frontiers in Genetics

ORIGINAL RESEARCH
published: 25 October 2021
doi: 10.3389/fgene.2021.728900

Identification of Non-Canonical Translation Products in *C. elegans* Using Tandem Mass Spectrometry

Bhavesh S. Parmar[1], Marlies K. R. Peeters[2], Kurt Boonen[3], Ellie C. Clark[1], Geert Baggerman[3], Gerben Menschaert[2] and Liesbet Temmerman[1]**

[1]*Animal Physiology and Neurobiology, University of Leuven (KU Leuven), Leuven, Belgium,* [2]*Laboratory of Bioinformatics and Computational Genomics (BioBix), Department of Mathematical Modelling, Ghent University, Ghent, Belgium,* [3]*Centre for Proteomics (CFP), University of Antwerp, Antwerp, Belgium*

Transcriptome and ribosome sequencing have revealed the existence of many non-canonical transcripts, mainly containing splice variants, ncRNA, sORFs and altORFs. However, identification and characterization of products that may be translated out of these remains a challenge. Addressing this, we here report on 552 non-canonical proteins and splice variants in the model organism *C. elegans* using tandem mass spectrometry. Aided by sequencing-based prediction, we generated a custom proteome database tailored to search for non-canonical translation products of *C. elegans*. Using this database, we mined available mass spectrometric resources of *C. elegans*, from which 51 novel, non-canonical proteins could be identified. Furthermore, we utilized diverse proteomic and peptidomic strategies to detect 40 novel non-canonical proteins in *C. elegans* by LC-TIMS-MS/MS, of which 6 were common with our meta-analysis of existing resources. Together, this permits us to provide a resource with detailed annotation of 467 splice variants and 85 novel proteins mapped onto UTRs, non-coding regions and alternative open reading frames of the *C. elegans* genome.

Keywords: *C. elegans*, altORFs, LC-MS/MS, mass spectrometry, timsTOF, MSFragger, PEAKS, sORFs

OPEN ACCESS

Edited by:
Chi-Ming Wong,
Hong Kong Polytechnic University,
Hong Kong, SAR China

Reviewed by:
Shardul Kulkarni,
The Pennsylvania State University (PSU), United States
Yoshihiro Shimizu,
RIKEN, Japan

****Correspondence:***
Liesbet Temmerman
liesbet.temmerman@kuleuven.be

Specialty section:
This article was submitted to RNA, a section of the journal Frontiers in Genetics

Received: *22 June 2021*
Accepted: *16 September 2021*
Published: *25 October 2021*

Citation:
Parmar BS, Peeters MKR, Boonen K, Clark EC, Baggerman G, Menschaert G and Temmerman L (2021) Identification of Non-Canonical Translation Products in C. elegans *Using Tandem Mass Spectrometry. Front. Genet. 12:728900. doi: 10.3389/fgene.2021.728900*

INTRODUCTION

Translation is a key biochemical process that produces a functional protein out of an open reading frame (ORF). While alternative definitions of an ORF exist (Sieber et al., 2018; Fermin et al., 2006; Claverie, 1997), we here use the term to indicate any mature mRNA sequence contained between a START and a STOP codon. With the advent of high-throughput sequencing and advanced computation, an *ad hoc* rule was established to restrict genomic annotation of ORFs to >100 codons unless previously characterised, as the small ORFs <100 codons (sORFs) posed higher probability of being false positives or biologically meaningless (Basrai et al., 1997). Moreover, most classical genome annotation pipelines enforce a stringent rule for monocistronic annotation of the longest possible ORF within an mRNA, further omitting sORFs and alternative ORFs (altORFs) beyond codon length restriction (Brunet et al., 2018). However, these notions have since been challenged with increasing evidence of non-canonical translation across eukaryotic life (Crowe et al., 2006; Kastenmayer et al., 2006; Ladoukakis et al., 2011).

For assignment of sORFs and altORFS, several bioinformatic and machine learning tools have been developed to predict 3- and 6- frame *in-silico* translation (Omasits et al., 2017; Brunet et al., 2018; Guruceaga et al., 2020). In addition, advances in ribosome sequencing have aided accurate

annotation of codon triplet periodicity and non-AUG start sites, increasing the potential translatome (Mackowiak et al., 2015; Hao et al., 2018; Cesnik et al., 2020). Consequently, ORF annotation has gradually moved from the classical pipeline towards a larger theoretical premise that includes prediction of non-canonical translations from high resolution nucleotide sequencing and efficient signal scoring algorithms. Such ORF annotation pipelines provide searchable databases for discovery proteomics, of which databases following the classical pipeline tend to be more concise than non-canonical sORFs and altORFs prediction databases.

One of the principal tools in proteomics is mass spectrometry (MS), which works in conjunction with proteome databases. In bottom-up proteomics, the mass spectra of (often tryptic) peptides are matched against their *in silico* digested counterparts generated from a database. Under a broader proteogenomic framework, various computational strategies have been developed to integrate proteomic data with (canonical and non-canonical) genomic annotation pipelines or to generate standalone *in silico* translation databases for discovery of novel proteins (Risk et al., 2013; Jagtap et al., 2014; Mackowiak et al., 2015; Nagaraj et al., 2015; Zickmann and Renard, 2015; Kolmogorov et al., 2016; Olexiouk et al., 2016; Brunet et al., 2018; Guillot et al., 2019). At the MS-based experimental front, various fractionation and small protein enrichment methods have been employed to successfully identify novel non-canonical proteins in eukaryotic cell lines and tissues (Ma et al., 2016a; Li et al., 2017; He et al., 2018; Cao et al., 2020; Cardon et al., 2020; Kaulich et al., 2020; Cassidy et al., 2021; Wang et al., 2021). Together, the advances in genome annotation pipelines and high throughput mass spectrometry highlight the importance of both computational and technical approaches to successfully identify novel proteins. With massive data generated from nucleotide sequencing and mass spectrometry, refined annotation of model species genomes has been an ongoing endeavour in upgrading our understanding of molecular processes and discovering missing pieces.

C. elegans is an invertebrate model organism widely used for fundamental biological research (Brenner, 1974). In classical terms it has a well-annotated genome, yet little is known about its non-canonical content (Mackowiak et al., 2015; Casimiro-Soriguer et al., 2020). With more than 1,500 publications annually for the past decade, *C. elegans* research nonetheless benefits from extensive resources of sequencing and mass spectrometry data. The current Ensembl (release 101-August 2020) *C. elegans* annotation comprises approximately 61,000 transcripts classified as protein coding (20,000), non-coding (25,000), or - nonetheless - arising from presumed pseudogenes (2,000) (Yates et al., 2020). Two evidence-mapping pipelines utilizing sequencing-based predictions and *C. elegans* LC-MS/MS datasets identified 9 novel sORFs (Mackowiak et al., 2015) and about a 100 alternative ORFs (Brunet et al., 2018), respectively. However, no dedicated proteomic investigation has focused on profiling the non-canonical proteome in *C. elegans*. To our knowledge, the only focused research of sORFs in *C. elegans* originates from an evolutionary conservation-based genetic screen of intergenic ORFs wherein the authors identified 82 novel proteins (Casimiro-Soriguer et al., 2020).

We here report on the identification of 552 novel proteins within the *C. elegans* proteome. Combining a meta-analysis of available data with mass spectrometry-based discovery of protein extracts prepared especially for this purpose, we present a repertoire of novel *C. elegans* proteome members, including several with orthologs in other model organisms. These results provide a valuable resource for functional biological research using model organisms.

MATERIALS AND METHODS

Construction of allORF Database

To identify unannotated novel proteins in the *C. elegans* proteome, an allORF database was constructed by complementing this organism's reference proteome (Ensembl version 97, downloaded on August 21, 2019) with its alternative proteome predicted by two publicly available repositories, sORFs.org (Olexiouk et al., 2016; Olexiouk et al., 2018a) and Openprot (Brunet et al., 2019).

To obtain the putative coding sequences according to the method in sORFs.org, the public ribo-seq datasets (GSE62859 (Arnold et al., 2014), GSE52910 (Hendriks et al., 2014), GSE67387 (Nedialkova and Leidel, 2015), GSE65948 (Aeschimann et al., 2015)) were downloaded from NCBI Gene Expression Omnibus or NCBI Sequence Read Archive (SRA055804 (Stadler and Fire, 2011), SRA049309 (Stadler et al., 2012)) and processed according to a previously described pipeline (Olexiouk et al., 2018a) with minor adaptations. Here, raw reads were aligned using the STAR splice site aware mapper on the reference genome retrieved from the iGenomes repository with a P-site pinpointed by Plastid (Dunn and Weissman, 2016). After quality assessment with the mQC tool (Verbruggen and Menschaert, 2019), translation initiation sites were delineated using only the data of elongating ribosomes (Olexiouk et al., 2018a). Subsequently, sORFs with a maximum length threshold of 100 codons were assembled by the previously published sORFs.org pipeline with minor code modifications (Olexiouk et al., 2018b). To take the compact genome of *C. elegans* into account, the noise filtering settings were set at alpha = 0.2. Only the longest sORF of candidates which shared a stop site was retained in order to reduce redundancy. Finally, duplicated sequences were removed to obtain the final database of predicted putative sequences based on the sORFs.org method. To reduce the sORF predictions overlapping with known or predicted proteins longer than 100 amino acids, the sequences of the predicted sORFs (excluding the first and the last amino acid) were tested for identical overlap with the protein sequences of the reference Ensembl and the *C. elegans* AltProts and Isoforms from the Openprot repository (Brunet et al., 2019) (release 1.3, downloaded 5th of November 2019) with an in-house scripting module (written in *Python*, available upon request). Finally, the reference Ensembl, the downloaded OpenProt, and the filtered sORFs.org predictions were

concatenated with the cRAP database (downloaded on 6th of November 2019) containing commonly identified contaminants in MS analysis, and the proteome of *E. coli* (strain K12, version 45, downloaded on 6th of November 2019) to account for contaminants introduced by the feeding conditions to obtain the final allORF search database.

Evaluation of allORF Database Characteristics

The features of the proteins present in the different database parts of the final allORF database were analysed using the Peptides package (Osorio et al., 2015) in R 4.0.2 (R Core Team, 2020) and visualized with the ggplot2 package (Wickham, 2016) (**Supplementary Datasheet 5 Figure S1**).

The final allORF database was *in silico* digested by trypsin allowing one missed cleavage and a mass limit between 600 and 4,000 Da, followed by redundancy clearance using the DBToolkit 2.0 (Martens et al., 2005) (version 4.2.5, downloaded on the August 11, 2020). Subsequently, only non-redundant peptides with a length between 7 and 30 amino acids were kept and mapped against the allORF database to identify the number of unique *in silico* peptides per alllORF protein. Finally, the results were visualized with the ggplot2 package (Wickham, 2016).

Worm Culture

All animals (*C. elegans* LSC 1918) were cultured at 20 °C on Nematode Growth Medium plates supplemented with *E. coli* OP50 as described previously (Lewis and Fleming, 1995).

