ORIGINAL RESEARCH
published: 26 May 2022
doi: 10.3389/fcell.2022.901351

In Depth Exploration of the Alternative Proteome of *Drosophila melanogaster*

Bertrand Fabre[1,2], Sebastien A. Choteau[3], Carine Duboé[1], Carole Pichereaux[4,5,6], Audrey Montigny[1], Dagmara Korona[7], Michael J. Deery[2], Mylène Camus[5,6], Christine Brun[3,8], Odile Burlet-Schiltz[5,6], Steven Russell[7], Jean-Philippe Combier[1], Kathryn S. Lilley[2] and Serge Plaza[1]**

[1]Laboratoire de Recherche en Sciences Végétales, UMR5546, Université de Toulouse, UPS, INP, CNRS, Auzeville-Tolosane, France, [2]Cambridge Centre for Proteomics, Cambridge Systems Biology Centre and Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom, [3]Aix-Marseille Université, INSERM, TAGC, Turing Centre for Living Systems, Marseille, France, [4]Fédération de Recherche (FR3450), Agrobiosciences, Interactions et Biodiversité (AIB), CNRS, Toulouse, France, [5]Institut de Pharmacologie et de Biologie Structurale (IPBS), Université de Toulouse, CNRS, UPS, Toulouse, France, [6]Infrastructure Nationale de Protéomique, ProFI, FR 2048, Toulouse, France, [7]Cambridge Systems Biology Centre and Department of Genetics, University of Cambridge, Cambridge, United Kingdom, [8]CNRS, Marseille, France

OPEN ACCESS

Edited by:
*Suman S. Thakur,
Centre for Cellular and Molecular Biology (CCMB), India*

Reviewed by:
*Paul Lasko,
McGill University, Canada
Ken Moberg,
Emory University School of Medicine, United States*

***Correspondence:**
*Bertrand Fabre
bertrand.fabre@univ-tlse3.fr
Serge Plaza
serge.plaza@univ-tlse3.fr*

Specialty section:
This article was submitted to Cellular Biochemistry, a section of the journal Frontiers in Cell and Developmental Biology

Received: 21 March 2022
Accepted: 25 April 2022
Published: 26 May 2022

Citation:
Fabre B, Choteau SA, Duboé C, Pichereaux C, Montigny A, Korona D, Deery MJ, Camus M, Brun C, Burlet-Schiltz O, Russell S, Combier J-P, Lilley KS and Plaza S (2022) In Depth Exploration of the Alternative Proteome of Drosophila melanogaster. Front. Cell Dev. Biol. 10:901351. doi: 10.3389/fcell.2022.901351

Recent studies have shown that hundreds of small proteins were occulted when protein-coding genes were annotated. These proteins, called alternative proteins, have failed to be annotated notably due to the short length of their open reading frame (less than 100 codons) or the enforced rule establishing that messenger RNAs (mRNAs) are monocistronic. Several alternative proteins were shown to be biologically active molecules and seem to be involved in a wide range of biological functions. However, genome-wide exploration of the alternative proteome is still limited to a few species. In the present article, we describe a deep peptidomics workflow which enabled the identification of 401 alternative proteins in *Drosophila melanogaster*. Subcellular localization, protein domains, and short linear motifs were predicted for 235 of the alternative proteins identified and point toward specific functions of these small proteins. Several alternative proteins had approximated abundances higher than their canonical counterparts, suggesting that these alternative proteins are actually the main products of their corresponding genes. Finally, we observed 14 alternative proteins with developmentally regulated expression patterns and 10 induced upon the heat-shock treatment of embryos, demonstrating stage or stress-specific production of alternative proteins.

Keywords: alternative proteins, short open reading frame–encoded polypeptide, microprotein, peptidomics, mass spectrometry

INTRODUCTION

Almost 20 years after the completion of the sequencing of the genomes of *Saccharomyces cerevisiae*, *Drosophila melanogaster*, and *Homo sapiens*, precise gene annotation still remains challenging. Initiatives such as the Human Proteome Project (HPP) (Omenn et al., 2021) or ProteomicsDB (Lautenbacher et al., 2021) aim at defining the ensemble of proteins actually expressed in humans or other organisms using mass spectrometry (MS) based approaches. These projects have reached impressive milestones but they are centered around the protein database that is used to mine the experimental MS data in order to identify expressed proteins (Brunet et al., 2020). So far, these databases mainly comprise genes annotated in UniProtKB (Bateman et al., 2021). However, recent studies have suggested that hundreds of small yet to be annotated proteins, might be expressed across the kingdom of life (Fabre et al., 2021). These proteins, called alternative prot0065ins (AltProts, or short open reading frame (ORF) encoding polypeptides (SEPs) or microproteins), have failed to be annotated notably due to the short length of their open reading frame (less than 100 codons), alternative start codon (other than AUG) or the enforced rule establishing that messenger RNAs (mRNAs) are monocistronic (Brunet et al., 2020). Almost two decades of pioneering work have highlighted that AltProts can be produced from ORFs on long non–coding RNA (lncRNA) or the different regions of mRNAs, within the 5′ or 3′ untranslated regions or alternative frames in canonical coding sequences (called uORFs, dORFs, and intORFs, respectively) (Plaza et al., 2017). Databases such as OpenProt (Brunet et al., 2019), sORFs.org (Olexiouk et al., 2018), SmProt (Li et al., 2021), ARA-PEPs (Hazarika et al., 2017), PsORF (Chen Y. et al., 2020), or MetamORF (Choteau et al., 2021) constitute repositories predicting the existence of potentially thousands of AltProts based mainly on ribosome footprints determined via ribosome profiling experiments. However, in most cases, we still lack unambiguous empirical evidence for the existence of most of these predicted short proteins. Although ribosome profiling approaches clearly established the binding of ribosome to alternative ORFs, it is in fact difficult to deduce the productive translation of the ORFs, resulting in the expression of stable proteins (Patraquim et al., 2020). Mass spectrometry is generally the method of choice for large scale identification of proteins and peptides (Cassidy et al., 2021). MS data demonstrating the genome-wide expression of AltProts are still limited to few species (Fabre et al., 2021). The roles of only few alternative proteins, less than 50 across all species, have been characterized to date (Plaza et al., 2017; Wright et al., 2021). The alternative proteins, whose function has been determined, seem to be involved in a wide range of key biological processes (Plaza et al., 2017; Wright et al., 2021). Due to their large spectrum of functions, alternative proteins represent an attractive new repertoire of molecules for drug development and agricultural applications.

In an effort to assess the pervasive production of alternative proteins in the model organism *Drosophila melanogaster*, we describe here the development of a deep peptidomics workflow combining different protein extraction methods, small protein enrichment steps, state of the art mass spectrometry, and optimized bioinformatics analysis using the well-curated OpenProt database. We were able to identify 401 yet unannotated alternative proteins, substantially increasing (twice) the repertoire of alternative proteins in *Drosophila melanogaster*. The majority of these proteins are produced from alternative reading frames in the canonical coding sequences (CDS), highlighting the fact that the proteome is more complex than previously anticipated. AltProts produced from different types of RNA (lncRNA or mRNA) or different regions of mRNA (5′ or 3′ untranslated regions or alternative frames within canonical CDS) have different amino acid compositions, isoelectric points, predicted protein domains, or disordered regions. Surprisingly, AltProts are predicted to be localized mainly in the cell nucleus, mitochondria, or secreted. We identified several AltProts for which the approximated abundances were higher than their canonical counterparts, suggesting that these AltProts are actually the main products of their corresponding genes. Finally, we observed 14 AltProts with developmentally regulated expression patterns and 10 induced upon the heat-shock treatment of embryos, demonstrating stage, or stress specific production of alternative proteins.

