quality control. Ribo-seq is a specialized technique that requires complex library preparation steps and is more expensive per sample than RNA-seq. These biological, technical, and economic constraints make it difficult to achieve routine, combined integration at the sample level in most research laboratories that perform proteomic analyses.

TIS Transformer is a deep learning model that uses transformer-based architectures, originally developed for natural language processing, to predict TISs directly from transcript nucleotide sequences[34]. Hence, such a predictor can be used to determine TISs without requiring experimental data. The model was trained on a dataset comprising Ensembl transcripts and labels corresponding to annotated TISs. The dataset contained 96,655 mRNAs, each with a unique annotated TIS (all AUG), and 154,466 lncRNAs labeled as not containing any TIS. Although it was able to identify some ncORFs embedded in lncRNAs and occasionally additional ORFs appearing alongside canonical CDSs, the

model's performance is fundamentally limited by its reliance on the Ensembl annotation framework. This dependence means the model has very few relevant ncORF examples to learn from. This causes the model to overlook sequence features unique to ncORFs. A striking example of this is the prediction of only 0.006% of TISs initiating at non-AUG codons[34], although 30%–60% of non-AUG initiation is observed in Ribo-seq analyses[13,22]. These discrepancies between the predictions and experimental results were expected, given that non-AUG start sites were incorrectly labelled as non-TISs in the training set.

We propose that incorporating the characteristics of experimentally detected ncORFs into the training data, rather than exclusively canonical TISs, would improve the performance of the model and allow it to overcome the current limitations. To test this hypothesis, we created a training set comprising 96,555 canonical TISs from Ensembl and 40,832 additional TISs associated with ncORFs, as detected by Ribo-seq across 78 independent studies. We trained TIS Transformer using both the new training and the original Ensembl sets. Integrating ncORFs into the model maintained TIS Transformer's predictive performance for canonical ORFs while simultaneously expanding its coverage to include 48,265 ncORFs, representing non-canonical translation products or microproteins.

To support the routine identification of microproteins without compromising the detection of canonical proteins in standard shotgun proteomics, we assembled SwissProt/MicroProt, a database combining 48,265 ncORF-derived microproteins with canonical Swiss-Prot sequences. By including only microprotein entries predicted by TIS Transformer, rather than attempting exhaustive coverage of all possible ncORFs, the database is expected to remain well-suited to conventional proteomics workflows. This strategy also aims to preserve robust canonical protein identifications, providing an accessible intermediary between exploratory microproteome resources and widely used reference databases. When processing MS raw data generated from HeLa samples and performing database searches using MaxQuant (Andromeda)[35], this expanded database maintained near-complete concordance of peptide-spectrum matches (PSMs), with 5,947,380 shared PSMs when using the standard Swiss-Prot database (98.83% and 99.33% overlap versus Swiss-Prot and Swiss-Prot/MicroProt searches, respectively). This process resulted in only a marginal decrease in PSMs (~ 0.49%) and an increase in the number of protein groups by 413, all without disrupting a core set of 13,291 Swiss-Prot proteins. Notably, this resulted in the identification of 539 microproteins.

MATERIALS AND METHODS

Ribo-seq

We selected 108 experiments from 74 Ribo-seq initiation and elongation studies that had at least two libraries per sample. This represented a total of 771 FASTQ files. These studies were performed on 50 different human cell lines and tissues (Supplementary Table 1). A detailed description of the studies is available in the sample sheets provided in the GitHub repository of this pipeline. The Ribo-seq pipeline was developed in three stages to facilitate manual validation and quality control (Fig. 1B).

FASTQ preprocessing

FastQC[36] 0.12.1 was used to ensure the quality of raw sequence reads prior to downstream analyses. Adaptors and potential unique molecular identifiers were trimmed using Trim-Galore[37] 0.6.10. Manual checks of the FastQC reports were performed before and after trimming to identify any incomplete adapter trimming or locate any remaining unique molecular identifiers in the raw fastq files. The trimming arguments used to optimize sample quality are annotated in the sample sheet file shared with the pipeline in the “trimming_arg” column.

Reads alignment

Bowtie[38] v1.3.1 was used to filter out non-ribosome-protected fragment reads from the samples using the default settings. A Bowtie index for contaminants was constructed, and the reads were aligned to it. Reads that mapped to the contaminant fasta were rejected. The split-aware aligner STAR[39] v2.7.11b was then used to map the remaining reads on the Ensembl v114 human transcriptome, which was based on the grch.38 human genome. Study files with a mapping depth of less than 20% were excluded.

ORF calling

Given the large-scale nature of this project, we required an ORF calling framework that could efficiently identify translation events from the ribosome-protected fragments. For this purpose, we used the state-of-the-art tool RiboTIE[40], integrated within *transcript_transformer* v1.0.1. RiboTIE identifies translated ORFs by scoring candidate start–stop codon pairs using a transformer model trained on ribosome occupancy patterns along the transcripts. Its ability to identify both canonical ORFs and ncORFs, combined with its robust performance across heterogeneous ribosome profiling datasets of varying quality, makes it optimal for our analysis.

Retraining of TIS Transformer

TIS Transformer[34] is a deep learning model specifically designed to predict TISs in transcripts using nucleotide sequence information alone. ORFs are subsequently defined by extending the predicted TIS positions to the first downstream in-frame stop codon. Transcriptomic data were obtained from Ensembl[41] (v114; Homo_sapiens.GRCh38.114. gtf and Homo_sapiens.GRCh38.dna.primary_assembly.fa) were processed using TIS transformer (transcript_transformer v1.0.1) with the argument --data to generate a baseline HDF5 (h5) file. For our retraining methodology, a modified version of this file was created using a custom Python script (*add_tis.py*). This script incorporated 40,832 non-canonical TIS derived from our Ribo-seq pipeline (Fig. 1B), as positive labels (Fig. 1C). Finally, the machine learning model was trained using both the original and

modified h5 files to evaluate the impact of the added non-canonical TIS labels on the model performance. For simplicity, the retrained model is referred to as TISnc Transformer in this study.

To ensure a standardized comparison between the two training methodologies, the prediction thresholds were calibrated to achieve a consistent recall rate of 91–92% for canonical TISs. Under these parameters, the required thresholds were determined to be 0.05 for the baseline model and 0.10 for TISnc Transformer.