Construction of HiBit Strain LSC1918

A nucleotide sequence coding for the 11 amino acid HiBit tag (gtg agcggctggcgcctgtttaaaaaaattagc) was inserted at the C-terminus of an 84 amino acid annotated protein R06C1.4.1 (strain: LSC 1917) and a predicted 59 amino acid altORF on C05C9.3 (Openprot ID: IP_1,500,296, strain: LSC 1916), using the CRISPR-cas9 system with the *dpy-10* co-CRISPR marker as described previously (Paix et al., 2017); all oligo sequences can be found in **Supplementary Datasheet 5 Table S1**. Briefly, guide RNA was designed based on sequence scoring (Integrated DNA Technologies CRISPR design checker tool) and proximity to the R06C1.4.1 and IP_1,500,296 stop codon. The repair template comprised ~35bp DNA homology arms with the HiBit nucleotide sequence inserted before the stop codon. The injection mix comprised 2.5 µL Cas9 enzyme (15 µg/µL), 2.5 µL tracrRNA (0.17 nmol/µL), 1 µL *dpy-10* crRNA (0.6 nmol/µL), 1 µL R06C1.4.1 or IP_1,500,296 crRNA (0.6 nmol/µL), 1 µL *dpy-10* repair template (0.5 µg/µL), 1 µL R06C1.4.1 or IP_1,500,296 repair template (1 µg/µL). For LSC 1916, injected hermaphrodites were allowed to lay eggs, F1 offspring were singled out based on the roller phenotype characteristic of heterozygous insertion of the *dpy-10* co-CRISPR marker and cultured until sufficient F2 progeny were present. For LSC 1917, injected hermaphrodites were allowed to lay eggs, F1 offspring were singled out based on the dumpy phenotype characteristic of homozygous insertion of the *dpy-10* co-CRISPR marker and cultured until sufficient F2 progeny were present. Each singled out F1 worm from LSC1916 and LSC1917 was then lysed, and PCR verified for HiBit sequence insertion using digestion of the PCR product with Mbil (Thermo fisher, FastDigest) as per the manufacturer's protocol. The HiBit nucleotide sequence contains the MbiI target cleavage sequence, hence, successful digestion is a readout for successful integration of the HiBit repair template. LSC1916 non-roller progeny (F2) of the HiBit-containing roller F1 were singled out to restore the co-CRISPR locus to its wild type allele, cultured, and screened for homologous HiBit insertion using PCR amplification and enzymatic digestion as described above. LSC1916 HiBit insertion strain with wild-type background was then crossed (Fay, 2006) with homozygous LSC1917 HiBit insert with dumpy background and progeny were screened for homologous HiBit insertion on both locus with restored wild-type co-CRISPR locus. The resulting LSC1918 strain was verified by sequencing the genomic region of R06C1.4.1 and C05C9.3 containing the Hibit sequence (**Supplementary Datasheet 5 Table S1**).

Worm Sampling and Protein Extraction

LSC1918 worms were synchronized by standard hypochlorite treatment (Porta-De-La-Riva et al., 2012). Following overnight incubation in S-basal (5.85 g NaCl, 1 g K_2 HPO_4, 6 g KH_2PO_4 in 1 L milliQ), L1 arrested animals were cultured on Nematode Growth Medium plates (Nematode Growth Medium (NGM), 2014) with an *E. coli* OP50 lawn (~3,000 worms/plate, 5 plates/sample), at 20°C for 52 h, which is until the worms had reached the young adult stage. Worms were washed off the plates with S-basal and allowed to settle in 15 ml tubes for 10 min at room temperature. To remove bacteria, the pellet was washed three times with S-basal while allowing the worms to settle for 5 min after each wash. After a final wash with ultrapure (Milli-Q) water, worm pellets were snap frozen in liquid nitrogen and stored at −80°C until further use.

For protein extraction, the frozen pellet was thawed by adding a double volume of lysis buffer (8 M urea, 2 M thiourea, 1 mM dithiothreitol, 1x cOmplete™ protease inhibitor cocktail (Roche)). The mixture was first homogenised in a Precellys-Cryolys homogenizer (Bertin Instruments) using an equal volume of ceramic beads (1.4 mm zirconium oxide, Bertin Technologies) beaten at 6,800 rpm for 10 cycles of 10 s each, with 20 s pause between each cycle at a temperature below 4°C. The homogenised lysate was collected and further sonicated on ice, using a probe sonicator for 12 cycles (5 s ON, 10 s OFF). The lysate was then spun at 16,000 *g* for 30 min at 8°C. The supernatant was collected and protein concentration of samples was estimated using a standard Bradford assay (Harlow and Lane, 2006).

Enrichment of Low Molecular Weight Proteins

Broadly, two strategies were employed for enrichment of low-molecular weight proteins and peptides, *viz.* (C8 and C18) reversed phase chromatography and gel electrophoresis using Tris-Tricine SDS-PAGE. Enriched fractions were either enzymatically digested (with trypsin or chymotrypsin) or loaded undigested onto the mass spectrometer after C18 cleanup. Additionally, 20 µg worm lysate pre-enrichment was

also digested for assessment of enrichment strategies compared to a whole-sample shotgun proteomic approach. All experiments were conducted with four biological replicates.

A) **Whole mount digestion**: 20 µg worm lysate pre-enrichment was reduced with 5 mM dithiothreitol at 56°C for 30 min, alkylated with 25 mM iodoacetamide at room temperature in the dark for 20 min. The volume was adjusted to 1 M urea with 50 mM triethyl ammonium bicarbonate and digested overnight with 1 µg of trypsin (at 37°C) or chymotrpysin (at RT) (Promega BNL, Netherlands). The reaction was then stopped by acidifying the samples to 0.1% formic acid. Following that, the sample was cleaned using C18 spin columns (Pierce™). Briefly, the column was rinsed with 50% methanol, equilibrated 3 times with 200 µL of 5% acetonitrile, 0.1% formic acid by spinning at 1,500 *g* for 1 min. The digested sample was loaded, washed 4 times with 5% acetonitrile, 0.1% formic acid and eluted in 50 µL of 30% acetonitrile and 50 µL of 60% acetonitrile with 0.1% formic acid, and dried using a Savant SpeedVac concentrator.

B) **C8 reversed phase enrichment**: Bond Elut C8 solid phase extraction cartridges (Agilent Technologies, United States) were coupled with a vacuum manifold with the pressure set to 1,000 mbar. The column was first rinsed with 6 ml of 50% methanol and equilibrated thrice with 6 ml buffer A (20 mM ammonium acetate in Milli-Q, pH 7.0). 2.5 mg of protein lysate was diluted 1:7 with buffer A and samples were loaded onto the equilibrated sorbent at room temperature and left undisturbed for 4 min. Following that, the sample was allowed to flow through the sorbent and the column was washed five times with 6 ml buffer A. Bound proteins and peptides were eluted with 3 ml of buffer B (75% acetonitrile in 20 mM ammonium acetate pH 7.0). The eluent was dried in a SpeedVac concentrator (Savant) and redissolved in 100 µL of 50 mM triethyl ammonium bicarbonate. Recovery concentration was estimated using a Bradford assay (Harlow and Lane, 2006). 10 µg of sample were reduced with 5 mM dithiothreitol at 56°C for 30 min and alkylated with 25 mM iodoacetamide at room temperature in the dark for 20 min. Samples were then digested overnight with 0.5 µg of trpysin (at 37°C) or chymotrpysin (at room temperature) (Promega, Netherlands). The reaction was stopped by acidifying the samples to 0.1% formic acid. 20 µg of undigested C8-enriched sample were also acidified to 0.1% formic acid and cleaned to capture peptides not amenable to bottom-up proteomics. The cleanup for digested and undigested peptides was performed as described above (*cf.* A).

C) **Acid precipitation enrichment:** for enrichment of peptides on C18, we adapted the protocol previously published by Secher et al. (2016). 500 µg of lysate were acidified to 0.5% acetic acid, vortexed for 1 h at 4°C and spun at 16,000 *g* for 15 min at 4°C. The supernatant was collected and filtered through a 10,000 Da (Da) molecular weight cut-off filter for 20 min at 4,000 *g* at 4°C (Amicon® Ultra-4 centrifuge filters, Merck Millipore, pre-rinsed twice with 50% methanol). The flow-through was then cleaned on C18 spin columns as described above (*cf.* A).

D) **Tris-tricine SDS PAGE and in-gel digestion**: 200 µg of lysate was run on 16.5% tris-tricine SDS-PAGE (Criterion™, Biorad) at 150 V for 30 min. The gel was washed twice with Milli-Q and the unstained gel was cut using a sterile scalpel to collect the fraction between 2,000 and 12,000 Da according to the Precision Plus Protein™ Dual Xtra Prestained Protein Standards (Bio-Rad). The gel fraction was diced into 1 mm^3 pieces, washed with 500 µL 50% acetonitrile in 25 mM ammonium bicarbonate and dehydrated with 100% acetonitrile by vortexing for 10 min. The dehydrated gel pieces containing proteins were allowed to rehydrate in 25 mM ammonium bicarbonate with 5 mM dithiothreitol and incubated at 56°C for 30 min. Unabsorbed buffer was removed and replaced with 25 mM iodoacetamide in 25 mM ammonium bicarbonate and incubated at room temperature in the dark for 45 min. Following that, the gel pieces were dehydrated again with 100% acetonitrile and rehydrated for 30 min on ice with 200 µL 25 mM ammonium bicarbonate, 10% acetonitrile and 3 µg of trypsin or chymotrypsin (Promega, Netherlands). 25 µL of 25 mM ammonium bicarbonate were added to cover the gel pieces, which were then incubated overnight at 37°C (trypsin) or room temperature (chymotrypsin). The following day, the unabsorbed mixture was collected in a fresh LoBind tube (Eppendorf), and digested peptide was extracted from the gel pieces by vortexing in 300 µL of 80% acetonitrile, 5% formic acid for 30 min. The extract was pooled with the unabsorbed mixture and the sample was dried using a Savant SpeedVac concentrator. The dried peptides were redissolved in 5% acetonitrile, 0.1% formic acid and cleaned as described above (*cf.* A).