MATERIALS AND METHODS

Drosophila Collection and S2 Cell Culture

D. melanogaster adult flies and embryos were maintained and collected as previously described (Fabre et al., 2019). S2 cells were cultured as described in Montigny et al. (2021).

Protein Extraction and Alternative Protein Enrichment

Several approaches were used to extract and enrich alternative proteins:

1) Embryo (100 µl equivalent of embryo per replicate), adult flies (10 adult flies per replicate), or S2 cell pellets (5×10^8 cells per replicate) were resuspended in an SDS buffer (Tris 50 mM pH 7.5, 5% SDS), then immediately sonicated and boiled for 10 min at 95°C. A detergent compatible protein assay (Bio-Rad) was used to measure the protein concentration. Loading buffer (Tris 40 mM pH 7.5, 2% SDS, 10% glycerol, and 25 mM DTT final concentration) was added to 100 µg of protein per condition and samples were boiled for 5 min at 95°C. The proteins were alkylated using chloroacetamide at a final concentration of 60 mM for 30 min at room temperature in the dark. The samples were loaded on an SDS-PAGE gel (acrylamide concentration of 4% for the stacking gel and 12% for the resolving gel). After protein migration, staining with InstantBlue™ (Merck) was performed and bands were excised between 15 kDa and the dye front (three bands for S2 cells and two bands for embryos and adult flies). The proteins were then digested over night at 37°C with trypsin (or glu-C or chymotrypsin in the case of S2 cells) using in-gel digestion as previously described (Fabre et al., 2016b). The resulting peptides were injected on a Thermofisher Q Exactive plus (S2 cells samples only) or a Thermofisher Fusion (embryo and adult flies samples only). Three biological replicates were performed for each condition.
2) Embryo (200 µl equivalent of embryo per replicate) and adult flies (50 adult flies per replicate) were lysed and proteins were reduced and alkylated as described in the approach 1 and 1 mg of protein were digested using in-gel digestion (trypsin for adult flies, or trypsin, Glu-C, or chymotrypsin for embryos). The resulting peptides were then separated by high pH reverse phase fractionation as described in Fabre et al. (2017). Each fraction was analyzed either on a Sciex TripleTOF 6600 (both embryo and adult fly samples), a Thermofisher Q Exactive

(embryo samples only), or a Thermofisher Fusion Lumos (embryo samples only). Three biological replicates were performed for each condition.

3) Embryos (100 µl equivalent of embryo per replicate) were incubated at 37°C to induce heat-shock or maintained at 25°C as described previously (Fabre et al., 2016c). The proteins were extracted, reduced, and alkylated as described in protocol 1 and 100 µg were loaded on an SDS-PAGE gel (acrylamide concentration of 4% for the stacking gel and 12% for the resolving gel). After a short migration, each gel lane was cut in three bands and in-gel digestion was performed with trypsin as previously described (Fabre et al., 2016b).The resulting peptides were injected on a Thermofisher Q Exactive. Three biological replicates were performed for each condition.
4) Embryos (100 µl equivalent of embryo per replicate) staged every 4.5 h as previously described (Fabre et al., 2016a) were lysed in a buffer containing 20 mM HEPES pH 8, 150 mM KCl, and 10 mM $MgCl_2$ and proteins were first digested with proteinase K and boiled for 10 min after the addition of guanidine hydrochloride (GnHCl) at a 6 M final concentration. The proteins were then reduced with 25 mM dithiothreitol (DTT), alkylated with chloroacetamide at a final concentration of 60 mM, and digested with trypsin, glu-C, or chymotrypsin over night at 37°C. The peptides were desalted on a C18 SepPak column (Waters), dried down using a speed-vac, labeled with Tandem Mass Tag (TMT) 10-plex (Thermo Scientific) according to the manufacturer's instructions, pooled and fractionated using the High pH Reversed-Phase Peptide Fractionation Kit (Pierce). Each fraction was analyzed on a Thermofisher Fusion Lumos. Three biological replicates were performed for each condition.
5) 5×10^8 S2 cells were boiled at 95°C for 20 min in water and sonicated. Then acetic acid and acetonitrile were added to the sample both at a final concentration of 20 and 5%, respectively. The samples were centrifuged at 20,000 g for 20 min at 4°C and the pellet was discarded. The supernatant was dried using a speed-vac and proteins were resuspended in 6 M GnHCl and 50 mM ammonium bicarbonate. A BCA assay (Pierce) was used to measure the protein concentration. The proteins were reduced in 5 mM TCEP (tris 2-carboxyethylphosphine hydrochloride) for 1 h at 37°C and alkylated in 10 mM chloroacetamide for 30 min at RT in the dark. The samples were diluted with 50 mM ammonium bicarbonate at a final concentration of GnHCl of 0.5 M. The proteins were digested with trypsin (at a 1:50 trypsin to protein ratio) and resulting peptides were desalted on a C18 Hypersep column (Thermo Scientific) and dried down using a speed-vac. The samples were injected on a Thermofisher Fusion. Two biological replicates were performed.
6) 5×10^8 S2 cells were boiled at 95°C for 20 min in GnHCl lysis buffer (6 M guanidine hydrochloride, Tris 50 mM pH 7.5, and 100 mM NaCl) and sonicated. The samples were centrifuged at 20,000 g for 20 min at RT and the pellet was discarded. Trifluoroacetic acid (TFA) was added to the supernatant at a final concentration of 0.4% before loading the sample on a C8 column (Pierce) preconditioned with acetonitrile (ACN) and equilibrated with 0.1% TFA. The column was washed twice with 0.1% TFA and proteins were eluted with 75% ACN and 0.1% TFA. The samples were dried down using a speed-vac and resuspended in 6 M GnHCl and 50 mM ammonium bicarbonate. A BCA assay (Pierce) was used to measure the protein concentration. The proteins were reduced in 5 mM TCEP (tris 2-carboxyethylphosphine hydrochloride) for 1 h at 37°C and alkylated in 10 mM chloroacetamide for 30 min at RT in the dark. The samples were diluted with 50 mM ammonium bicarbonate at a final concentration of GnHCl of 0.5 M. The proteins were digested with trypsin (at a 1:50 trypsin to protein ratio) and the resulting peptides were desalted on a C18 Hypersep column (Thermo Scientific) and dried down using a speed-vac. The samples were injected on a Thermofisher Fusion. Three biological replicates were performed.
7) 5×10^8 S2 cells were boiled at 95°C for 20 min in GnHCl lysis buffer (6 M guanidine hydrochloride, Tris 50 mM pH 7.5, and 100 mM NaCl) and sonicated. The sample was centrifuged at 20,000 g for 20 min at RT and the pellet was discarded. The supernatant was loaded on an ultrafiltration device with a molecular weight cut-off of 30 kDa (Millipore) and the fraction retained (above 30 kDa) was discarded. A BCA assay (Pierce) was used to measure the protein concentration. The proteins were reduced in 5 mM TCEP (tris 2-carboxyethylphosphine hydrochloride) for 1 h at 37°C and alkylated in 10 mM chloroacetamide for 30 min at RT in the dark. The samples were diluted with 50 mM ammonium bicarbonate at a final concentration of GnHCl of 0.5 M. The proteins were digested with trypsin (at a 1:50 trypsin to protein ratio) and resulting peptides were desalted on a C18 Hypersep column (Thermo Scientific) and dried down using a speed-vac. The samples were injected on a Thermofisher Orbitrap Velos. One biological replicate was performed.