Swiss-Prot/MicroProt database

Swiss-Prot/MicroProt is a customized database that integrates Swiss-Prot[42] entries and microproteins predicted by TIS Transformer retrained with ncTISs (Fig. 1C,D). The database enables the identification of microproteins that fall within the detectable range of conventional shotgun proteomics workflows alongside canonical proteins, without the need for exhaustive microproteome coverage. Therefore, identifications obtained using this database should be interpreted as reflecting a subset of microproteins represented in the Swiss-Prot/MicroProt sequence space that can be detected under the employed experimental conditions.

Statistical analyses

To evaluate whether the TIS Transformer and TISnc Transformer models differed significantly in their ability to recover TransCODE ncORFs, we compared the proportion of model-predicted ncORFs that overlapped with the 7,264 entries from the TransCODE catalog[15]. Overlap analysis was performed at the amino acid level following *in silico* translation of ncORFs. Only microproteins with amino acid sequences identical to those of the TransCODE entries were considered a match. We created the following 2×2 contingency table showing the number of ncORFs that overlap with TransCODE and the number that do not for each model:

TIS Transformer	**Overlap with TransCODE ncORFs**	**No overlap with TransCODE ncORFs**	**Row total**
	a = 548	**b = 18,031–548 = 17,483**	**18,031**
TISnc Transformer	c = 2,788	d = 48,265–2,788 = 45,477	48,265
Grand total	3,336	62,960	66,296

The statistical significance of the differences in recovery rates was first assessed using a two-proportion z-test. Simultaneously, we performed a chi-squared test of independence on the same contingency table to verify the robustness of the results. For all tests, the assumptions of large sample sizes and expected cell counts were satisfied. Statistical significance was set at $p < 0.05$.

To evaluate the similarity of the sets of leading proteins returned by the two searches with Swiss-Prot or Swiss-Prot/MicroProt, we calculated the degree of proteome concordance using the Jaccard index. This metric was chosen because it provides a size-independent measure of set similarity, which is useful for evaluating the stability of the identifications when the underlying search space differs substantially in size.

MS reanalysis

The study by Bekker-Jenssen et al.[43] was selected for reanalysis of the data. This generated a comprehensive HeLa canonical proteome, identifying 584,000 unique peptides and over 12,200 protein-coding genes. Raw MS data files were downloaded from the ProteomeXchange/PRIDE repository (identifier PXD004452). We used MaxQuant[35] (v2.7.5.0) together with its built-in search engine Andromeda, one of the most widely used software packages in proteomics data processing. Two sequence databases were used: Swiss-Prot and Swiss-Prot/MicroProt (Fig. 1C,D).

Sequence database searches were performed using the same parameters as those used in the original study[43]. Microproteins with redundant sequences in Swiss-Prot were removed. In brief, the contaminant database was set to 'default', and all corresponding results were excluded from further analysis. The data were searched using the following generic variable modifications: methionine oxidation, protein N-terminal acetylation, and pyro-glutamic acid formation from glutamic acid or glutamine. Additionally, asparagine and glutamine deamidation were considered. Alkylation modification was set as a fixed modification (typically carbamidomethylated cysteine). The false discovery rate was set at 1% for PSM and protein levels.

Microprotein validation

Orthogonal validation with Ribo-Seq

RiboCRYPT[44] - This interactive database is designed for the visual exploration and analysis of Ribo-seq data and a powerful tool for finding ncORFs. The default ribosome-protected fragment (RPF) signal corresponds to elongating ribosomes. The validation criteria were as follows: (1) The RPF signal in the reading frame of the predicted ncORF was greater than 50 across its entire length; (2) The RPF signal showed a clear increase (10x the average noise signal or 1.5 × the out-of-frame signal inside another overlapping ORF) at the predicted TIS. If the increased signal started upstream or downstream of the predicted TIS, but criteria 1 and 3 were met, the validation was considered positive, as this was most likely due to initiation occurring upstream or downstream of the predicted TIS; (3) The RPF signal showed a clear decrease back down to a noise signal or the out-of-frame signal of another overlapping ORF at the predicted stop codon.

TransCODE v35[15] – This is a reference database of ncORFs with Ribo-seq evidence that likely encodes microproteins. The protein sequences of MaxQuant leading microproteins were matched to the orf_sequence columns in Supplementary Tables 2 and 3 from the TransCODE v35 manuscript, and the orf_name from the tables was added to the column “transcodev35 ORF mapping” of Supplementary Table 5 when a match was found.

RiboTIE[40] - ncORFs are listed in Supplementary Table 2. The leading microprotein sequence was compared to the set of unique protein sequences detected in the aggregation of ORFs returned by RiboTIE. A match is annotated as True in Supplementary Table 5.

Independent validation

Evidence of microprotein expression in MS in other studies specifically focused on microprotein detection.

Wacholder et al.[19] - This study benchmarked MS methods to establish a robust standard for the detection of previously unannotated human microproteins. Peptides in the protein group with a leading microprotein were matched to the unmodified peptide column of Supplementary Table 1 of the study. The orf_names associated to the matched peptides in the table was then annotated when there was one or multiple match

Deutsch et al.[16] - This manuscript determines a high-confidence, MS-based landscape for TransCODE v35 ncORFs and their associated microproteins. For the microproteins identified by Deutch et al. in Supplementary Table 3, the protein sequences were retrieved by matching the column Ribo-Seq_ORF to the ORF in the TransCODE v35 supplementary table s2 and s3. The sequences were then matched to the sequences of the leading proteins detected in the MaxQuant analysis, and the TransCODE v35 id was applied when there was a match.

OpenProt[27] is a proteogenomic resource that uses a machine learning-based method to recalculate peptide scores, improving peptide identification and increasing confidence in microprotein identification. The sequences of the leading microproteins were matched with the microproteins (IP_ and II_) sequences in the OpenProt human v2.2 fasta file. The OpenProt v2.2 list of human proteins was used to return any MS evidence for each OpenProt ID.

RESULTS

Distribution of Ribo-seq ORFs and TISs

Our Ribo-seq pipeline identified 87,110 unique ORFs (Supplementary Table 2). The distribution of ORF types identified by RiboTIE remained consistent across the 118 experiments (Supplementary Fig. 1), confirming its ability to provide reliable results for translated ORFs under various ribo-seq analysis conditions, including those with different sequencing depths[40]. In addition to annotated CDSs

representing 50.4% of ORFs, 27.4% corresponded to N-terminal truncation and N-terminal extension of canonical CDSs (Supplementary Fig. 2A). The number of N-terminal truncation was approximately twice as high as the number of N-terminal extensions (Supplementary Fig. 2B). These proportions were similar to those previously reported[40]. As this study focused on ncORFs rather than CDS isoforms, N-terminal extensions and truncations were excluded from subsequent analyses. The distribution of ncORFs indicated that uORFs, uoORFs, and lncRNA ORFs are the three most abundant ncORF types, which is consistent with RiboTIE[40] and TransCODE[15] v35, as shown in Supplementary Fig. 2C. With regard to the distribution of non-ATG TISs, CTG, TTG, and GTG represented the most abundant categories in descending order, as previously reported (Supplementary Fig. 2D). In total, the pipeline identified 40,397 ncTISs associated with 19,585 distinct ncORFs (Supplementary Table 3). As each TIS is associated with a transcript and the same ORF can appear on multiple transcripts, there are more TISs than unique ORFs.