LC-MS/MS

The sample was dissolved in 10 µL of 6% ACN and 0.1% FA and separated on a ACQUITY UPLC M-Class System (Waters), fitted with a nanoEase™ M/Z Symmetry C18 trap column (100 Å, 5 µm, 180 µm × 20 mm) and a nanoEase™ M/Z HSS C18 T3 Column (100 Å, 1.8 µm, 75 µm × 250 mm, both from Waters). The sample was loaded onto the trap column in 2 min at 5 µL/min in 94% buffer A, 6% buffer B (buffer A is 0.1% FA in MilliQ, buffer B 0.1% FA in 80% ACN). The flow over the main column was 0.4 µL/min and the column was heated to 40°C. After an isocratic flow of 4 min at 6% B, the concentration of B increased in 36 min–50% B, then to 94% B in 4 min, using linear gradients. After again an isocratic flow of 4 min at 94% B, the concentration of B decreased in 4 min–6% which was followed by 15 min of equilibration at an isocratic flow of 6% B. The column was online with a timsTOF Pro operating in positive ion mode, coupled with a CaptiveSpray ion source (both from Bruker Daltonik GmbH, Bremen). The timsTOF Pro was calibrated according to the manufacturer's guidelines. The temperature of the ion transfer capillary was 180°C. The Parallel Accumulation–Serial Fragmentation DDA method was used to select precursor ions for fragmentation with 1 TIMS-MS scan and 10 PASEF MS/MS scans, as described by Meier et al. (2018). The TIMS-MS survey scan was acquired between 0.70–1.45 V s/cm^2 and 100–1700 m/z with a ramp time of 166 ms. The 10 PASEF scans contained

maximum of 12 MS/MS scans per PASEF scan with a collision energy of 10 eV. Precursors with 1 – 5 charges were selected with the target value set to 20,000 a. u. and intensity threshold to 2,500 a. u. Precursors were dynamically excluded for 0.4 s. The timsTOF Pro was controlled by the OtofControl 5.1 software (Bruker Daltonik GmbH). 10 PASEF scans contained on average 12 MS/MS scans per PASEF scan. Raw data were analysed with the DataAnalysis 5.1 software (Bruker Daltonik). All mass spectrometry raw and spectrum files can be downloaded from MassIVE with identifier MSV000087909.

Data Analysis

For meta-analysis of proteomic datasets of larval, adult and Stress *C. elegans*, raw files were accessed from ProteomeXchange (http://www.proteomexchange.org/) via their respective IDs (PXD006676 (Xia et al., 2018), PXD004584 (Narayan et al., 2016), PXD005649 (Edifizi et al., 2017)). The raw files were then analysed on PEAKS Studio X (Bioinformatics Solutions Inc., Canada) with precursor tolerance set to 10 ppm and a fragment tolerance of 0.02 Da with fully tryptic digestion and 2 allowed missed cleavages. The fixed and variable modifications were set according to the information published in the respective original studies. For experiments conducted in this study using LC-TIMS-MS/MS, data were analysed using MSFragger based Fragpipe (Yu et al., 2020a) and PEAKS Online Xpro (Bioinformatics Solutions Inc., Canada). For C8 enriched, in-gel digested and whole-sample digests, carbamidomethylation (C, +57.02) was set as a fixed modification and carbamidomethylation (DHEK &N-term, +57.02), N-terminal acetylation (+42.01), oxidation (M, +15.99), pyroglutamation (N-term E, -18.01), pyroglutamation (N-term Q, -17.02) and deamidation (NQ, +0.98) were set as variable modifications with only 3 allowed modifications per peptide. The enzyme (trypsin or chymotrypsin) was chosen corresponding to the respective digests, with full specificity. Precursor tolerance was set to 20 ppm and fragment tolerance was set to 0.05 Da. For undigested samples, variable modifications were N-terminal acetylation (+42.01), oxidation (M, +15.99), pyroglutamation (N-term E, -18.01), pyroglutamation (N-term Q, -17.02) and deamidation (NQ, +0.98) with non-specific cleavage and no fixed modification. The d raw files were searched against our *C. elegans* allORF database (**Supplementary Data Sheet 1.FASTA**) on both the search engines and filtered to remove contaminants and non-razor (subgroup) proteins and only top group proteins were considered. Protein identifications with at least 1 peptide at 1% FDR were considered and further analysis was performed in R studio (http://www.rstudio.com/) using custom script for statistics and visualization. For **Figure 3A,B**, the data was fitted into a generalized linear mixed model and significance calculated with Type II Anova. Pairwise posthoc comparison was performed with least square means and Benjamini-Hochberg correction. For interaction statistics for non-canonical identifications (**Figure 4B,D** and **Supplementary Datasheet 5 Figures S2B,D**), superexact test (Wang et al., 2015) was performed. Other statistical tests conducted in this study are mentioned in the respectively text and/or caption along with resulting *p*-values.

Custom BLASTp Candidate List

A local BLAST search database was created with the BLAST + application (Camacho et al., 2009) using the alternative proteome of four model organisms, namely fruit fly, human, house mouse and zebrafish, downloaded from Openprot (release 1.3, downloaded on 6th of September 2018), sORFs.org (downloaded on 6th of September 2018) and described by Mackowiak et al. (2015). The BLASTp algorithm was applied to search the candidate list against the individual search database of each model organism. For every search, only the hit with the lowest Expect value (minimum E = 10^{-10}) and highest sequence identity per model and database was retained by manual inspection (**Supplementary Datasheet 3**).

RESULTS

Construction of a *C. elegans* allORF Database

In this work, we set out to expand and describe the *C. elegans* non-canonical proteome. We built a custom database to provide a comprehensive proteomic search space, encompassing all theoretically possible translational outcomes based on available transcript and ribosome sequencing data. To that end, we used predictions from Openprot and sORFs.org concatenated with the Ensembl *C. elegans* proteome. Openprot predictions comprise altORFs and isoforms without codon length cut-off, whereas sORFs.org contributed to prediction of sORFs <100 codons. The resulting allORF database in total comprises 137,194

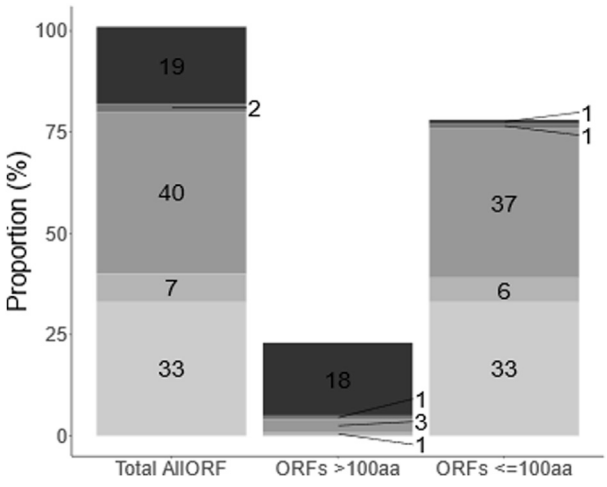

FIGURE 1 | Composition of allORF database. The allORF database combines sequence information from Ensembl, Openprot and sORFs.org. (From top to bottom: Ensembl (19), Ensembl + sORFs.org (2), OpenProt (40), OpenProt + sORFs.org (7), sORFs.org (33). OpenProt + Ensembl + sORFs.org (3 ORFs on a total of 137,166 ORFs) and OpenProt + Ensembl (25 ORFs on a total of 137,166 ORFs) are not shown due to the low number). Numbers indicate the percentage of ORFs in the allORF database retrieved from each sequence source.) Second and third bar show the same information grouped proteins (>100 codons) and small (<100 codons) ORF types.

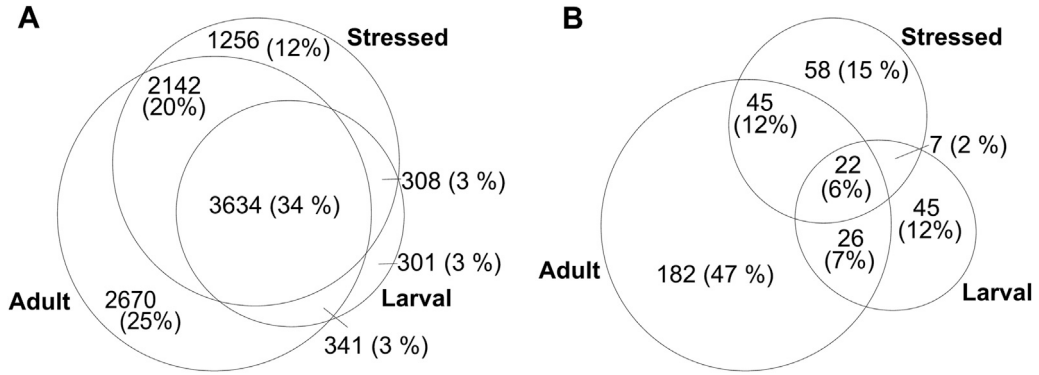

FIGURE 2 | Numbers of identified proteins (grouped) in larval Xia et al., 2018, adult Narayan et al., 2016 and stressed Edifizi et al., 2017 *C. elegans* LC-MS/MS datasets, with at least 1 unique peptide at 1% FDR. **(A)** Total number of unique identifications **(B)** non-canonical proteins as predicted by Openprot and sORFs.org.

protein sequences (**Figure 1** and **Supplementary Datasheet 5 Table S2**), which for this species is about five times more than the Ensembl data alone (25,886 protein sequences). Briefly, about 77% of the allORF database comprises small (<100 codons) ORF predictions from Openprot and sORFs.org, whereas 2% of the database are sORFs that originate from Ensembl. Additionally, about 18% of all ORFs are >100 codons and retrieved from Ensembl, with the remaining 4% also longer than 100 codons but derived from Openprot and sORFs.org (**Figure 3** and **Supplementary Datasheet 5 Table S2**). For allORF database assembly, the calculated median ORF length within the sORFs.org predictions was 25 codons whereas for Openprot predictions, this was 45 codons (**Supplementary Datasheet 5 Figure S1**). With about 70% of allORF sequences unique to either one of these two prediction algorithms, the newly assembled database efficiently exploits the complementarity of Ensembl, sORFs.org and Openprot.

The allORF database unveils 385 novel non-canonical proteins in published proteomics data of *C. elegans*

Despite stringent concatenation of redundancy in the allORF database, we anticipated that the sheer number of entries might increase false positives in proteomic mapping. To test this, we utilized available *C. elegans* raw data from Xia et al. (2018) (larval - PXD006676), Narayan et al. (2016) (adult - PXD004584) and Edifizi et al. (2017) (stressed - PXD005649) and re-analysed over 120 raw files of 240 min gradient runs each against our allORF database using PEAKS. These datasets are interesting from a discovery perspective because they cover a range of physiological changes within the worm: the larval dataset contains samples of all 4 larval stages (L1-L4), the adult dataset includes day 1, day 5 and day 10 adult worms, and stressed samples were subjected to UV-irradiation or starvation. In total, 10,651 proteins were identified across all three datasets with at least one unique peptide at 1% FDR (**Supplementary Datasheet 1**). This is in the same range as originally identified in the respective studies, with minor differences due to differences in database, analysis pipeline and because for the adult dataset, we only considered one biological replicate (with 80 LC-MS/MS runs).

About 82.5% of the proteins were identified in the adult-stage dataset (**Figure 2A**). Although there is a clear complementarity of the diverse *C. elegans* sampling conditions that benefits discovery of novel proteins, this observation is likely due to technical differences. The adult dataset comprises five extraction methods and 80 fractions as compared to a shotgun approach used for the larval and Stress datasets, increasing the depth of identification in the adult *C. elegans* dataset. Of the 10,651 proteins in total, 385 identifications relied on either Openprot or sORFs.org contributions to the allORF database (**Figure 2B**). 85% of these non-canonical identifications categorize as non-canonical isoforms, whereas the remaining 51 novel proteins originate from genomic regions such as UTRs, ncRNA and polycistronic mRNA with alternative open reading frames (**Supplementary Datasheet 3**).