Mass Spectrometry Analysis

Sciex TripleTOF 6600 and Thermofisher Q Exactive were operated as described in Mata et al. (2017). The Thermofisher OrbiTrap Fusion Lumos was used as in Geladaki et al. (2019). The Thermofisher OrbiTrap Velos and Q Exactive plus were operated as described in Menneteau et al. (2019). The Thermofisher OrbiTrap Fusion was used as described in Payros et al. (2021).

Mass Spectrometry Data Analysis

The raw files generated during this work and previous studies (Wan et al., 2015; Wessels et al., 2016; Müller et al., 2020) were analyzed using MaxQuant (Cox et al., 2014) version 1.6.15.0. The minimal peptide length was set to 7. Trypsin/P, GluC, or chymotrypsin were used as the digestive enzymes. Search criteria included carbamidomethylation of cysteine as a fixed modification, oxidation of methionine, and N-terminal acetylation as variable modifications. Up to two missed cleavages were allowed. The mass tolerance for the precursor was set to 20 and 4.5 ppm for the first and the main searches, respectively, and 20 ppm for the fragment ions for Thermofisher instruments. The mass tolerance for the precursor was 0.07 and 0.006 Da for the first and the main searches, respectively, and for the fragment ions was 50 ppm and TOF recalibration was enabled for the Sciex TripleTOF 6600 instrument. The raw files were searched against the OpenProt fasta *Drosophila melanogaster* database (release 1.6, Altprots, isoforms, and Refprots). For the identification of RefProts, default MaxQuant settings were used (1% FDR both at the protein and PSM levels). Regarding AltProts identification, a minimum score of 70 was set for both modified and unmodified peptides (corresponding to the first quartile of the distribution of the score of RefProts from an analysis of the raw files with MaxQuant default settings). The candidates were filtered to obtain an FDR of 1% at the peptide level. Because alternative proteins are generally shorter than canonical proteins, no FDR was set at the protein level and no filter was applied to the number of peptides per protein. A minimum sequence coverage of 70% of the peptide sequence was required for the alternative protein identification. MSMS spectra were manually inspected by two independent operators. Peptides matching both a novel predicted protein and a RefProt were discarded. As implemented in OpenProt (Brunet et al., 2019), peptide matching two AltProts, two novel isoforms or an AltProt, and a novel isoform were assigned to both proteins in each case. For quantification, the match between runs and iBAQ modules of MaxQuant was enabled. Quantitative comparisons between AltProts and RefProts were performed on samples from the high pH reverse phase experiments only [protein extraction and alternative protein enrichment protocol number 2, and data from Müller et al. (2020)]. As iBAQ represents an approximation of the absolute abundance of a protein (Fabre et al., 2014) and given the low number of observable peptides for AltProts, we considered that an AltProt was more (or less) abundant than its corresponding RefProt if the ratio between their iBAQ values was at least 10-fold different. Otherwise, AltProt and RefProt were considered to have similar expression levels. STRING v11.5 (Szklarczyk et al., 2021) was used for network generation and GO term/KEGG pathway analysis.

Confocal Microscopy

For imaging experiments, S2 cells were co-transfected using an actin-GAL4 driver with UAS-CG34150-GFP and UAS-AltProtCG34150-RFP or UAS-CG265z-GFP and UAS-AltProtCG2650-RFP (both constructions encoding AltProts also contain the start codons and sequence of the canonical proteins). The cells were transfected with effectene (Qiagen) according to manufacturer specification and as described in Montigny et al. (2021). After 48 h of transfection, the cells were fixed in 4% formaldehyde in phosphate buffer saline (PBS) at room temperature for 30 min. They were rinsed three times in PBS for 10 min. Nuclei were stained with DAPI and samples were rinsed several times in PBS. Coverslides were mounted in Prolong (Invitrogen) and images were acquired using a SP8 Leica confocal microscope. Three biological replicates were performed for each condition.

Bioinformatic Analyses

Detection of the AltProt and RefProt Domains

The RefProt domains have been collected from the InterPro database (Blum et al., 2021). All domains identified on UniProtKB reviewed proteins of *Drosophila melanogaster* (Proteome identifier: UP000000803) have been recovered using the EBI REST API.

The domains on AltProt sequences have been identified using InterProScan (Jones et al., 2014) (v5.52-87.0) looking for signatures in the Pfam database (Mistry et al., 2021). The signatures with an e-value lower than 10^{-5} have been selected. Pfam identifiers have been mapped to InterPro accessions using the InterPro cross-references collected through the EBI REST API.

Detection of the Short Linear Motifs

The classes of SLiMs have been downloaded from the Eukaryotic Linear Motif (ELM) database (Kumar et al., 2020). The classes with a pattern probability lower than 0.01 and having at least one true positive instance detected in *D. melanogaster* in the ELM database have been selected.

The short linear motifs (SLiMs) have then been detected in the disordered regions of the AltProts using the IUPred2A (Mészáros et al., 2018) and the Short Linear Motif Probability tool (SLiMProb) of SLiMSuite (Edwards et al., 2020), using the following SLiMProb parameters: iumethod = long, iucut = 0.2, and minregion = 5.

Associations Between Short Linear Motifs and Domain Usage and AltProt Classes

To check whether AltProt classes were preferentially associated with SLiM or domain usage, chi-squared tests of independence have been performed.

Short Linear Motifs and Domain Enrichments and Depletions Among AltProt Classes

For each class type of motif (LIG, DOC, TRG, MOD, CLV, and DEG), and for each class of AltProt (ncRNA, isoform, 5′UTR, CDS, and 3′UTR), enrichment and depletion in AltProt with at least one motif of the class type among the AltProt of the class have been assessed, using one-sided Fisher's exact tests. The *p*-values computed have, then, been adjusted for multiple comparisons using the Benjamini–Hochberg procedure.

For each class of motif, and for each class of AltProt (ncRNA, isoform, 5′UTR, CDS, and 3′UTR), enrichment and depletion in AltProt with at least one motif of the class among the AltProt of the class have been assessed, using one-sided Fisher's exact tests. The *p*-values computed have then been adjusted for multiple comparisons using the Benjamini–Hochberg procedure.

Disorder regions (sequence of at least five amino acids) were predicted using IUPred2A (Mészáros et al., 2018) using the long disorder setting. The prediction of transmembrane helices and signal peptides were performed using TMHMM—2.0 (Krogh et al., 2001) and SignalP—5.0 (Almagro Armenteros et al., 2019), respectively. DeepLoc (Almagro Armenteros et al., 2017) was used to predict AltProts subcellular localization.