Overall, these results revealed that the ncORFs identified by our ribo-seq pipeline in the 118 analyzed experiments exhibited characteristics similar to those previously reported.

The integration of ncORFs in the training set does not affect TIS Transformer's ability to predict canonical ORFs and increases coverage of ncORFs at inference

It is important to determine the impact of adding ncTISs on the stability of TIS Transformer. For simplicity, TIS Transformer retrained using a combination of Ensembl TIS and ncTISs is termed TISnc Transformer. Retraining slightly degraded the performance of the model, with an area under the precision-recall curve of 0,79 for the original training with Ensembl, and 0,76 for retraining (Fig. 2A). However, TISnc Transformer retained a large stable core of the original predictions (*i.e.*, 60,371 canonical ORFs) (Fig. 2B). We observed a modest loss of 1254 canonical ORFs, representing 2.03%. This indicates that the model learned stable patterns for canonical TISs that persisted after integrating 40,397 ncTISs into the training dataset.

TISnc Transformer predicted 48,265 ncORFs compared to 18,031 for TIS Transformer (Fig. 2B; Supplementary Table 4). This significant increase in the volume of predicted ncORFs at a stable number of canonical ORFs confirmed that retraining was not redundant and allowed the model to capture the new patterns associated with ncTISs. Notably, the model successfully predicted non-ATG ncRNAs and a larger number of uORFs (Fig. 2C), in agreement with experimental Ribo-seq data (Fig. S1C). Unlike TIS Transformer, it also captured intORFs (Fig. 2C).

The TransCODE consortium provides a consensus catalog of 7,264 human ncORFs with strong translation signatures that are now annotated by GENCODE[15]. With the initial TIS Transformer model, 3.04% (548/18,031) of the predictions overlapped with the TransCODE catalog (Fig. 2D). With TISnc Transformer, the proportion of TransCODE ncORFs within the total prediction set increased to 5.78% (2,788/48,265). To test whether the larger number of TransCODE ncORFs recovered by TISnc Transformer was merely a reflection of its higher total prediction volume or a consequence of improved model sensitivity, we evaluated the significance of this difference between the two models using both a

two-proportion z-test and a chi-squared test. Both statistical tests indicated a highly significant enrichment of curated ncORFs in the predictions of the TISnc Transformer (two-proportion z-test: $z = -14.35$, $p = 1.1 \times 10^{-46}$; chi-squared test: $\chi^2 = 205.24$, $p = 1.5 \times 10^{-46}$). These findings reject the hypothesis that the increased identification of TransCODE ncORFs is a stochastic byproduct of a larger search space. Instead, they suggested that retraining enhanced the ability of TISnc Transformer to distinguish biologically relevant ncORFs from the background.

Collectively, these results support our hypothesis that incorporating ncTISs in the training set of TIS Transformer improves the model's predictive performance on ncORFs.

Database search using MaxQuant (Andromeda)

Our customized Swiss-Prot/MicroProt database for microprotein detection was built by including the protein sequences of ncORFs predicted by TISnc Transformer into the Swiss-Prot database, which is typically used for standard MS analyses. Swiss-Prot/MicroProt aims to support the identification of microproteins without compromising the detection of canonical proteins, enabling both to be analyzed using conventional shotgun proteomic workflows typically optimized for canonical proteins.

To evaluate the impact of expanding the database from 20,431 to 68,531 entries, we compared the search results of Swiss-Prot and Swiss-Prot/MicroProt. The vast majority of identifications remained stable across both searches, with 5,947,380 shared peptide-spectrum matches (PSMs) (Fig. 3A). The search using Swiss-Prot showed a 98.83% overlap, while the search using Swiss-Prot/MicroProt showed a 99.83% overlap relative to the total number of identified proteins. We observed a marginal decrease in total PSMs when using the Swiss-Prot/MicroProt database (-0.49%). This minor reduction is consistent with the increased search space inherent in larger databases; the addition of target sequences can lead to a higher frequency of alternative matches, requiring a more stringent score threshold to maintain a 1% FDR. Thus, a fraction of spectra loses confident identifications in the expanded search[45–47] MaxQuant reports all identifications as protein groups[48,35]. Despite a slight decrease in the number of PSMs, the number of protein groups increased by 215, indicating new protein-level identification (Fig. 3B). The two searches shared a core of 12,020 protein groups, accounting for 87.77% and 84.24% of the total number of protein groups, respectively, indicating a strong consistency. When using the canonical Swiss-Prot database alone, 1,481 protein groups were uniquely identified, whereas the Swiss-Prot/MicroProt database yielded 1,894 specific protein groups, reflecting the successful assignment of spectra to microprotein entries (Fig. 3D) without substantially affecting Swiss-Prot protein detection. Interestingly, 135 PSMs initially allocated to peptides from Swiss-Prot proteins were reallocated to peptides from microproteins with higher MaxQuant scores (Fig. 3D; see Supplementary Fig. 3 for examples of reallocated MS2 spectra).

For protein inference, MaxQuant[35] assigns a leading protein to represent each protein group, selecting the entry with the strongest peptide-level support (e.g., number of peptides, scores, and sequence coverage). The concordance between the two database searches was high: 13,291 leading proteins (or 13,265 genes) were shared, corresponding to 98.55% of all leading proteins identified in the Swiss-Prot

database (13,291/13,501) and 98.98% of those identified in the Swiss-Prot/MicroProt database (13,291/13,914). When considering the combined set of leading proteins (14,124 total), the Jaccard index was 94.1%, indicating that the inferred canonical proteome remained highly stable despite the substantial expansion of the search space. Among the 623 leading proteins unique to the Swiss-Prot/MicroProt search, 539 (or 403 genes) corresponded to microproteins (Fig. 3D, Supplementary Table 5). Overall, microproteins accounted for 3.87% of the total (i.e., 539 out of 13,914). In contrast, canonical Swiss-Prot proteins exhibited a modest contraction, with 210 leading proteins detected only in the Swiss-Prot search versus 84 detected only in the expanded search. This asymmetry is consistent with the expected effects of a larger database, including slightly stricter protein-level FDR thresholds and changes in peptide parsimony when additional candidate sequences are available[45–47].