The experimental data used here were retrieved from studies that followed well-established bottom-up proteomic sample preparation strategies with depth achieved mainly by peptide fractionation (adult dataset) and long liquid chromatographic separation coupled online with orbitrap-based mass spectrometric detection. Thus, this meta-analysis of available data provided an efficient means for testing the efficacy of ORF predictions in our database, and of existing proteomic strategies to capture novel proteins.

Enrichment Strategies and Alternative Cleavage for Shotgun Mass Spectrometry Analysis Are Complementary to Standard Proteomic Digest

Most non-canonical predictions reside in the low molecular weight range, with median length of 25 amino acids for sORFs.org and 45 amino acids for OpenProt predictions (**Figure 1** and **Supplementary Datasheet 5 Figure 1**). To evaluate whether enrichment strategies could be beneficial for cost and time-efficient shotgun identification of novel non-canonical proteins, we combined eight different strategies with LC-TIMS-MS/MS (**Supplementary Datasheet 5 Table S3**). Next to using strategies

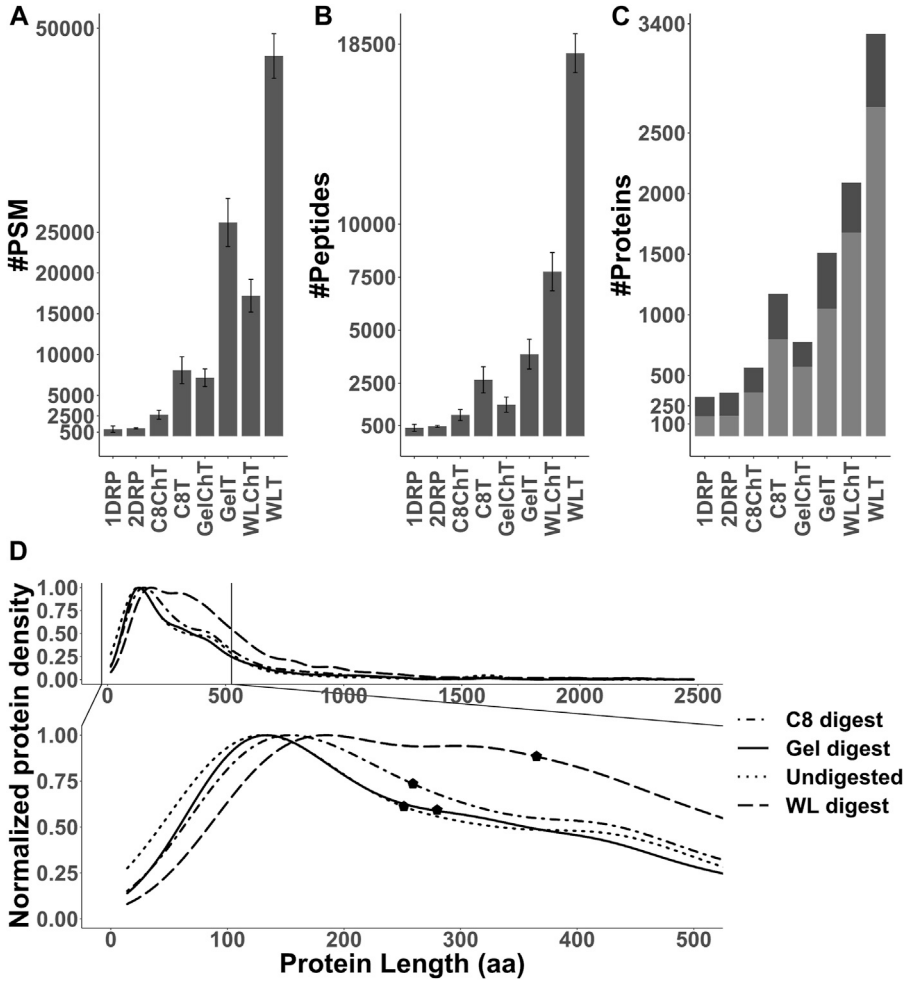

FIGURE 3 | (A,B) Number of peptide spectrum matches (PSM) and peptides identified for each sample processing method used in this study (p-value <0.001). Error bars indicate standard deviation between replicates (n = 4). **(C)** Total number of proteins for all replicates combined, with unique (stripped) peptides >1 (light gray) or = 1 (dark gray) per condition. **(D)** Distribution of identified proteins and their length normalized to the total of identifications per group and median markers ($p_{Fligner\text{-}Killeen}$: 2e-16).

described in literature, *viz.* Tris-tricine SDS-PAGE (Wang et al., 2021) and C8-reversed phase (Ma et al., 2016a) enrichments, we enzymatically digested the whole worm lysate under denaturing conditions. To further improve the protein sequence coverage and the expected depth of our analysis, we utilized complementary cleavage enzymes, trypsin and chymotrypsin, in combination with the three strategies (conditions named GelT, GelChT, C8T, C8ChT, WLT, WLChT henceforward). Finally, we also included undigested peptidome analysis of C8-reversed phase enriched samples (henceforward 2DRP) and an adaptation of the method by Secher et al. (2016), which aims at capturing endogenous peptides and peptides not amenable to enzymatic digestion due to their small size and lack of cleavage site (henceforward 1DRP).

Out of the 8 strategies, the undigested fractions (1DRP, 2DRP) aimed at capturing endogenous peptides led to the least number of peptide spectrum matches (PSMs), peptides and protein identifications (**Figures 3A–C**). In the range of 2.5–12 kDa proteins, numbers from C8-enrichment and digestion with trypsin and chymotrypsin (C8T, C8ChT) were superseded by those of in-gel digestion (GelT, GelChT), indicating that a well-optimized in-gel digestion method is far more efficient at size separation of proteins compared to a standard C8 reversed-phase solid phase extraction (**Figures 3A,B**, all p-values < 0.001). As anticipated, the whole worm lysate digest (WLT, WLChT) yielded maximum identifications in total, whereas relatively lower numbers of proteins were matched from enriched samples (**Figure 3C**, **4A**). Compared to whole lysate, the C8 digests, gel digests and undigested samples performed relatively better at targeting proteins below 150 amino acids (**Figure 3D**, $p_{Fligner\text{-}Killeen}$: 2e-16). With WL, Gel, C8 and undigested runs combined, a total of 3,969 proteins groups were identified (**Supplementary Datasheet 2**). The WL digest (Tryptic + Chymotryptic) accounts for >90% of these identifications (**Figure 4A**) while only less than 10% were derived uniquely from enriched and digested samples. However, this ratio is increased more than 2-fold for non-canonical identifications, with 22.7% uniquely contributed by the total of all enrichment methods (**Figure 4B**). Enriched samples do contribute unique

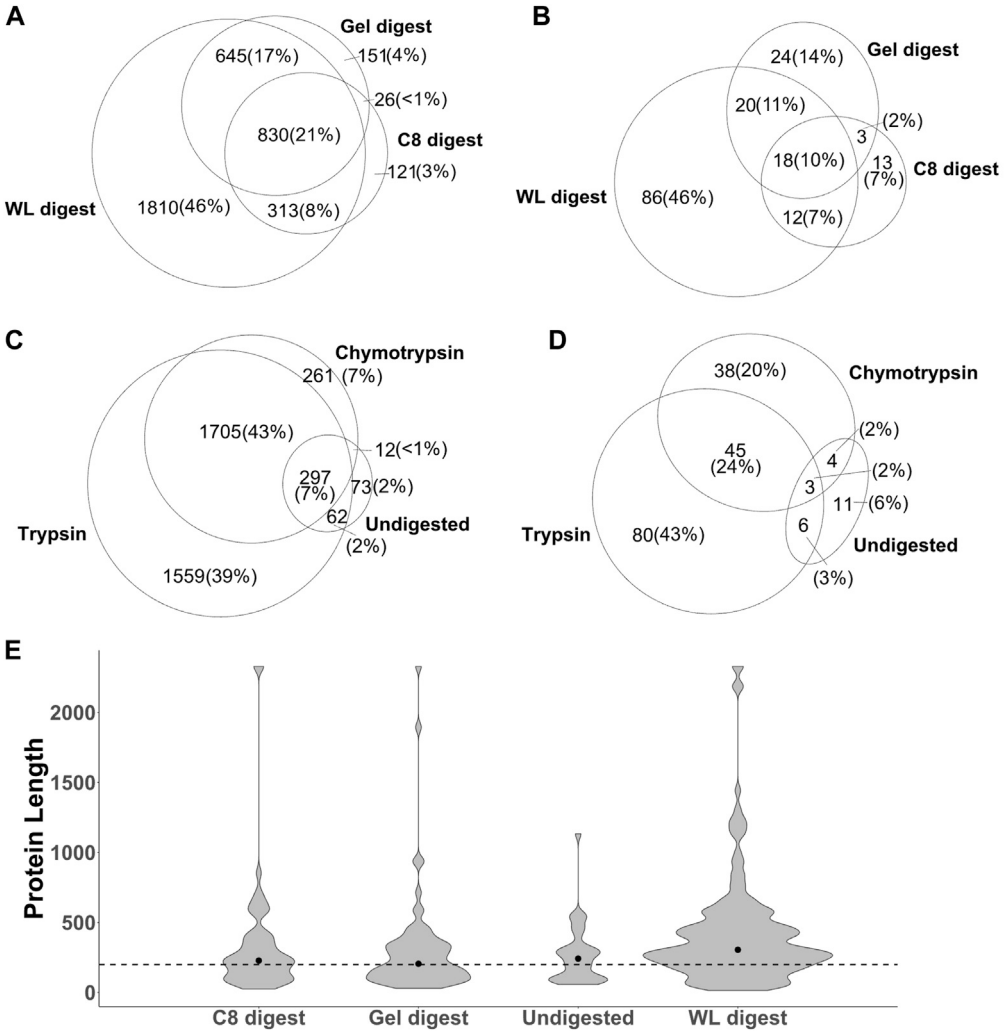

FIGURE 4 | The identification of total proteins **(A,C)** and non-canonical proteins (**(B,D)** p-value: <0.001) is influenced by sample processing methods **(A,B)** and enzymatic cleavage **(C,D)**. WL: whole lysate, C8: C8-reversed phase enriched, Gel: in-gel digested samples, Undigested: 1DRP + 2DRP. **(F)** Absolute count of non-canonical proteins identified in each group and distribution along protein amino acids length with dotted line at 200 amino acids and median represented by a dot.