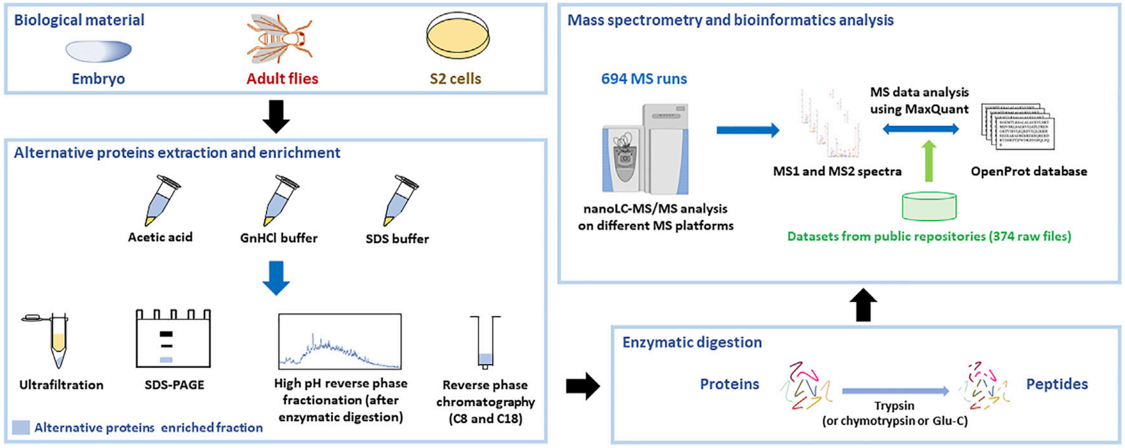
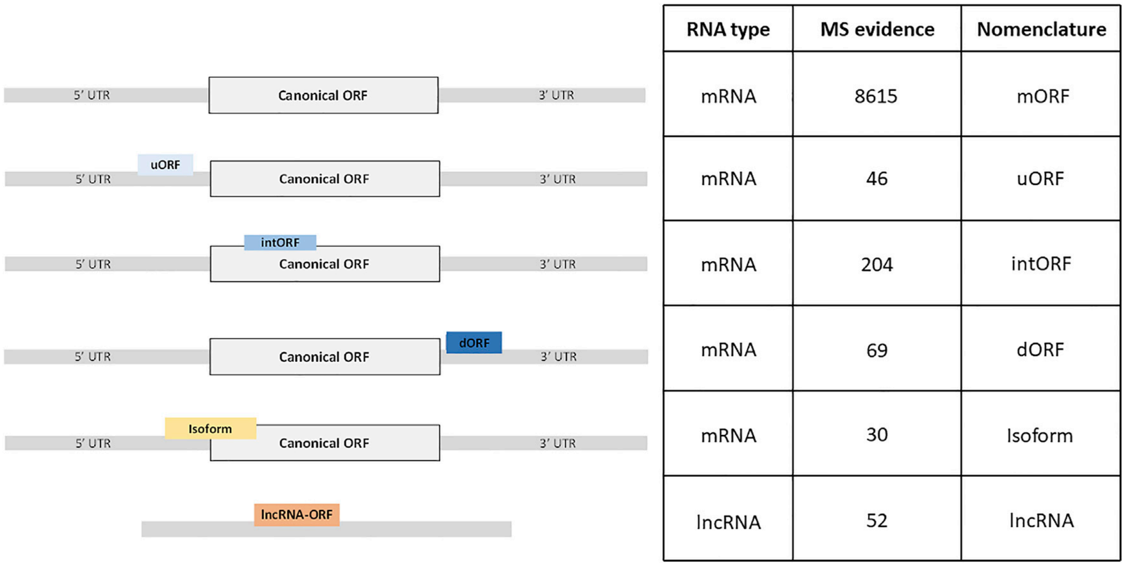

Biological material
Embryo
Adult flies
S2 cells
Alternative proteins extraction and enrichment
Acetic acid
GnHCl buffer
SDS buffer
Ultrafiltration
SDS-PAGE
High pH reverse phase fractionation (after enzymatic digestion)
Reverse phase chromatography (C8 and C18)
Alternative proteins enriched fraction
Mass spectrometry and bioinformatics analysis
694 MS runs
nanoLC-MS/MS analysis on different MS platforms
MS1 and MS2 spectra
MS data analysis using MaxQuant
OpenProt database
Datasets from public repositories (374 raw files)
Enzymatic digestion
Proteins
Trypsin (or chymotrypsin or Glu-C)
Peptides

FIGURE 1 | Peptidomics workflow to identify alternative proteins in *Drosophila melanogaster*. Proteins were extracted from embryos, adult flies, or S2 cultured cells using different extraction protocols. Alternative proteins were then enriched from the total protein pool and digested with trypsin (or other enzymes). The resulting peptides were injected on different mass spectrometry platforms and the generated MS data, as well as datasets available from public repositories, were analyzed using MaxQuant with the OpenProt database.



RNA type	MS evidence	Nomenclature
mRNA	8615	mORF
mRNA	46	uORF
mRNA	204	intORF
mRNA	69	dORF
mRNA	30	Isoform
lncRNA	52	lncRNA

FIGURE 2 | Different classes of ORFs produce yet non-annotated alternative proteins. Hundreds of alternative proteins are produced from ORFs from novel isoforms, ORFs located in the 5′UTR (uORF), coding sequence (intORF) or 3′UTR (dORF) of mRNAs, or from non-annotated ORFs on long non–coding RNAs (lncRNA). The nomenclature of the different classes of ORFs was adapted from Mudge et al. (2021).

RESULTS AND DISCUSSION

Genome-Wide Identification of Alternative Proteins in *Drosophila melanogaster*

In order to identify new alternative proteins in *Drosophila melanogaster*, we developed a customized peptidomics workflow (**Figure 1**). We used a combination of protein extraction and small proteins enrichment protocol as it was previously shown to increase the number of AltProts identified by mass spectrometry (Ma et al., 2016; Cardon et al., 2020). Extensive fractionation, using high pH reverse-phase chromatography, as well as specific enrichment of short proteins, through SDS-PAGE, ultrafiltration, acid precipitation, and reverse-phase chromatography, was employed to retrieve AltProts from adult flies, 0–24 h embryos, and S2 cells (**Figure 1**). We also re-analyzed MS data available in public repositories. In total, 1,068 MS files were analyzed using optimized MaxQuant parameters and the OpenProt predicted AltProts database

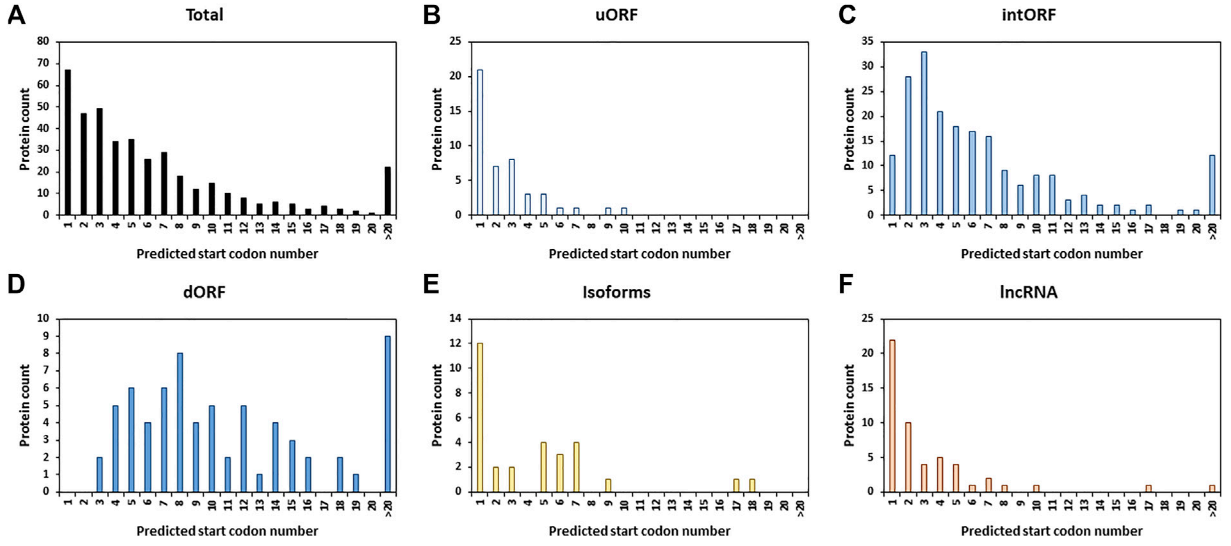


FIGURE 3 | Alternative proteins are not mainly produced from the first predicted ORF. Distribution of the alternative proteins counts depending on the order of the position of the predicted start codons for all the newly identified proteins (total) **(A)**, isoforms **(B)**, proteins identified from ORFs encoded by lncRNA **(C)**, uORF **(D)**, AltCDS **(E),** or dORF **(F)** on mRNA.