All ncORF classes were represented among the ncORFs encoding microproteins, with doORF and lncRNA-ORFs being the most frequently observed (Fig. 3E). The initiation codons included 518 ATG, 14 CTG, five GTG, and two TTG. Examples of annotated MS2 spectra corresponding to microproteins encoded by these ncORF classes are presented in Supplementary Fig. 4.

Together, these results show that expanding the database from 20,431 entries in Swiss-Prot to 68,531 in Swiss-Prot/MicroProt induces only minor reshaping of canonical protein groups while enabling substantive recovery of microproteins that are inaccessible to searches using Swiss-Prot alone.

Microproteins

Before exploring the biological functions of the microproteins identified by MaxQuant in more detail, it is essential to confirm that they are indeed translated and exist as stable protein products within cells. This can be achieved using methods such as Ribo-seq and targeted proteomics. This is particularly important because microproteins are very short, tend to be present in low quantities, and often produce only one detectable peptide, making them difficult to identify with confidence using shotgun MS alone[25,26]. Although these validations are beyond the scope of our study, we cross-referenced our data with current proteogenomic resources to assess the quality of our results. Of the 539 microproteins identified by MaxQuant, 395 have previously shown evidence of orthogonal (*i.e.*, Ribo-Seq) or independent (*i.e.*, MS) expression (Fig. 3F, Supplementary Table 5). Of these, 364 were detected at the translational level by Ribo-seq, 90 at the peptide level by MS, and 39 by both methods. This level of overlap with external sources, based on a cursory comparison, is encouraging and suggests that the proposed analytical method can be used to discover microproteins with evidence of expression in the human proteome.

DISCUSSION

The significant differences between standard shotgun proteomic workflows and proteogenomic pipelines dedicated to microprotein discovery hinder the simultaneous detection of canonical proteins and microproteins, thereby impeding the functional study of microproteins in routine proteomics experiments. Here, we propose pairing an improved TIS prediction framework with a practical search database to address this issue. Specifically, we showed that expanding TIS Transformer training set to

incorporate Ribo-seq-supported ncTISs enhanced the detection of both canonical TISs and ncTISs without compromising the performance on canonical CDSs. In parallel, we leveraged these gains in prediction for routine proteomics by constructing a SwissProt/MicroProt database. In the reanalysis of MS data for the HeLa proteome, this database maintained near-complete PSM concordance with conventional SwissProt searches (98.8–99.3% overlap), incurred only a 0.5% decrease in PSMs, and increased protein groups by 413. This yielded 539 microprotein identifications, accounting for 3.90% of the total number of proteins identified. A significant proportion of these genes have previously been shown to exhibit evidence of translation via Ribo-seq. These results indicate that the Swiss-Prot/MicroProt database can be used to systematically detect microproteins in shotgun proteomic experiments.

MS-based proteomics has become a routinely used core technique in functional biology, a trend reflected in the significant recent increase in the generation and sharing of MS-based datasets through ProteomeXchange[49]. Continuing to investigate microproteins and canonical proteins separately has important conceptual and practical implications for functional proteomics. This division perpetuates methodological bias toward larger, well-annotated proteins and systematically underestimates the regulatory impact of microproteins in cellular and disease contexts[5]. The resulting fragmentation of datasets and resources further limits the integration of microproteins into protein–protein interaction networks, pathway representations, and system-level models, hindering efforts to construct comprehensive and quantitative views of cellular function[50,51]. The Swiss-Prot/MicroProt database is expected to help bridge the gap between these two fields, thereby improving the quality of biological information.

An alternative approach to achieving a standard strategy for detecting both canonical proteins and microproteins within an inclusive database compatible with common search engines, such as MaxQuant/Andromeda, would be to restrict shotgun MS searches to ncORFs supported by matched Ribo-seq evidence. However, this would be impractical for routine workflows, as generating Ribo-seq data for every sample would significantly increase costs and labour. In addition, unlike proteomics pipelines, which typically accommodate standardized sample preparation, diverse biological materials such as tissues, patient biopsies, archived specimens, or large cohorts cannot readily be adapted to Ribo-seq protocols. Therefore, incorporating Ribo-seq into every MS experiment would be either too resource-intensive and low-throughput or not feasible for many sample types.

Despite the growing recognition of their prevalence and importance[52], microproteins are still commonly studied separately from canonical proteins, largely because they typically require tailored proteogenomic searches rather than sequence-based predictions and standard proteomics, which are sufficient for canonical proteins[19]. However, investigating the microproteome separately from the canonical proteome inevitably leads to fragmented functional proteomics, limiting our ability to recover complete biological pathways and regulatory interactions in the proteome[50,51]. We suggest that analyzing canonical proteins

and microproteins within a unified framework using the Swiss-Prot/MicroProt database will help recover a significant fraction of this biological information.

Limits of the study - Our approach is not designed for cutting-edge microprotein research, which typically requires a multidisciplinary approach integrating dedicated methods such as low–molecular-weight protein enrichment, high-resolution proteomics, ribosome profiling, and highly specialized computational analysis pipelines[6,53]. It only allows the detection of microproteins that are successfully extracted under the experimental conditions used for sample preparation, are sufficiently abundant, generate peptides that can be analyzed by MS, and are present in the Swiss-Prot/MicroProt database. Additionally, because TISnc Transformer was trained with ncORFs derived from Ribo-seq data obtained through large-scale meta-analyses, our approach has the advantage of leveraging existing data without the need for new Ribo-seq sequences. However, it may fail to account for ncORFs that are translated only in rare circumstances. Finally, factors such as RNA secondary and tertiary structures, RNA-binding proteins, ribosome pausing, and uORFs strongly influence translation initiation[54,55]. Although the TIS Transformer model achieves state-of-the-art performance in learning sequence-encoded determinants of translation initiation[34], its predictions are still limited by biological features that cannot be captured by the primary sequence alone.

Declarations

FUNDING

This research was supported by a grant from the Canadian Institutes for Health Research (PJT-175322), the Digital Research Alliance of Canada and the Canada Research Chair in Functional Proteomics and Discovery of Novel Proteins to X.R.