identifications, indicating that these methods may add distinctive complementarities to the WL digests. Our data also suggest that enzymatic digestion choices may influence peptide identification: undigested samples uniquely contributed to only 2% and chymotrypsin-digested to 6.8% of the total number of identifications, meaning less than 10% of the total went undetected in their tryptic counterparts (**Figure 4C**). However, chymotrypsin-treated and undigested peptidomes accounted for 53 unique identifications of the total 187 novel non-canonical proteins identified across all groups, underscoring the added value of experimental diversity in discovery approaches (**Figure 4D**). In absolute numbers, the novel non-canonical proteins identified across all conditions span a wide protein length range with median close to 200 amino acids for the total of all enriched samples (**Figure 2E**), with a significant distinction in non-canonical proteome coverage depending on the sample type (**Supplementary Datasheet 5 Figure S2E**, $p_{Kruskal\text{-}Wallis}$ 0.01). Together, enrichment for low molecular weight proteins and an alternative cleavage strategy contributed to a quarter of all novel non-canonical proteins identified in our experiments. Moreover, trypsin - the workhorse enzyme for bottom-up proteomics - outperforms chymotrypsin in the number of peptides and protein identifications across sample processing conditions (**Figure 3B,C**), but each enzyme also contributes unique identifications of novel non-canonical peptides to our resource (**Figure 4D**). Enrichment strategies facilitated the identification of an additional 305 proteins (8.4% of total) for tryptic (**Supplementary Datasheet 5 Figure S2A**) and 185 proteins (8.1% of total) for chymotryptic (**Supplementary Datasheet 5 Figure S2B**) digests, that were missed in their whole lysate digested counterparts. The advantage of enrichment with SDS-PAGE and C8-reversed phase is more prominent for non-canonical identifications, with a distinct

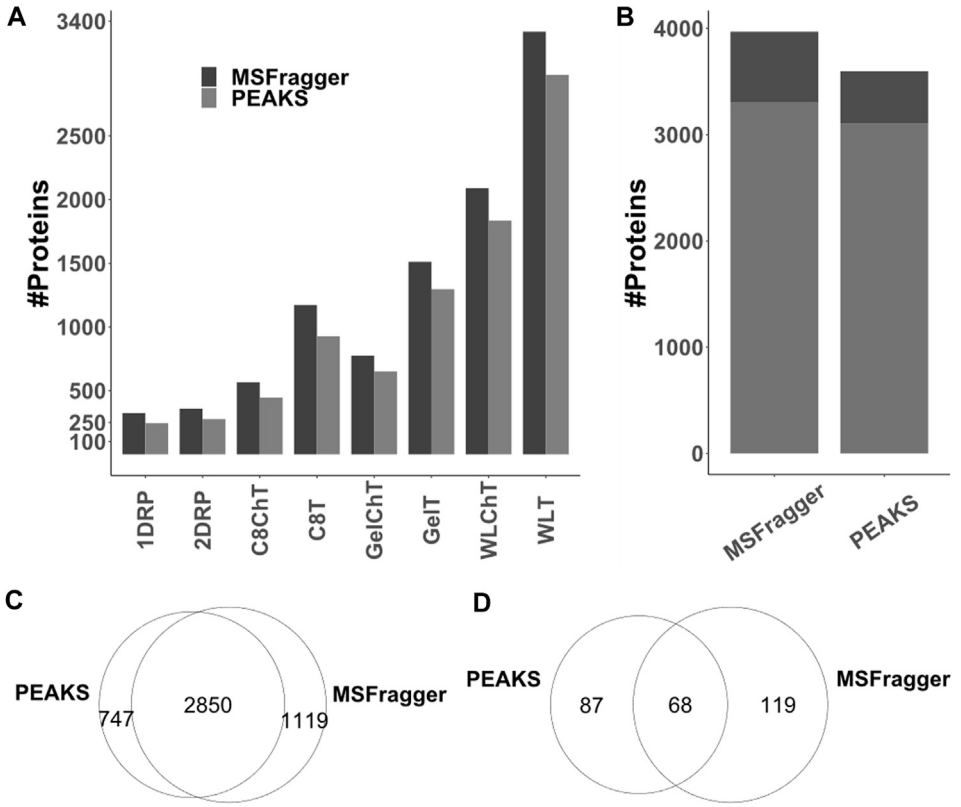

FIGURE 5 | Comparison of PEAKS and MSFragger protein identifications for LC-TIMS-MS/MS data acquired in this study **(A)** total protein numbers as identified per condition **(B)** combined total protein numbers across all conditions for proteins with unique (stripped) peptides >1 (light gray) and >1 (dark gray). **(C,D)** While there is certainly overlap, each algorithm contributes unique identifications to the list of total **(C)** and non-canonical **(D)** proteins.

complementarity achieved in either case in comparison to whole lysate digests (**Supplementary Datasheet 5 Figures S2C,D**).

Alternative Search Engines Further Improve the Detection of Novel Proteins

Multiple peptide search algorithms exist to aid efficient mapping of experimental LC-MS/MS data against proteome databases; however, underlying principles and data training may differ widely. For example, MSFragger (Kong et al., 2017) employs a fragment ion indexing method and supports open searches for mapping peptides onto a database, whereas PEAKS (Zhang et al., 2012) generates *de novo* tags from MS data for peptide identification and prediction of post-translational modifications. To test whether differences in algorithm would affect the identification success, hence discovery of novel peptides, we processed the same raw files (*i.e.,* 32 × 60 min runs) already analysed via MSFragger (*cf.* above) with PEAKS Online Xpro using the same parameters. Across all sample processing conditions, the number PSMs (**Supplementary Datasheet 5 Figure S3A**) and unique peptide sequences (**Supplementary Datasheet 5 Figure S3B**) identified by MSFragger and PEAKS differed slightly. However, MSFragger consistently identified more proteins across all sample processing conditions (**Figure 5A**, **Supplementary Datasheet 5 Table S4**). For combined results from all the experiments in this study, the difference in sheer total number of identified proteins between the two search engines is 372 (approximately 10%) with relatively more proteins identified based on only 1 unique peptide by MSFragger (**Figure 5B**, dark gray). A huge difference is observed in distinct protein identifications reported by the two search algorithms (**Figure 5C**) with 747 and 1,119 unique proteins with at least 1 peptide identified by PEAKS *vs* MSFragger, respectively. Similarly, 87 unique non-canonical proteins were added by PEAKS to our previous list of 187 from MSFragger, with 68 common between the two search algorithms, taking the total count of non-canonical proteins identified in this study to 274 (**Figure 5D**). As was the case for complementary sample processing, a complementarity is observed in data analysis pipelines as well. Although similar strategies have been utilized previously and search engines do perform differently (Shteynberg et al., 2013), the differences observed in our study amount to nearly 40% of the total protein identifications that are unique to either of the search engines. This suggests that efficient spectrum matching algorithms are yet to be standardized for search engine-wide consistency and the combination of search engines is likely to be more informative.

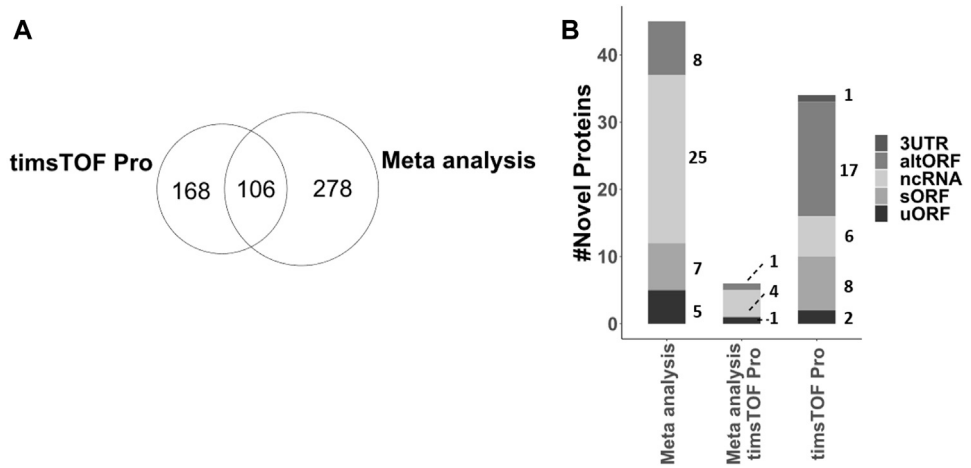

FIGURE 6 | (A) Comparison of all 552 non-canonical proteins, including 467 splice variants and 85 novel proteins, identified from meta-analysis of larval, adult and Stress proteomic datasets acquired on Orbitrap platform and experiments conducted in this study on timsTOF Pro platform. **(B)** Annotation of 85 novel proteins identified either via meta-analysis, timsTOF Pro acquisition of our experiments or both, mapped on C. elegans genome.

Annotation of Identified Novel Non-Canonical Proteins

Using a custom *C. elegans* database, existing *C. elegans* LC-MS/MS data and experiments aimed at capturing non-canonical proteins, we report the identification of 552 non-canonical proteins based on predictions from Openprot and sORFs.org. 324 out of these were identified with more than 1 unique peptide, 42 with 1 unique peptide but across multiple samples, and 186 with only 1 unique peptide. Of the total of non-canonical proteins identified in this study, 106 identifications were common between the samples analysed with timsTOF Pro and the meta-analysis performed on three proteomic datasets (acquired via Orbitrap platform), while 168 proteins were unique to our experiments and 278 mined from pre-existing *C. elegans* datasets via our meta-analysis (**Figure 6A**). 467 of these proteins correspond to low-novelty isoforms/splice variants. From the remaining 85, 8 were mapped onto 5′UTR (uORF) of annotated genes, 1 in 3′UTR, 35 in non-coding RNA (ncRNA), and 41 are alternative polycistronic translation products from annotated genes (15 sORF and 26 altORF respectively, **Supplementary Datasheet 3**). This group represents the novel proteins that are crucial for further investigation and functional annotation. To further assess the significance of this group from a model organism perspective, we searched these 85 candidates against the human and three other model organism proteomes (*H. sapiens*, *M. musculus*, *D. melanogaster* and *D. rerio*) via a custom BLASTp script. 18 out of 85 were found to be conserved across these species (**Supplementary Datasheet 3**). Together with the less-conserved novel proteins and protein variants, these non-canonical proteins add to the proteogenomic repertoire of interest for functional genetics research.

DISCUSSION

Protein translation is a major driver of organismal phenotypes and plasticity, and for decades now, breakthroughs in all domains of biological sciences have relied on understanding (differences in) translation - hence, indirectly on genome annotation. However, genome annotation is mired with inaccuracies and many genomes, as well as the process of annotation itself, are iteratively updated based on new evidence. In recent years, increased interests in non-canonical translation products such as sORF- and ncRNA-encoded peptides or proteins have motivated researchers to revisit the full complexity of functional genome annotation. Given the anomalous nature of their coding potential and small size, non-canonical ORFs tend to escape the annotation radar. This is also true for the otherwise well-annotated genome of the model organism *Caenorhabditis elegans*. In an effort towards unveiling non-canonical translation products in *C. elegans*, we therefore combined omics strategies to build a resource of reliably detectable non-canonical translations. In general, our efforts acted at three conceptual levels: tailoring searchable databases towards the needs of non-canonical identification, diversifying sample preparation in order to capture non-canonical translations, and high-end mass spectrometric detection aided by different spectrum matching algorithms to maximize discovery.