(**Figure 1**). In total, 401 AltProts and 8,615 RefProts (including 267 RefProts containing less than 100 amino acids annotated in UniProtKB) were identified (**Figure 2** and **Supplementary Tables S1, S2**). The identification scores obtained for the AltProts were similar to the ones measured for a typical proteomics analysis (median Andromeda score of 98.52 for AltProts and 101.97 for RefProts) (**Supplementary Figure S1A**). The majority of the AltProts identified here are short proteins (88.8% of AltProts are less than 150 amino acids) (**Supplementary Figure S1B**). Comparing the AltProts identified to the ones with MS evidence in OpenProt and a recent article (Wang et al., 2022), only two were common to the three datasets, 29 were found in at least two datasets and 374 new AltProts were identified in this study (**Supplementary Figure S1C**). The low overlap between the datasets might be explained by the different sample types and extraction and fractionation protocols used (Cardon et al., 2020). Amongst the 401 non-annotated proteins identified, 30 were new isoforms (**Figure 2**). As defined in OpenProt (https://www.openprot.org/), we refer here as isoform (or novel isoform) to any non-annotated proteins that share some homology with a RefProt (either partially overlapping coding sequences, although only isoform unique peptides are used for their identification). Next, we looked at the RNA types and regions from which AltProts are produced (**Figure 2**). We used a recently suggested nomenclature (Mudge et al., 2021) to refer to the types of ORFs encoding the AltProts (**Figure 2**). Surprisingly, whereas pioneering studies identified non-coding RNA or untranslated regions of mRNAs as the main sources of AltProts (Plaza et al., 2017), the majority of AltProts identified in our study are produced from mRNA and more particularly from alternative reading frames in canonical CDS (intORFs). With more than 300 AltProts produced from uORFs, intORFs, or dORFs, our data advocate toward a model in which several proteins can be produced from one mRNA in *Drosophila melanogaster* (**Figure 2**). Of note, 52 AltProts are produced from previously predicted long non–coding RNA (**Figure 2**), including one AltProt encoded by a precursor of miRNA (pri-miRNA) (**Figure 2**), supporting the idea that miPEPs (miRNA-encoded peptides) are expressed in flies (Immarigeon et al., 2021; Montigny et al., 2021). Regarding the sources of the production of AltProts, and more particularly the chromosomes they are produced from, a distribution similar to the predicted AltProts distribution from OpenProt was observed (**Supplementary Figures S2A,B**), although slight differences could be noticed. The proportion of AltProts produced from the chromosomes 2R and 3L was higher than expected contrary to the chromosomes four and X where a lower proportion of AltProts was identified (**Supplementary Figures S2A,B**). Interestingly, the proportion of new isoforms and Altprots synthesized from uORFs and lncRNA was more represented than expected (**Supplementary Figures S2C–H**).

We next looked at the position of the start codon of the 401 AltProts identified. Surprisingly, only 16.7% of the AltProts identified here are produced from the first predicted start codon (**Figure 3A**). As expected, AltProts produced from uORFs are synthesized from the first start codon more frequently than AltProts produced from intORFs and dORFs (45.7 versus 5.9% and 0%, respectively) (**Figures 3B–D** and **Supplementary Figure S3**). Interestingly, new isoforms and AltProts produced from lncRNA follow a pattern similar to uORFs with 40 and 42.3% of these proteins being synthesized from the first start codon (**Figures 3B,E–F** and **Supplementary Figure S3**). These data highlight that, although the translation of AltProts from the first ORF on an RNA is the most probable (notably for AltProts produced from lncRNA), 334 of the new