Author Contribution

Conceptualization: N.P., S.L., and X.R.; investigation: N.P., S.L., and J-F. J: methodology, formal analysis, data curation, writing – original draft, and visualization; N.P. and X.R.: resources and validation; X.R. and J-F.J.: writing – review and editing; X.R.: funding acquisition.

Acknowledgement

This research was supported by the Canada Research Chair in Functional Proteomics and Discovery of Novel Proteins to X.R. The computational resources for this study were provided by the Digital Research Alliance of Canada. We would like to thank Danny Bergeron of the RNomics platform at the University of Sherbrooke for his assistance with the Ribo-Seq workflow.

Data Availability

The GitHub repository associated with this paper is available at: https://github.com/NicolasProvencher/Provencher_et_al_2026 . All codes written and used in the Ribosome Profiling reanalysis pipeline are available in the repository. The script used to supplement TIS Transformer training data is also available in the repository. The protein FASTA files used for MS analysis are also available in the repository.

References

1. Basrai, M. A., Hieter, P. & Boeke, J. D. Small open reading frames: beautiful needles in the haystack. *Genome Res.* **7**, 768–771. https://doi.org/10.1101/gr.7.8.768 (1997).
2. Okazaki, Y. et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573. https://doi.org/10.1038/nature01266 (2002).
3. Brent, M. R. Genome annotation past, present, and future: how to define an ORF at each locus. *Genome Res.* **15**, 1777–1786. https://doi.org/10.1101/gr.3866105 (2005).
4. Dinger, M. E., Pang, K. C., Mercer, T. R. & Mattick, J. S. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput. Biol.* **4**, e1000176. https://doi.org/10.1371/journal.pcbi.1000176 (2008).
5. Hassel, K. R., Brito-Estrada, O. & Makarewich, C. A. Microproteins: overlooked regulators of physiology and disease. iScience 26, 106781. (2023). https://doi.org/10.1016/j.isci.2023.106781
6. Jaunbocus, N., Ebenki, V., Su, H. & Slavoff, S. A. Eukaryotic microproteins. *Annu. Rev. Biochem.* **94**, 1–28. https://doi.org/10.1146/annurev-biochem-080124-012840 (2025).
7. Mouilleron, H., Delcourt, V. & Roucou, X. Death of a dogma: eukaryotic mRNAs can code for more than one protein. *Nucleic Acids Res.* **44**, 14–23. https://doi.org/10.1093/nar/gkv1218 (2016).
8. Chen, J. et al. Pervasive functional translation of noncanonical human open reading frames. *Science* **367**, 1140–1146. https://doi.org/10.1126/science.aay0262 (2020).
9. Kute, P. M., Soukarieh, O., Tjekdnes, H., Trégouët, D. A. & Valen, E. *Front. Genet.* **12**, 796060. https://doi.org/10.3389/fgene.2021.796060. (2022).
10. Ingolia, N. T., Hussmann, J. A. & Weissman, J. S. Ribosome profiling: global views of translation. *Cold Spring Harb Perspect. Biol.* **11**, a032698. https://doi.org/10.1101/cshperspect.a032698 (2019).
11. Prensner, J. R. et al. What can ribo-seq, immunopeptidomics, and proteomics tell us about the noncanonical proteome? Mol. *Cell. Proteom.* **22**, 100631. https://doi.org/10.1016/j.mcpro.2023.100631 (2023).
12. Chothani, S., Ho, L., Schafer, S. & Rackham, O. Discovering microproteins: making the most of ribosome profiling data. *RNA Biol.* **20**, 943–954. https://doi.org/10.1080/15476286.2023.2279845 (2023).

13. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**, 789–802. https://doi.org/10.1016/j.cell.2011.10.002 (2011).
14. Ingolia, N. T. Ribosome footprint profiling of translation throughout the genome. *Cell* **165**, 22–33. https://doi.org/10.1016/j.cell.2016.02.066 (2016).
15. Mudge, J. M. et al. Standardized annotation of translated open reading frames. *Nat. Biotechnol.* **40**, 994–999. https://doi.org/10.1038/s41587-022-01369-0 (2022).
16. Deutsch, E. W. et al. High-quality peptide evidence for annotating non-canonical open reading frames as human proteins. bioRxiv 2024.09.09.612016. (2025). https://doi.org/10.1101/2024.09.09.612016
17. He, C., Jia, C., Zhang, Y. & Xu, P. Enrichment-based proteogenomics identifies microproteins, missing proteins, and novel smORFs in Saccharomyces cerevisiae. *J. Proteome Res.* **17**, 2335–2344. https://doi.org/10.1021/acs.jproteome.8b00032 (2018).
18. Rathore, A., Martinez, T. F., Chu, Q. & Saghatelian, A. Small, but mighty? Searching for human microproteins and their potential for understanding health and disease. *Expert Rev. Proteom.* **15**, 963–965. https://doi.org/10.1080/14789450.2018.1547194 (2018).
19. Wacholder, A. et al. Community benchmarking and evaluation of human unannotated microprotein detection by mass spectrometry based proteomics. *Nat. Commun.* **17**, 1241. https://doi.org/10.1038/s41467-025-68002-x (2026).
20. Lee, S., Liu, B., Huang, S. H., Shen, B. & Qian, S. B. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proc. Natl. Acad. Sci. USA 109, E2424-2432. (2012). https://doi.org/10.1073/pnas.1207846109
21. Cao, X. & Slavoff, S. A. Non-AUG start codons: expanding and regulating the small and alternative ORFeome. *Exp. Cell. Res.* **391**, 111973. https://doi.org/10.1016/j.yexcr.2020.111973 (2020).
22. Andreev, D. E. et al. Non-AUG translation initiation in mammals. *Genome Biol.* **23**, 111. https://doi.org/10.1186/s13059-022-02674-2 (2022).
23. Valdivia-Francia, F. & Sendoel, A. No country for old methods: new tools for studying microproteins. *iScience* **27**, 108972. https://doi.org/10.1016/j.isci.2024.108972 (2024).
24. Tong, G. & Martinez, T. F. Ribosome profiling reveals hidden world of small proteins. *Trends Genet.* **41**, 101–103. https://doi.org/10.1016/j.tig.2024.12.010 (2025).
25. Ma, J. et al. Improved identification and analysis of small open reading frame encoded polypeptides. *Anal. Chem.* **88**, 3967–3975. https://doi.org/10.1021/acs.analchem.6b00191 (2016).
26. Wacholder, A. & Carvunis, A. R. Biological factors and statistical limitations prevent detection of most noncanonical proteins by mass spectrometry. *PLoS Biol.* **21**, e3002409. https://doi.org/10.1371/journal.pbio.3002409 (2023).
27. Leblanc, S. et al. OpenProt 2.0 builds a path to the functional characterization of alternative proteins. *Nucleic Acids Res.* **52**, D522–D528. https://doi.org/10.1093/nar/gkad1050 (2024).