At the database level, we relied on concatenation of sequencing-based predictions with the *C. elegans* Ensembl proteome (**Figure 1**). As such, and in line with similar strategies used in other studies (Chu et al., 2015; Ma et al., 2016b; Budamgunta et al., 2018; Wang et al., 2021), our custom database provides a less biased search space for downstream LC-MS/MS peptide mapping than the proteome database alone, while remaining within reasonable limits of the genome as determined by the transcriptome. However, the percentage of positives is likely to be lower as compared to a well-curated reference proteome such as Ensembl or Uniprot. The issue of theoretical peptide search space size has been addressed before (Borges et al., 2013; NesvizhskiiProteogenomics, 2014; Chatterjee et al., 2016)

and with machine learning algorithms these are likely to be circumvented in the near future (Chatterjee et al., 2016; Bouwmeester et al., 2020).

Our results show that a custom database tailored from sequencing-based predictions has a significant impact on novel protein discovery. Using existing *C. elegans* deep proteome datasets in combination with our custom allORF database, we could identify 385 new proteins (**Figure 2B**) across larval (Xia et al., 2018), adult (Narayan et al., 2016) and stressed (Edifizi et al., 2017) datasets. With extensive MS analysis covering wide physiological stages in *C. elegans*, the number of non-canonical proteins identified was only 0.3% of total predictions, whereas 40% of the Ensembl reference proteome was identified in our meta-analysis. This highlights the challenges in proteome coverage for non-canonical translation products under standard tryptic digest. Furthermore, based on *in silico* digest, 20% of the non-canonical predictions in our allORF database are not expected to produce unique tryptic peptides. Additionally, ribosome sequencing is also prone to detecting regulatory RNA-ribosome interactions that do not correspond to translation of a functional protein, therefore adding spurious predictions to our database (Verbruggen et al., 2021; Ingolia, 2016; Raj et al., 2016). Thus, using existing mass spectrometric tools, standard proteomic sample preparation and a custom database of non-canonical ORF prediction, it is feasible to identify novel proteins, and revisiting available data can be a valuable effort from a discovery perspective.

To test whether the technical pitfalls of standard proteomic workflow could be overcome, we investigated prominent low-MW protein enrichment strategies and alternative digestion against standard tryptic digest (*cf.* Results). Since the median length for predicted sORFs (which comprises approximately 40% of our allORF database) was 25 amino acids (**Supplementary Datasheet 5 Figure S1**), we expected our own samples relying on endogenous peptidome enrichment strategies (1DRP,2DRP) to yield maximum novel identifications. However, only 24 non-canonical proteins were identified (**Figures 4D,E**) with a varying length distribution (**Supplementary Datasheet 5 Figure S2E**) instead of expected <25 amino acids enrichment. Overall, the data from peptidome analysis points towards three possibilities; *1)* sORF-encoded peptides <25 amino acids might be labile, *2)* low abundant and/or *3)* the extraction and enrichment of this group of peptides begs further investigation.

From a discovery perspective, combining different sample preparation methods with enzymatic digestion worked well in our hands. For digested samples, we observed comparatively higher number of PSMs, peptides and proteins identified with tryptic digests (**Figure 3A–C**, **Figure 4C**). However, chymotryptic digests identified 42 unique non-canonical proteins that were missed in tryptic digests (**Figure 4D**). Although more non-canonical proteins are identified in standard digests (WL) as compared to enrichments aimed at targeting 2–12 kDa proteins, a clear advantage is observed with SDS-PAGE and C8-SPE enrichment with unique identifications (**Figure 4B**), highlighting the complementarity of different sample processing approaches. The non-canonical proteins identified in all sample processing methods span a wide protein length range (**Figure 4E**, **Supplementary Datasheet 5 Figure S2E**) although, 96% of the Openprot and sORFs.org predictions are below 100 amino acids (**Figure 1**). Moreover, we also identified non-canonical proteins of higher MW in the low-MW enriched samples (**Figure 4E**, **Supplementary Datasheet 5 Figure S2E**), which could be due to protein instability *in vitro* or *in vivo*. In all, our data supports that alternative approaches for sample processing are complementary to classical tryptic digest in the detection of novel proteins, however, efficient enrichment of endogenous polypeptides below 100 amino acids remains a challenge for the bulk of predicted sORF and altORF annotations. Some studies have used extensive sample fractionation strategies and top-down mass spectrometry (Cardon et al., 2020; Cassidy et al., 2021) to detect non-canonical proteins, however, that is likely to raise the experimental cost and duration by multiple folds. To further chisel the enrichment of low-MW proteins for cost-efficient shotgun proteomics, we generated a strain (LSC 1918) with two HiBit-tag (Schwinn et al., 2020) knock-ins, one at the C-terminus of R06C1.4.1 (84 + 11 amino acids) and one at the predicted C05C9.3 altORF IP_1,500,296 (59 + 11 amino acids). This strain was used for shotgun experiments in this study, and we envision it to be resourceful for further optimization of enrichment for low-MW proteins in the future. Using a blotting system in combination with this strain, sample preparation strategies may be compared for their performance in the low-MW range prior to moving towards LC-MS/MS for unbiased discovery and identification.

Protein identifications are also influenced by the instrumentation and downstream analysis pipeline, and this was the third conceptual level of interest in our current study. Previous work by Shteynberg et al. already highlighted the advantages of combining multiple search algorithms to improve total peptide spectral matches (Shteynberg et al., 2013). Moreover, the recent development of trapped ion mobility coupled with parallel accumulation serial fragmentation on timsTOF platforms (used in this study) have also extended the sensitivity and depth of mass spectrometric data multifold (Meier et al., 2018). However, the downstream analysis pipelines are yet to catch up with the instrumental advances and, to our knowledge, only three analysis pipelines can process the raw files generated by LC-TIMS-MS/MS proteomic experiments (Yu et al., 2020b). We utilized PEAKS and MSFragger to analyse the same data acquired on a timsTOF Pro platform and observed substantial differences in protein identifications (**Figure 5**). This could be due to various factors, as the underlying algorithms of both search engines differ considerably, however, in combination, that gives an advantage for identification of peptides that might be missed otherwise.

In conclusion, thanks to optimisations at the database, peptide extraction and analysis levels of an omics-based discovery strategy, we can provide mass spectrometric evidence of 467 putative splice variants and 85 novel proteins in *C. elegans*. 18 of these novel proteins were found to be conserved across vertebrates, of which 14 are annotated as ncRNA, 2 mapped onto 5′UTR (uORFs) and 2 are in alternative ORFs in *C. elegans* (**Supplementary Datasheet 3**,

WormBase release WS280). This highlights how genomic annotation can be improved with proteogenomic strategies. These 18 proteins have annotated paralogs in *C. elegans*, indicative of a shared genetic ancestry, however, their functional relevance remains to be investigated. Utilization of sequencing (RNA and Ribosome) in conjunction with proteomics/peptidomics as presented here is likely to further contribute to species-wide genome annotation and our understanding of genetic divergence and compensation. Interestingly, a total of 8 novel proteins belong to uORFs, a regulatory class of proteins (Chew et al., 2016; Johnstone et al., 2016; Zhang et al., 2019) that we propose to define as PEU family (Proteins Encoded in uORFs) for *C. elegans*. Question remains whether living systems tend to follow the robust canonical translation and the non-canonical translations, for large part, are mere slips through reading frames (Verheggen et al., 2017; Chen et al., 2020). However, based on some noteworthy discoveries (Pauli et al., 2014; Anderson et al., 2015; Makarewich et al., 2018; Rathore et al., 2018; Chu et al., 2019; Na et al., 2020), it is certainly clear that nature does not always conform to canonical rules, with discrete prevalence of non-canonical translations, either camouflaged as isoforms/splice variants or anomalous translation of presumed ncRNA, UTRs and polycistronic alternative ORFs. Based on resources like the one presented here, the coming years will certainly witness an increase in functional research into non-canonical translation products and their contribution to organismal phenotypes and plasticity.

DATA AVAILABILITY STATEMENT

All mass spectrometry raw and spectrum files that are part of this study are available from the MassIVE online repository with identifier MSV000087909.

AUTHOR CONTRIBUTIONS

BP contributed to the conception, design of the study, experiments, data analysis, manuscript writing and revision. MP contributed to database construction and manuscript writing/revision. KB contributed to experiments and manuscript revision. EC contributed to experiments. GM, GB and LT contributed to conception and manuscript revision.

FUNDING

This work was funded by FWO grant G052217N and KU Leuven grant C16/19/003. The funders had no role in study design, data collection or analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

LSC1918 was generated based on the N2 wild type strain as provided by the CGC, which is funded by NIH Office of Research Infrastructure Programs (P40 OD010440). We are grateful to N. De Fruyt (KU Leuven) and B. Driesschaert (KU Leuven) for help with statistics.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.728900/full#supplementary-material

REFERENCES

Aeschimann, F., Xiong, J., Arnold, A., Dieterich, C., and Großhans, H. (2015). Transcriptome-Wide Measurement of Ribosomal Occupancy by Ribosome Profiling. *Methods* 85, 75–89. doi:10.1016/j.ymeth.2015.06.013

Anderson, D. M., Anderson, K. M., Chang, C.-L., Makarewich, C. A., Nelson, B. R., McAnally, J. R., et al. (2015). A Micropeptide Encoded by a Putative Long Noncoding RNA Regulates Muscle Performance. *Cell* 160 (4), 595–606. doi:10.1016/j.cell.2015.01.009

Arnold, A., Rahman, M. M., Lee, M. C., Muehlhaeusser, S., Katic, I., Gaidatzis, D., et al. (2014). Functional Characterization of *C. elegans* Y-Box-Binding Proteins Reveals Tissue-specific Functions and a Critical Role in the Formation of Polysomes. *Nucleic Acids Res.* 42 (21), 13353–13369. doi:10.1093/nar/gku1077

Basrai, M. A., Hieter, P., and Boeke, J. D. (1997). Small Open Reading Frames: Beautiful Needles in the Haystack, *Genome Research*., 7. , 1997 Cold Spring Harbor Laboratory Press August 1, 768–771. doi:10.1101/gr.7.8.768

Borges, D., Perez-Riverol, Y., Nogueira, F. C. S., Domont, G. B., Noda, J., Da Veiga Leprevost, F., et al. (2013). Effectively Addressing Complex Proteomic Search Spaces with Peptide Spectrum Matching. *Bioinformatics* 29 (10), 1343–1344. doi:10.1093/bioinformatics/btt106

Bouwmeester, R., Gabriels, R., Van Den Bossche, T., Martens, L., and Degroeve, S. (2020). The Age of Data-Driven Proteomics: How Machine Learning Enables Novel Workflows. *Proteomics* 20 (21–22), 1900351. doi:10.1002/pmic.201900351

Brenner, S. (1974). The Genetics of Caenorhabditis Elegans. *Genetics* 77 (1), 71–94. doi:10.1093/genetics/77.1.71