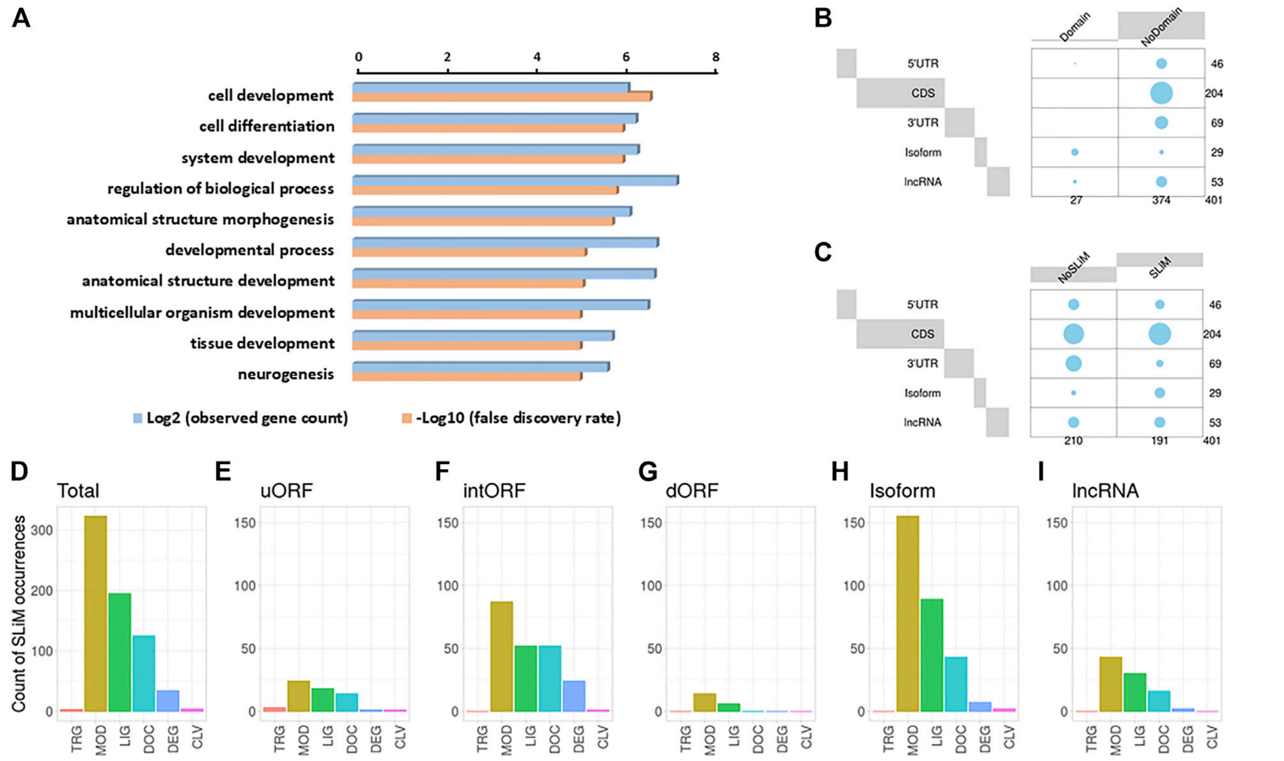


FIGURE 4 | Identification of protein domains and short linear motifs in alternative proteins. **(A)**. Gene Ontology term analysis of the host genes of the alternative proteins identified. **(B)**. Balloon plots showing the presence or lack of protein domain for the different types of ORFs, from which the AltProts are produced. The area is proportional to the frequency. **(C)**. Balloon plots showing the presence or lack of SLiMs for the different types of ORFs, from which the AltProts are produced. The area is proportional to the frequency. D-I. Counts of the different classes of SLiMs in the AltProts identified in this study **(D)** or identified from uORFs **(E)**, intORFs **(F)**, dORFs **(G)**, isoforms **(H),** or lncRNA **(I)**. SLiM classes are targeting sites for subcellular localization (TRG), post-translational modification sites (MOD), ligand binding sites (LIG), docking sites (DOC), degradation sites (DEG), and proteolytic cleavage sites (CLV).

proteins identified here are translated from further ORFs on RNAs. Notably, 22 AltProts are synthesized from the 20th predicted ORF or beyond (**Figure 3A**).

Structural Properties of Alternative Proteins in *Drosophila melanogaster*

Next, the chemical characteristics of the AltProts identified were investigated. First, we looked at the size distribution of the AltProts depending on the type of ORF they are synthesized from (**Supplementary Figures S4A–C**). As expected, isoforms are longer than other AltProts (median length of 221.5 and 52 amino acids, respectively) (**Supplementary Figures S4A–C**). Within AltProts, proteins produced from lncRNA are slightly longer than the alternative proteins synthesized from mRNA (median length of 67, 49, 52.5, and 44 for AltProts from lncRNA, uORFs, intORFs, and dORFs, respectively) (**Supplementary Figures S4A–C**).

Comparing the isoelectric point (pI) of the different classes of AltProts revealed that isoforms have lower pI than other AltProts ($p < 9.04 \times 10^{-5}$) (**Supplementary Figure S4D**). In addition, AltProts produced from intORFs tend to have higher pI than the other AltProts ($p < 0.0013$) (**Supplementary Figure S4D**). This might be explained by the fact that the overall amino acid composition of AltProts produced from intORFs differs from other AltProts (**Supplementary Figure S5**). The former has more arginine, alanine, and tryptophan and less asparagine, lysine, and glutamic acid ($p < 2.2 \times 10^{-16}$) (**Supplementary Figure S5**). This difference in the composition might point toward specific functions of AltProts produced from intORFs.

We then performed a Gene Ontology (GO) analysis on the host genes, from which the AltProts are produced, to gain some insight into the possible functions of the newly discovered proteins (**Figure 4A**). Interestingly, the most significant terms enriched were cell development (FDR = 2.4×10^{-7}) and cell differentiation (FDR = 9.6×10^{-7}), suggesting that the AltProts identified in this study might have functions related to developmental processes (**Figure 4A** and **Supplementary Figure S6**). These pathways are mainly enriched in host genes from AltProts produced from intORFs and dORFs (**Supplementary Figures S7, S8**). No pathway was found enriched in AltProts produced from uORFs or isoforms (**Supplementary Figure S9**).

In order to dig deeper into the possible role of the AltProts of *Drosophila melanogaster*, several prediction tools were used to identify potential protein domains, disordered regions, or subcellular localization signals. Looking at protein domains, InterPro (Blum et al., 2021) predicted that 27 of the AltProts identified might have one or more protein domains (**Figure 4B** and **Supplementary Table S3**). Analysis using the TMHMM—2.0 (Krogh et al., 2001) and SignalP—5.0 (Almagro Armenteros et al., 2019) software identified possible transmembrane domains and signal peptides for 33 and nine AltProts, respectively. Interestingly, the AltProt IP_1410397 encoded by the host gene CG15784 had both a signal peptide and a transmembrane domain predicted (both with probabilities >0.8) in the first 30 amino acids of its sequence (**Supplementary Figure S10**).

We next looked for the presence of short linear motifs (SLiMs) in the AltProts identified. SLiMs are functional short stretches of protein sequence that are generally involved in protein–protein interactions (Hraber et al., 2020). A total of 684 SLiMs were mapped on 191 AltProts (**Figures 4C,D** and **Supplementary Table S4**). Most of the SLiMs retrieved belong to the post-translational modification sites (MOD) (enriched in isoforms, Benjamini–Hochberg adjusted *p*-value = 3.5 0.10^{-4}, and odd ratio = 5.41), ligand binding sites (LIG) (especially enriched in isoforms, Benjamini–Hochberg adjusted *p*-value = 1.42 0.10^{-7}, and odds ratio = 11.04), and docking site (DOC) classes (47, 29, and 18% of the SLiMs identified, respectively) (**Figure 4D**, **Supplementary Figure S11A**, and **Supplementary Table S4**). The most represented SLiMs are Polo-like kinase1 and four phosphosite motifs (MOD_PlK_1 and MOD_Plk_4, found on 64 and 100 AltProts, respectively), cyclin N-terminal domain docking motifs (DOC_CYCLIN_RXL_1, found on 38 AltProts), and Atg8 protein family ligand motifs (LIG_LIR_Gen_1, found on 51 AltProts) (**Supplementary Figure S11A** and **Supplementary Table S4**), suggesting a possible role of these AltProts in *Drosophila* cell cycle and autophagy. Importantly, isoforms were the only class of AltProts which displayed a significant enrichment in SLiMs (**Supplementary Figures S11B,C** and **Supplementary Table S5**). Looking at each type of ORFs, slight differences in SLiM classes could be observed (**Figures 4E–I**). The SLiM class targeting sites for subcellular localization (TRG) were identified only on one AltProt produced from an uORF (**Figure 4E**). It was surprising to notice that AltProts from dORFs only have 20 SLiMs detected on 11 AltProts (**Figure 4G**). In addition, this type of AltProt does not seem to bear any protein domain and although all the SLiMs identified are from the MOD and LIG classes, the number of SLiMs from these classes detected is still lower than expected (Benjamini–Hochberg adjusted *p*-value = 0.009 and 0.0006; odds ratio = 0.35 and 0.17, respectively) (**Supplementary Table S5**). These results suggest that most of the AltProts produced from dORFs are less susceptible than other AltProts to carry particular functions based on domain prediction and that the ORF itself might be mainly involved in the regulation of protein translation as recently proposed (Wu et al., 2020).

Next, the IUPred2A software was used to predict disordered regions within AltProts. Around 50% of the AltProts identified in our study contained one or more predicted disordered region (**Supplementary Figure S12**). A higher proportion of isoforms (70%) and AltProts produced from uORFs (58.7%) and intORFs (55.9%) tend to have disordered regions compared to lncRNA (38.5%) or dORFs (26.1%) (**Supplementary Figure S12**). This is in agreement with a recent report on plants, which also predicted that numerous non-annotated short proteins might contain disordered regions, transmembrane domains, or signal peptides (Fesenko et al., 2021).