28. Castellana, N. & Bafna, V. Proteogenomics to discover the full coding content of genomes: a computational perspective. *J. Proteom.* **73**, 2124–2135. https://doi.org/10.1016/j.jprot.2010.06.007 (2010).
29. Renuse, S., Chaerkady, R. & Pandey, A. *Proteogenomics* **11**, 620–630. https://doi.org/10.1002/pmic.201000615. (2011).
30. Nesvizhskii, A. I. Proteogenomics: concepts, applications, and computational strategies. *Nat. Methods*. **11**, 11141125. https://doi.org/10.1038/nmeth.3144 (2015).
31. Declercq, A. et al. MS2Rescore: data-driven rescoring dramatically boosts immunopeptide identification rates. *Mol. Cell. Proteom.* **21**, 100266. https://doi.org/10.1016/j.mcpro.2022.100266 (2022).
32. Koch, A. et al. A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites. *Proteomics* **14**, 2688–2698. https://doi.org/10.1002/pmic.201400180 (2014).
33. de Vieira, E. et al. Rp3: ribosome-profiling-assisted proteogenomics improves coverage and confidence during microprotein discovery. *Nat. Commun.* **15**, 6839. https://doi.org/10.1038/s41467-024-50301-4 (2024).
34. Clauwaert, J., McVey, Z., Gupta, R. & Menschaert, G. TIS Transformer: remapping the human proteome using deep learning. *NAR Genom Bioinform*. **5**, Iqado21. https://doi.org/10.1093/nargab/lqad021 (2023).
35. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass-spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319. https://doi.org/10.1038/nprot.2016.136 (2016).
36. Andrews, S. FastQC: a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
37. Krueger, F. Trim galore. Wrapper Tool Cutadapt FastQC Consistently Apply Qual Adapt Trimming FastQ Files. https://github.com/FelixKrueger/TrimGalore
38. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25. https://doi.org/10.1186/gb-2009-10-3-r25 (2009).
39. Dobin, A. & Gingeras, T. Mapping RNA-seq reads with STAR. *Curr. Protoc. Bioinf.* **51** https://doi.org/10.1002/0471250953.bi1114s51 (2015). 11.14.1-11.14.19.
40. Clauwaert, J. et al. Deep learning to decode sites of RNA translation in normal and cancerous tissues. *Nat. Commun.* **16**, 1275. https://doi.org/10.1038/s41467-025-56543-0 (2025).
41. Yates, A. D. et al. Ensembl 2026. *Nucleic Acids Res.* **54**, D1053–D1060. https://doi.org/10.1093/nar/gkaf1239 (2026).
42. UniProt Consortium. UniProt: the universal protein knowledgebase in 2025. *Nucleic Acids Res.* **53**, D609–D617. https://doi.org/10.1093/nar/gkae1010 (2025).

43. Bekker-Jensen, D. B. et al. An optimized shotgun strategy for the rapid generation of comprehensive human proteoms. *Cell. Syst.* **4**, 587–599e4. https://doi.org/10.1016/j.cels.2017.05.009 (2017).
44. Swirski, M. & Tjeldnes, H. RiboCrypt: Interactive visualization in genomics. R package version 1.16.0, (2025). https://github.com/m-swirski/RiboCrypt
45. Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteom.* **73**, 2092–2123. https://doi.org/10.1016/j.jprot.2010.08.009 (2010).
46. Colaert, N., Degroeve, S., Helsens, K. & Martens. L Analysis of the resolution limitations of peptide identification algorithms. *J. Proteome Res.* **10**, 5555–5561. https://doi.org/10.1021/pr200913a (2011).
47. Verheggen, K., Raeder, H., Berven, F. S., Martens, L., Barsnes, H. & Vaudel. M Anatomy and evolution of database search engines-a central component of mass spectrometry based proteomic workflows. *Mass. Spectromm Rev.* **39**, 292–306. https://doi.org/10.1002/mas.21543 (2020).
48. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372. https://doi.org/10.1038/nbt.1511 (2008).
49. Deutsch, E. W. et al. The ProteomeXchange consortium in 2026: making proteomics data FAIR. *Nucleic Acids Res.* **54**, D459–D469. https://doi.org/10.1093/nar/gkaf1146 (2026).
50. Leblanc, S. & Brunet, M. A. Modelling of pathogen-host systems uding deeper ORF annotations and trancriptomics to inform proteomics analyses. *Comput. Struct. Biotechnol. J.* **18**, 2836–2850. https://doi.org/10.1016/j.csbj.2020.10.010 (2020).
51. Leblanc., S. et al. *Genomics Proteom. Bioinf.* **21**, 515–534. (2023).
52. Du, Z., Trifiro, F. A., Brunet, M. A. & Prensner, J. R. Microproteins in human physiology and pathology. *Biochemistry* **65**, 812–834. https://doi.org/10.1021/acs.biochem.6c00063 (2026).
53. Saghatelian, A. & Couso, J. P. Discovery and characterization of smORF-encoded bioactive peptides. *Nat. Chem. Biol.* **11**, 909–916. https://doi.org/10.1038/nchembio.1964 (2015).
54. Jackson, R. J., Hellen, C. U. T. & Pestova, T. V. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat. Rev. Mol. Cell. Biol.* **11**, 113–127. https://doi.org/10.1038/nrm2838 (2010).
55. Querido, J. B., Diaz-Lopez, I. & Ramakrishnan, V. The molecular basis of translation initiation and its regulation in eukaryotes. *Nat. Rev. Mol. Cel Biol.* **25**, 168–186. https://doi.org/10.1038/s41580-023-00624-9 (2024).