Brunet, M. A., Brunelle, M., Lucier, J.-F., Delcourt, V., Levesque, M., Grenier, F., et al. (2018). OpenProt: A More Comprehensive Guide to Explore Eukaryotic Coding Potential and Proteomes. *Nucleic Acids Res.* 47, 403–410. doi:10.1093/nar/gky936

Brunet, M. A., Brunelle, M., Lucier, J.-F., Delcourt, V., Levesque, M., Grenier, F., et al. (2019). OpenProt: A More Comprehensive Guide to Explore Eukaryotic Coding Potential and Proteomes. *Nucleic Acids Res.* 47 (D1), 403–410. doi:10.1093/nar/gky936

Budamgunta, H., Olexiouk, V., Luyten, W., Schildermans, K., Maes, E., Boonen, K., et al. (2018). Comprehensive Peptide Analysis of Mouse Brain Striatum Identifies Novel SORF-Encoded Polypeptides. *Proteomics* 18, 1700218. doi:10.1002/pmic.201700218

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: Architecture and Applications. *BMC Bioinformatics* 10, 421. doi:10.1186/1471-2105-10-421

Cao, X., Khitun, A., Na, Z., Dumitrescu, D. G., Kubica, M., Olatunji, E., et al. (2020). Comparative Proteomic Profiling of Unannotated Microproteins and Alternative Proteins in Human Cell Lines. *J. Proteome Res.* 19 (8), 3418–3426. doi:10.1021/acs.jproteome.0c00254

Cardon, T., Hervé, F., Delcourt, V., Roucou, X., Salzet, M., Franck, J., et al. (2020). Optimized Sample Preparation Workflow for Improved Identification of Ghost Proteins. *Anal. Chem.* 92, 1122–1129. doi:10.1021/acs.analchem.9b04188

Casimiro-Soriguer, C. S., Rigual, M. M., Brokate-Llanos, A. M., Muñoz, M. J., Garzón, A., Pérez-Pulido, A. J., et al. (2020). Using AnABlast for Intergenic

SORF Prediction in the Caenorhabditis Elegans Genome. *Bioinformatics* 36, 4827–4832. doi:10.1093/bioinformatics/btaa608

Cassidy, L., Helbig, A. O., Kaulich, P. T., Weidenbach, K., Schmitz, R. A., and Tholey, A. (2021). Multidimensional Separation Schemes Enhance the Identification and Molecular Characterization of Low Molecular Weight Proteomes and Short Open Reading Frame-Encoded Peptides in Top-Down Proteomics. *J. Proteomics* 230, 103988. doi:10.1016/j.jprot.2020.103988

Cesnik, A. J., Miller, R. M., Ibrahim, K., Lu, L., Millikin, R. J., Shortreed, M. R., et al. (2020). Spritz: A Proteogenomic Database Engine. *J. Proteome Res.* 2020. doi:10.1101/2020.06.08.140681

Chatterjee, S., Stupp, G. S., Park, S. K. R., Ducom, J.-C., Su, A. I., and Wolan, D. W. (2016). A Comprehensive and Scalable Database Search System for Metaproteomics. *BMC Genomics*, 17. 642. doi:10.1186/s12864-016-2855-3

Chen, J., Brunner, A.-D., Cogan, J. Z., Nuñez, J. K., Fields, A. P., Adamson, B., et al. (2020). Pervasive Functional Translation of Noncanonical Human Open Reading Frames. *Science* 367 (6482), 1140–1146. LP – 1146. doi:10.1126/science.aay0262

Chew, G. L., Pauli, A., and Schier, A. F. (2016). Conservation of UORF Repressiveness and Sequence Features in Mouse, Human and Zebrafish. *Nat. Commun.* 7 (1), 1–10. doi:10.1038/ncomms11663

Chu, Q., Ma, J., and Saghatelian, A. (2015). Identification and Characterization of SORF-Encoded Polypeptides. *Crit. Rev. Biochem. Mol. Biol.* 50, 134–141. doi:10.3109/10409238.2015.1016215

Chu, Q., Martinez, T. F., Novak, S. W., Donaldson, C. J., Tan, D., Vaughan, J. M., et al. (2019). Regulation of the ER Stress Response by a Mitochondrial Microprotein. *Nat. Commun.* 10 (1), 1–13. doi:10.1038/s41467-019-12816-z

Claverie, J. (1997). Computational Methods for the Identification of Genes in Vertebrate Genomic Sequences. *Hum. Mol. Genet.* 6 (10), 1735–1744. doi:10.1093/hmg/6.10.1735

Crowe, M. L., Wang, X.-Q., and Rothnagel, J. A. (2006). Evidence for Conservation and Selection of Upstream Open Reading Frames Suggests Probable Encoding of Bioactive Peptides. *BMC Genomics* 7, 16. doi:10.1186/1471-2164-7-16

Dunn, J. G., and Weissman, J. S. (2016). Plastid: Nucleotide-Resolution Analysis of Next-Generation Sequencing and Genomics Data. *BMC Genomics* 17, 958. doi:10.1186/s12864-016-3278-x

Edifizi, D., Nolte, H., Babu, V., Castells-Roca, L., Mueller, M. M., Brodesser, S., et al. (2017). Multilayered Reprogramming in Response to Persistent DNA Damage in C. Elegans. *Cel Rep.* 20 (9), 2026–2043. doi:10.1016/j.celrep.2017.08.028

Fay, D. (2006). Genetic Mapping and Manipulation: Chapter 1-Introduction and Basics, WormBook, *ed. The* C. elegans *Research Community*, , 1–12. doi:10.1895/wormbook.1.90.1

Fermin, D., Allen, B. B., Blackwell, T. W., Menon, R., Adamski, M., Xu, Y., et al. (2006). Novel Gene and Gene Model Detection Using a Whole Genome Open Reading Frame Analysis in Proteomics. *Genome Biol.* 7 (4), R35. doi:10.1186/gb-2006-7-4-r35

Guillot, L., Delage, L., Viari, A., Vandenbrouck, Y., Com, E., Ritter, A., et al. (2019). Peptimapper: Proteogenomics Workflow for the Expert Annotation of Eukaryotic Genomes. *BMC Genomics* 20 (1), 56. doi:10.1186/s12864-019-5431-9

Guruceaga, E., Garin-Muga, A., and Segura, V. (2020). MiTPeptideDB: A Proteogenomic Resource for the Discovery of Novel Peptides. *Bioinformatics* 36 (1), 205–211. doi:10.1093/bioinformatics/btz530

Hao, Y., Zhang, L., Niu, Y., Cai, T., Luo, J., He, S., et al. (2018). SmProt: A Database of Small Proteins Encoded by Annotated Coding and Non-coding RNA Loci. *Brief. Bioinform.* 19 (4), bbx005–643. doi:10.1093/bib/bbx005

Harlow, E., and Lane, D. (2006). Bradford Assay. *Cold Spring Harbor Protoc.* 2006 (6), prot4644–pdb. doi:10.1101/pdb.prot4644

He, C., Jia, C., Zhang, Y., and Xu, P. (2018). Enrichment-Based Proteogenomics Identifies Microproteins, Missing Proteins, and Novel SmORFs in Saccharomyces Cerevisiae. *J. Proteome Res.* 17 (7), 2335–2344. doi:10.1021/acs.jproteome.8b00032

Hendriks, G.-J., Gaidatzis, D., Aeschimann, F., and Großhans, H. (2014). Extensive Oscillatory Gene Expression during *C. elegans* Larval Development. *Mol. Cel* 53 (3), 380–392. doi:10.1016/j.molcel.2013.12.013

Ingolia, N. T. (2016). Ribosome Footprint Profiling of Translation throughout the Genome. *Cell* 165, 22–33. doi:10.1016/j.cell.2016.02.066

Jagtap, P. D., Johnson, J. E., Onsongo, G., Sadler, F. W., Murray, K., Wang, Y., et al. (2014). Flexible and Accessible Workflows for Improved Proteogenomic Analysis Using the Galaxy Framework, *J. Proteome Res.*.,13, 5898–5908. doi:10.1021/pr500812t

Johnstone, T. G., Bazzini, A. A., and Giraldez, A. J. (2016). Upstream ORF S Are Prevalent Translational Repressors in Vertebrates. *EMBO J.* 35 (7), 706–723. doi:10.15252/embj.201592759

Kastenmayer, J. P., Ni, L., Chu, A., Kitchen, L. E., Au, W. C., Yang, H., et al. (2006). Functional Genomics of Genes with Small Open Reading Frames (SORFs) in S. Cerevisiae. *Genome Res.* 16 (3), 365–373. doi:10.1101/gr.4355406

Kaulich, P. T., Cassidy, L., Weidenbach, K., Schmitz, R. A., and Tholey, A. (2020). Complementarity of Different SDS-PAGE Gel Staining Methods for the Identification of Short Open Reading Frame-Encoded Peptides. *Proteomics* 20 (19–20), 2000084. doi:10.1002/pmic.202000084

Kolmogorov, M., Liu, X., and Pevzner, P. A. (2016). SpectroGene: A Tool for Proteogenomic Annotations Using Top-Down Spectra. *J. Proteome Res.* 15 (1), 144–151. doi:10.1021/acs.jproteome.5b00610

Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D., and Nesvizhskii, A. I. MSFragger: Ultrafast and Comprehensive Peptide Identification in Mass Spectrometry-Based Proteomics. *Nat. Methods*. Methods 2017, 14 (5), 513–520. doi:10.1038/nmeth.4256

Ladoukakis, E., Pereira, V., Magny, E. G., Eyre-Walker, A., and Couso, J. (2011). Hundreds of Putatively Functional Small Open Reading Frames in Drosophila. *Genome Biol.* 12 (11), R118. doi:10.1186/gb-2011-12-11-r118

Lewis, J. A., and Fleming, J. T. (1995). Chapter 1 Basic Culture Methods. *Methods Cel Biol* 48 (C), 3–29. doi:10.1016/S0091-679X(08)61381-3

Li, W., Petruzziello, F., Zhao, N., Zhao, H., Ye, X., Zhang, X., et al. (2017). Separation and Identification of Mouse Brain Tissue Microproteins Using Top-down Method with High Resolution Nanocapillary Liquid Chromatography Mass Spectrometry, *Proteomics*, 17, 1600419. doi:10.1002/pmic.201600419

Ma, J., Diedrich, J. K., Jungreis, I., Donaldson, C., Vaughan, J., Kellis, M., et al. (2016). Improved Identification and Analysis of Small Open Reading Frame Encoded Polypeptides. *Anal. Chem.* 88 (7), 3967–3975. doi:10.1021/acs.analchem.6b00191

Ma, J., Diedrich, J. K., Jungreis, I., Donaldson, C., Vaughan, J., Kellis, M., et al. (2016). Improved Identification and Analysis of Small Open Reading Frame Encoded Polypeptides. *Anal. Chem.* 88 (7), 3967–3975. doi:10.1021/acs.analchem.6b00191