DeepLoc (Almagro Armenteros et al., 2017) was used to predict the possible subcellular localization of Altprots. A potential localization was assigned to 79 out of 401 with a probability higher than 0.8 (**Figure 5A** and **Supplementary Table S5**). Surprisingly, 35 of these AltProts were predicted to be mitochondrial, 18 extracellular, and 17 nuclear (**Figure 5B** and **Supplementary Table S6**). Only seven Altprots are predicted to be cytoplasmic, one potentially localized in the Golgi and one in the endoplasmic reticulum (**Figure 5B** and **Supplementary Table S6**). When comparing the predicted subcellular localization of Altprots produced from mRNAs and their corresponding RefProts, only 18% (12 out of 67) were concordant (**Supplementary Figure S13**). In order to validate the prediction from DeepLoc, the AltProt and RefProt of the gene *CG34150*, tagged with a red fluorescent protein (RFP) and a green fluorescent protein (GFP), respectively, were transfected in S2 cells and co-expressed under the same *actin* promoter. Confocal imaging revealed that, in agreement with the DeepLoc prediction, both proteins are colocalized in S2 cells (**Figure 5C**). Similarly, tagged versions of the AltProt and RefProt of the gene *CG2650*, for which no subcellular localizations were predicted in animal cells, were also expressed in S2 cells (**Figure 5D**). Surprisingly, colocalization could be observed in certain cells whereas other cells showed different localization patterns between the two proteins in the S2 cells within the same experiment (**Figure 5D**). This might be indicative that the AltProt and RefProt of CG2650 are colocalized under particular cellular conditions (e.g. specific cell cycle stages…). These experiments also showed that the two *CG34150* and *CG2650* AltProts are expressed despite the presence of the ATG of the canonical ORF, confirming peptide detection observed in MS analysis.

Overall, these data corroborate previous observations in humans suggesting that AltProts might have independent functions or roles related to their corresponding RefProts (Chen J. et al., 2020). Here, we identified 235 AltProts for which at least a protein domain, a SLiM, or a subcellular localization was predicted (**Figures 4B**,D, **5A**). Although further functional experiments would be necessary to better understand the role of these AltProts, these predictions provide first hints regarding the functions of the Altprots identified in this study.

Alternative Proteins Are Not Necessarily Less Abundant Than Canonical Proteins

We next wondered if AltProts can be more abundant than their corresponding RefProts as previously shown for the human alternative protein altMiD51 (Delcourt et al., 2018).

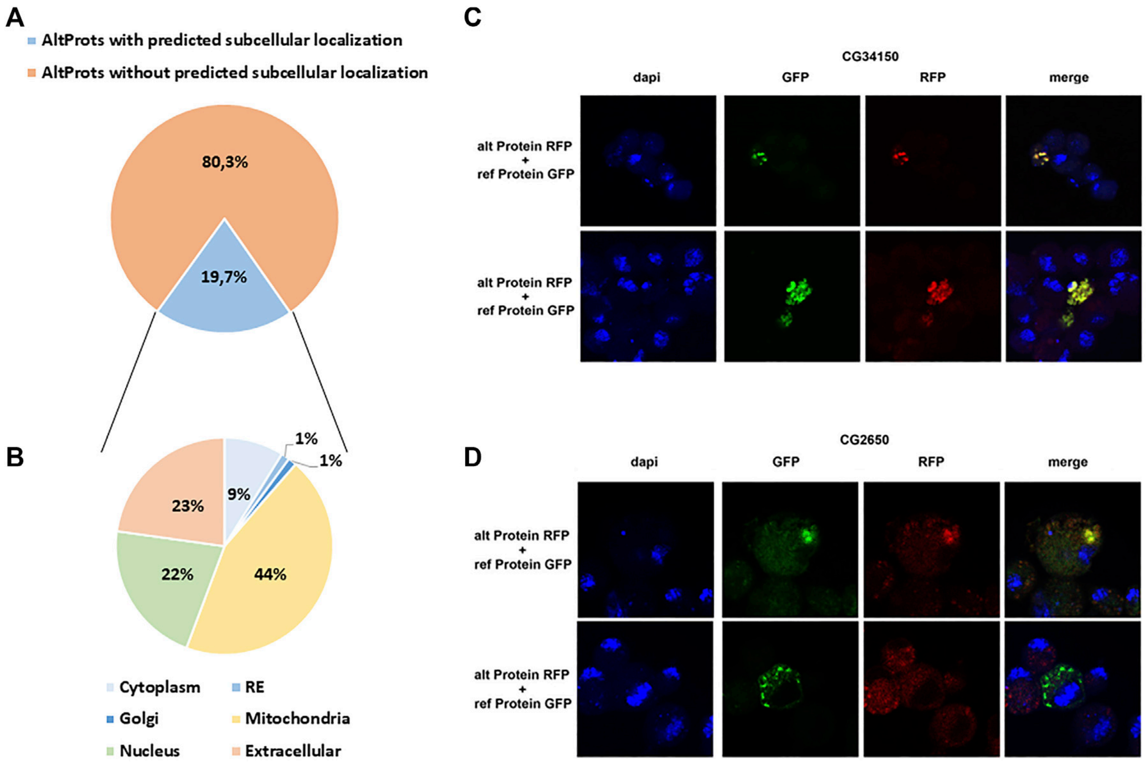


FIGURE 5 | Prediction of alternative proteins subcellular localization. **(A)**. Proportion of AltProts with or without predicted subcellular localization. **(B)**. Proportion of AltProts predicted to be localized in different organelles. **(C)**. Confocal microscopy analysis of the co-expression of the AltProt (alt Protein) (with the sequence of the RefProt before the AltProt starting codon) and RefProt (ref Protein) encoded by the CG34150 gene, tagged with RFP and GFP, respectively, in S2 cells (stained with DAPI). **(D)**. Confocal microscopy analysis of the co-expression of the AltProt (alt Protein) (with the sequence of the RefProt before the AltProt starting codon) and RefProt (ref Protein) encoded by the CG2650 gene, tagged with RFP and GFP, respectively, in S2 cells (stained with DAPI).

Comparing the intensities measured for peptides from AltProts and RefProts from total lysates and high pH reverse-phase fractionation (no specific AltProts enrichment, see Material and Methods section protocol 2 as well as data from Müller et al. (2020), the peptides from the latter were slightly more intense (1.96 fold difference of the average peptides intensities measured for AltProts and RefProts, **Supplementary Figure S14A**). This implies that some AltProts might be as abundant as RefProts in *Drosophila*. The iBAQ values, which represent an approximation of the abundance of a protein (Krey et al., 2014), were used to compare the abundance of AltProts with the abundance of their corresponding RefProts in total lysates and high pH reverse-phase fractionation experiments (**Figures 6A,B**). Out of the 39 pairs of AltProts/RefProts for which iBAQ values were measured, 22 did not show any difference in abundance between AltProts and RefProts expression levels (less than 10-fold difference between the iBAQ values), whereas 11 AltProts were more abundant (**Figures 6A,B**). Only six RefProts were more abundant than their corresponding AltProts (**Figures 6A,B**). This trend was observed in two independent datasets (**Supplementary Figure S14B**) and we did not observe any bias based on the length of the AltProts (**Supplementary Figure S15**). These data reveal that, in several cases, alternative proteins are actually the main protein produced from their corresponding genes.