Figures

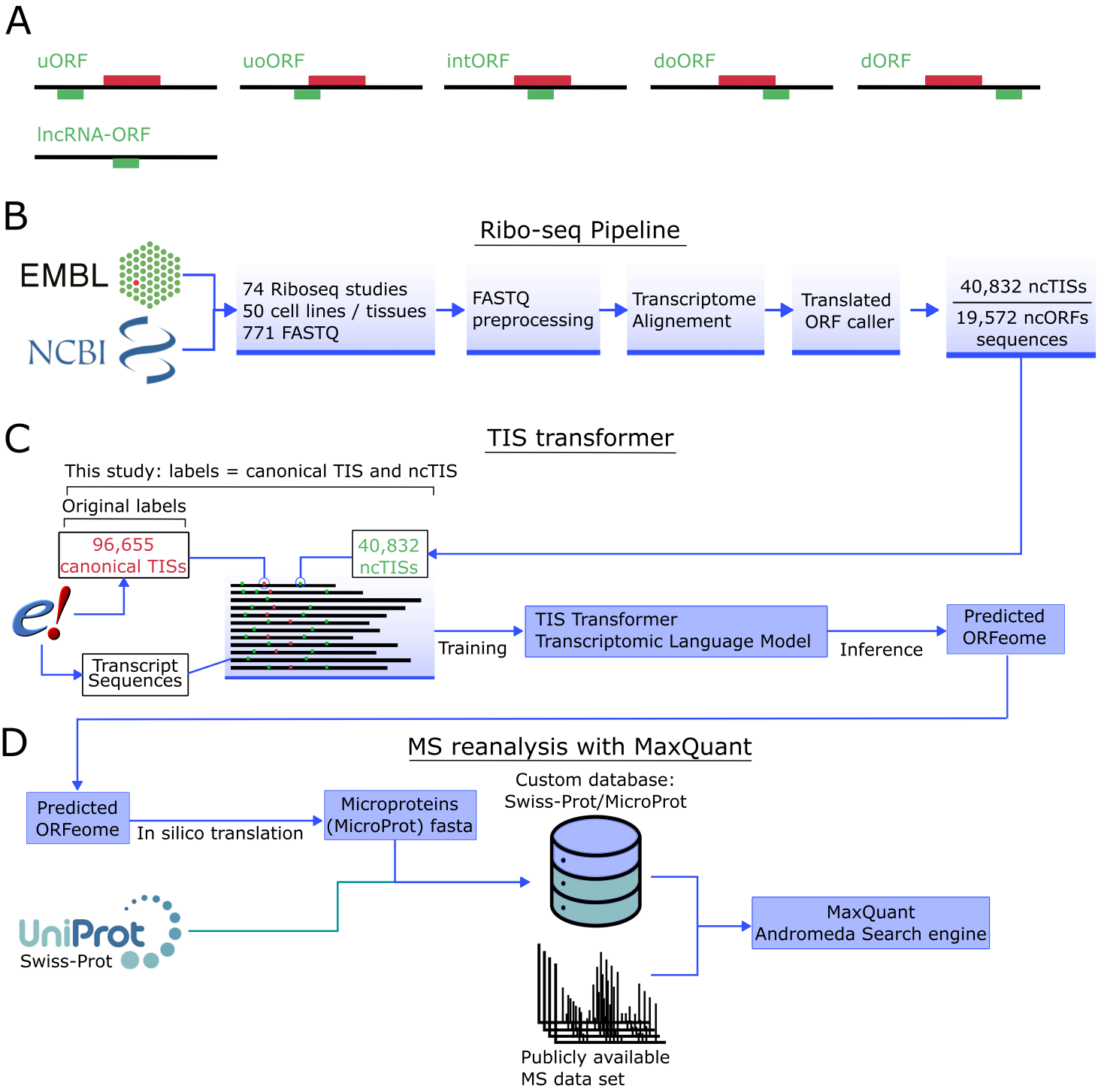


Figure 1

Integration of Ribo-seq canonical and non-canonical TISs (ncTISs) into a transcriptomic language model for the systematic detection of microproteins in a typical MS-based proteomics analysis. (A) Diagram depicting the standardized terminology for non-canonical ORFs (ncORF) types. The canonical CDS and ncORFs are represented by red and green boxes, respectively. uORF, upstream ORF; uoORF, upstream-overlapping ORF; intORF, internal ORF; doORF, downstream overlapping ORF; dORF, downstream ORF; lncRNA ORF, long non-coding ORF. (B) Human experimental ncTISs and ncORFs were identified by reanalyzing 74 ribo-seq studies integrating 108 experiments. As TISs are mapped onto

transcripts and the same ORF can be present on several transcripts originating from the same gene, the number of TISs (40,832) is necessarily higher than the number of ORFs (19,572). (C) TIS Transformer was retrained using Ensembl TISs (red) and ncTISs (green) from (B). (D) A customized Swiss-Prot/MicroProt database was generated and used to reanalyze publicly available MS datasets.