Mackowiak, S. D., Zauber, H., Bielow, C., Thiel, D., Kutz, K., Calviello, L., et al. (2015). Extensive Identification and Analysis of Conserved Small ORFs in Animals. *Genome Biol.* 16 (1), 179. doi:10.1186/s13059-015-0742-x

Makarewich, C. A., Baskin, K. K., Munir, A. Z., Bezprozvannaya, S., Sharma, G., Khemtong, C., et al. (2018). MOXI Is a Mitochondrial Micropeptide that Enhances Fatty Acid β-Oxidation. *Cel Rep.* 23 (13), 3701–3709. doi:10.1016/j.celrep.2018.05.058

Martens, L., Vandekerckhove, J., and Gevaert, K. (2005). DBToolkit: Processing Protein Databases for Peptide-Centric Proteomics. *Bioinformatics* 21 (17), 3584–3585. doi:10.1093/bioinformatics/bti588

Meier, F., Brunner, A.-D., Koch, S., Koch, H., Lubeck, M., Krause, M., et al. (2018). Online Parallel Accumulation-Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer. *Mol. Cell Proteomics* 17 (12), 2534–2545. doi:10.1074/mcp.TIR118.000900

Na, Z., Luo, Y., Schofield, J. A., Smelyansky, S., Khitun, A., Muthukumar, S., et al. (2020). The NBDY Microprotein Regulates Cellular RNA Decapping. *Biochemistry* 59 (42), 4131–4142. doi:10.1021/acs.biochem.0c00672

Nagaraj, S. H., Waddell, N., Madugundu, A. K., Wood, S., Jones, A., Mandyam, R. A., et al. (2015). PGTools: A Software Suite for Proteogenomic Data Analysis and Visualization, *J. Proteome Res.*.,14, 2255–2266. doi:10.1021/acs.jproteome.5b00029

Narayan, V., Ly, T., Pourkarimi, E., Murillo, A. B., Gartner, A., Lamond, A. I., et al. (2016). Deep Proteome Analysis Identifies Age-Related Processes in C. Elegans. *Cel Syst.* 3 (2), 144–159. doi:10.1016/j.cels.2016.06.011

Nedialkova, D. D., and Leidel, S. A. (2015). Optimization of Codon Translation Rates via TRNA Modifications Maintains Proteome Integrity. *Cell* 161 (7), 1606–1618. doi:10.1016/j.cell.2015.05.022

Nematode Growth Medium (Ngm). *Cold Spring Harbor Protoc.* 2014, Nematode Growth Medium (NGM), 2014 (3), pdb.rec081299. doi:10.1101/pdb.rec081299

NesvizhskiiProteogenomics, A. I. (2014). .Proteogenomics: Concepts, Applications and Computational Strategies, *Nature Methods*, 11. , 2014 Nature Publishing Group November, 1114–1125. doi:10.1038/NMETH.3144

Olexiouk, V., Crappé, J., Verbruggen, S., Verhegen, K., Martens, L., and MenschaertSORFs, G. (2016). sORFs.org: a Repository of Small ORFs Identified by Ribosome Profiling. *Nucleic Acids Res.* 44 (D1), D324–D329. doi:10.1093/nar/gkv1175

Olexiouk, V., Van Criekinge, W., and Menschaert, G. (2018). An Update on SORFs.Org: A Repository of Small ORFs Identified by Ribosome Profiling. *Nucleic Acids Res.* 46 (D1), D497–D502. doi:10.1093/nar/gkx1130

Olexiouk, V., Van Criekinge, W., and Menschaert, G. (2018). An Update on SORFs.Org: A Repository of Small ORFs Identified by Ribosome Profiling. *Nucleic Acids Res.* 46 (D1), D497–D502. doi:10.1093/nar/gkx1130

Omasits, U., Varadarajan, A. R., Schmid, M., Goetze, S., Melidis, D., Bourqui, M., et al. (2017). An Integrative Strategy to Identify the Entire Protein Coding Potential of Prokaryotic Genomes by Proteogenomics, *Genome Res.*, .27, 2083–2095. doi:10.1101/gr.218255.116

Osorio, D., Rondón-Villarreal, P., and Torres, R. (2015). Peptides: A Package for Data Mining of Antimicrobial Peptides. *R. J.* 7 (1), 4–14. doi:10.32614/rj-2015-001

Paix, A., Folkmann, A., and Seydoux, G. (2017). Precision Genome Editing Using CRISPR-Cas9 and Linear Repair Templates in C. Elegans. *Methods* 121-122, 86–93. doi:10.1016/j.ymeth.2017.03.023

Pauli, A., Norris, M. L., Valen, E., Chew, G.-L., Gagnon, J. A., Zimmerman, S., et al. (2014). Toddler: An Embryonic Signal that Promotes Cell Movement via Apelin Receptors. *Science* 343 (6172), 1248636. doi:10.1126/science.1248636

Porta-De-La-Riva, M., Fontrodona, L., Villanueva, A., and Cerón, J. (2012). Basic Caenorhabditis Elegans Methods: Synchronization and Observation. *JoVE* 64. e4019, No. doi:10.3791/4019

R Core Team, R. (2020). *A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raj, A., Wang, S. H., Shim, H., Harpak, A., Li, Y. I., Engelmann, B., et al. (2016). Thousands of Novel Translated Open Reading Frames in Humans Inferred by Ribosome Footprint Profiling. *Elife* 5 (MAY2016). doi:10.7554/eLife.13328

Rathore, A., Chu, Q., Tan, D., Martinez, T. F., Donaldson, C. J., Diedrich, J. K., et al. (2018). MIEF1 Microprotein Regulates Mitochondrial Translation. *Biochemistry* 57 (38), 5564–5575. doi:10.1021/acs.biochem.8b00726

Risk, B. A., Spitzer, W. J., and Giddings, M. C. (2013). *Peppy: Proteogenomic Search Software*, Peppy: Proteogenomic Search Software, *J. Proteome Res.*., 12, 3019–3025. doi:10.1021/pr400208w

Schwinn, M. K., Steffen, L. S., Zimmerman, K., Wood, K. V., and Machleidt, T. (2020). A Simple and Scalable Strategy for Analysis of Endogenous Protein Dynamics. *Sci. Rep.* 10 (1), 1–14. doi:10.1038/s41598-020-65832-1

Secher, A., Kelstrup, C. D., Conde-Frieboes, K. W., Pyke, C., Raun, K., Wulff, B. S., et al. (2016). Analytic Framework for Peptidomics Applied to Large-Scale Neuropeptide Identification. *Nat. Commun.* 7 (1), 1–10. doi:10.1038/ncomms11436

Shteynberg, D., Nesvizhskii, A. I., Moritz, R. L., and Deutsch, E. W. (2013). Combining Results of Multiple Search Engines in Proteomics, *Molecular and Cellular Proteomics*, 12. , 2013 Elsevier September, 2383–2393. doi:10.1074/mcp.R113.027797

Sieber, P., Platzer, M., and Schuster, S. (2018). The Definition of Open Reading Frame Revisited, *Trends in Genetics*, 34. , 2018 Elsevier Ltd March 1, 167–170. doi:10.1016/j.tig.2017.12.009

Stadler, M., Artiles, K., Pak, J., and Fire, A. (2012). Contributions of mRNA Abundance, Ribosome Loading, and post- or Peri-Translational Effects to Temporal Repression of *C. elegans* Heterochronic miRNA Targets. *Genome Res.* 22 (12), 2418–2426. doi:10.1101/gr.136515.111.influenced

Stadler, M., and Fire, A. (2011). Wobble Base-Pairing Slows *In Vivo* Translation Elongation in Metazoans. *RNA* 17 (12), 2063–2073. doi:10.1261/rna.02890211

Verbruggen, S., Gessulat, S., Gabriels, R., Matsaroki, A., Van de Voorde, H., Kuster, B., et al. (2021). Spectral Prediction Features as a Solution for the Search Space Size Problem in Proteogenomics. *Mol. Cell Proteomics* 20, 100076. doi:10.1016/j.mcpro.2021.100076

Verbruggen, S., and Menschaert, G. (2019). mQC: A post-mapping Data Exploration Tool for Ribosome Profiling. *Computer Methods Programs Biomed.* 181 (104806), 104806. doi:10.1016/j.cmpb.2018.10.018

Verheggen, K., Volders, P.-J., Mestdagh, P., Menschaert, G., Van Damme, P., et al. (2017). Noncoding after All: Biases in Proteomics Data Do Not Explain Observed Absence of LncRNA Translation Products. *J. Proteome Res.* 16 (7), 2508–2515. doi:10.1021/acs.jproteome.7b00085

Wang, B., Hao, J., Pan, N., Wang, Z., Chen, Y., and Wan, C. (2021). Identification and Analysis of Small Proteins and Short Open Reading Frame Encoded Peptides in Hep3B Cell. *J. Proteomics* 230, 103965. doi:10.1016/j.jprot.2020.103965

Wang, M., Zhao, Y., and Zhang, B. (2015). Efficient Test and Visualization of Multi-Set Intersections. *Sci. Rep.* , 5, 16923. doi:10.1038/srep16923

Wickham, H. (2016). *Ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

Xia, T., Horton, E. R., Salcini, A. E., Pocock, R., Cox, T. R., and Erler, J. T. (2018). Proteomic Characterization of Caenorhabditis Elegans Larval Development. *Proteomics* 18 (2), 1700238. doi:10.1002/pmic.201700238

Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., et al. (2020). Ensembl 2020. *Nucleic Acids Res.* 48 (D1), D682–D688. doi:10.1093/nar/gkz966

Yu, F., Haynes, S. E., Teo, G. C., Avtonomov, D. M., Polasky, D. A., and Nesvizhskii, A. I. (2020). Fast Quantitative Analysis of TimsTOF PASEF Data with MSFragger and IonQuant. *Mol. Cell Proteomics* 19 (9), 1575–1585. doi:10.1074/mcp.TIR120.002048

Yu, F., Haynes, S. E., Teo, G. C., Avtonomov, D. M., Polasky, D. A., and Nesvizhskii, A. I. (2020). Fast Quantitative Analysis of TimsTOF PASEF Data with MSFragger and IonQuant. *Mol. Cell Proteomics* 19 (9), 1575–1585. doi:10.1074/mcp.TIR120.002048

Zhang, H., Wang, Y., and Lu, J. (2019). Function and Evolution of Upstream ORFs in Eukaryotes. *Trends Biochem. Sci.* 44 (9), 782–794. doi:10.1016/J.TIBS.2019.03.002

Zhang, J., Xin, L., Shan, B., Chen, W., Xie, M., Yuen, D., et al. (2012). PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification. *Mol. Cell Proteomics* 11 (4). M111010587, doi:10.1074/mcp.M111.010587

Zickmann, F., and Renard, B. Y. (2015). MSProGene: Integrative Proteogenomics beyond Six-Frames and Single Nucleotide Polymorphisms. *Bioinformatics* 31 (12), i106–i115. doi:10.1093/BIOINFORMATICS/BTV236

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.