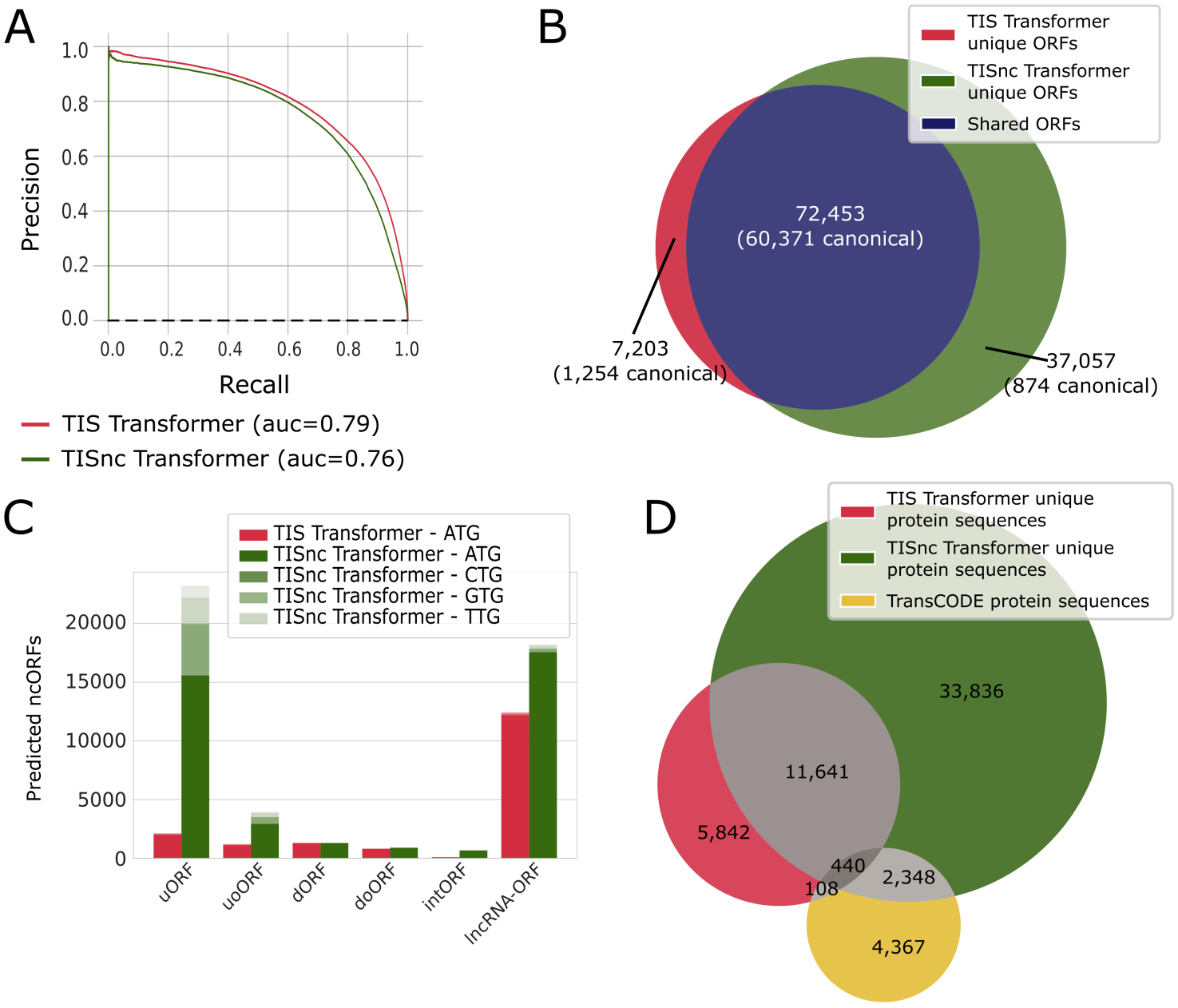


Figure 2

Retraining TIS Transformer with ncTISs improves its ability to predict ncORFs without affecting the prediction of CDSs. (A) Precision-recall curve for TIS Transformer (trained with Ensembl TISs) and TISnc Transformer (trained with Ensembl TISs and 40,832 ncTISs). For an accurate comparison of canonical prediction performance, ncTISs added to the TISnc training set were masked in the PR curve calculation. (B) Proportion of predicted unique ORF sequences common and specific to TIS Transformer and TISnc

Transformer. (C) Predicted ncORFs with ATG or non-ATG TISs. (D) TISnc Transformer is more effective than TIS Transformer at predicting ncORFs annotated in the TransCODE_v35 catalogue.

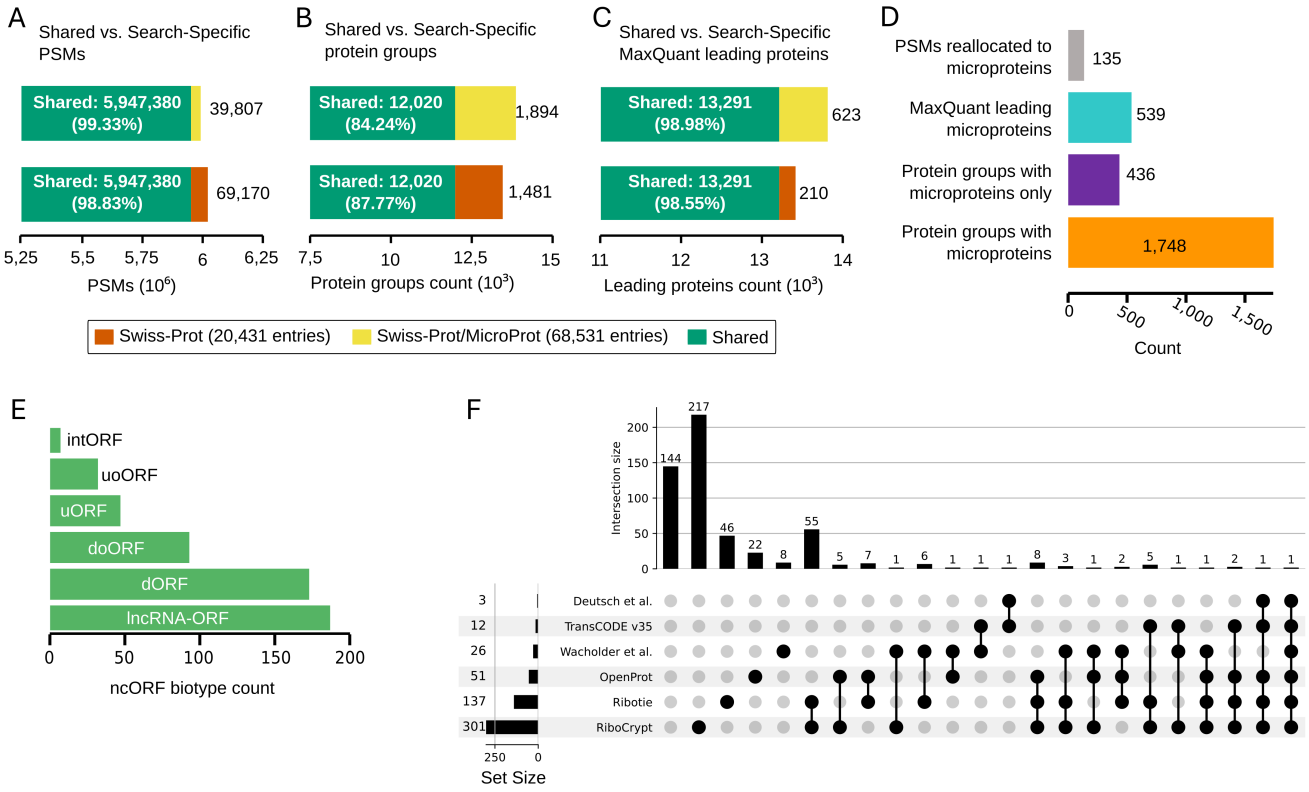


Figure 3

Main layers of information returned by the database search with either Swiss-Prot or Swiss-Prot/MicroProt using MaxQuant (Andromeda) and cross-validation.(A-C) Total PSMs (A), protein groups (B), and MaxQuant leading proteins (C) for each database search, with shared features (green) at the base and search-specific features stacked on top. (D) Four microprotein-specific metrics from the Swiss-Prot/MicroProt search. (E) Number of different ncORF classes encoding the 348 leading microproteins. (F) Cross-validation of the leading microproteins using various proteogenomic resources.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable1.xlsx
- SupplementaryTable2.xlsx
- SupplementaryTable3.csv.xlsx
- SupplementaryTable4.xlsx
- SupplementaryTable5.xlsx

- SUPLEMENTARYFIGURES.docx