Developmentally Timed and Stress-Specific Production of Alternative Proteins

Next, the expression of AltProts was compared to monitor potential changes between embryos and adult flies (Material and Methods protocol 1 and 2, **Figures 6C,D**). All 14 AltProts for which we obtained quantitative data in at least two biological replicates were more abundant in one developmental stage (**Figures 6C,D**). Three AltProts were identified both in embryos and adult flies but were at least three times more abundant in embryos (**Figures 6C,D**). The remaining 11 AltProts were identified only in one stage (**Figures 6C,D**), suggesting that the expression of most of the AltProts quantified here is developmentally timed. Four AltProts were identified only in adult flies whereas seven were specific to embryo samples, including two AltProts produced from lncRNA (**Figures 6C,D**).

We also tested whether the expression of AltProts varies upon stress. The embryos were treated with heat-shock at 37°C for up to 3 h or kept at 25°C and analyzed to identify alternative proteins.

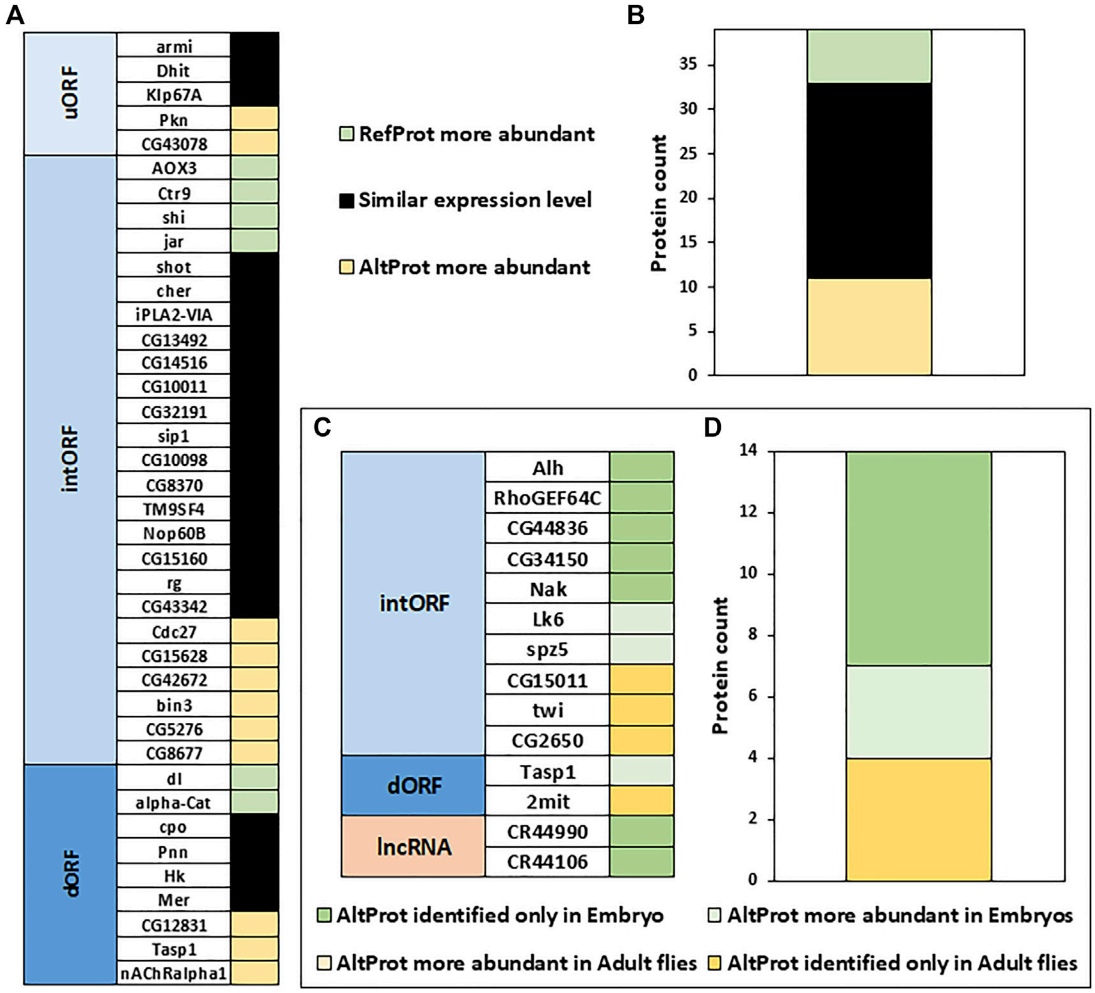


FIGURE 6 | Alternative proteins can be more abundant than their corresponding canonical proteins and their expression is developmentally regulated. **(A)**. Comparison of the approximation of the relative abundance measured for each couple of canonical (RefProts) and alternative (AltProts) proteins for each type of ORF using the intensity based absolute quantification approach (iBAQ). An AltProt/RefProt is considered more abundant than its counterpart if its iBAQ value is at least 10-fold higher. **(B)**. Graph shows the protein counts of AltProts are more abundant than RefProts, RefProts are more abundant than AltProts, and similar expression levels between AltProts and RefProts. **(C)**. Fold change between embryo and adult fly was calculated for each AltProt and normalized by the maximum intensity between the two samples. An AltProt/RefProt is considered more abundant in a specific developmental stage if its intensity value is at least 3-fold higher. **(D)**. Graph shows the protein counts of AltProts are more abundant, or identified only, in embryo, or in adult flies.

We were able to identify 22 AltProts in these samples, including 10 AltProts that were identified only in heat-shock–treated embryos (**Supplementary Table S7**). These results demonstrate that alternative proteins are produced under specific developmental stages or stress conditions in *D. melanogaster*.

CONCLUSION

Recent studies in humans suggested that the complexity of the genome was underestimated (Brunet et al., 2021; Ouspenskaia et al., 2021) and that many unannotated proteins might fulfill important functions, related or not to canonical proteins (Plaza et al., 2017; Chen J. et al., 2020). However, it is not clear whether this is specific to human or whether this characteristic is present in every species since we still lack deep analysis of this alternative proteome in many species, including the model organism *D. melanogaster*. In flies, until now, mainly data from ribosome profiling experiments were available to annotate putative translated alternative sORF (Aspden et al., 2014; Patraquim et al., 2020). In the present study, we developed a deep peptidomics workflow which combines several extraction methods and enrichment protocols with mass spectrometry and dedicated bioinformatics analysis to identify new alternative proteins in flies. We proved for the first time the existence of 374 AltProts predicted in OpenProt (**Figure 2**), significantly increasing the repertoire of not yet annotated proteins in *D. melanogaster*. Many of these AltProts even escaped from ribosome profiling experiments as they are encoded by alternative frames within the annotated CDS. Contrasting with these results, we did not find many unannotated proteins with a coding sequence of more than 100 codons, revealing that the annotation of proteins with ORF of 100 codons or more is precise and reliable. On the other hand, our study shows that many proteins of less than 100 amino acids remain to be discovered, especially considering the fact that we did not search for alternative proteins of less than 30 amino acids, which are known to be expressed and functional in *D. melanogaster* (Magny et al., 2013; Zanet et al., 2015; Immarigeon et al., 2021; Montigny et al., 2021) and would require further investigation. Interestingly, these AltProts are not necessarily produced from the first predicted ORF on a RNA (**Figure 3**), one spectacular result came from an AltProt being synthesized from the 134th predicted ORF on the dumpy mRNA (**Supplementary Table S1**). Another key observation is that more than 300 mRNAs actually encode more than one protein (**Figure 2**). The main source of production of AltProts in *Drosophila melanogaster* is alternative frames in canonical coding sequences (intORFs) (**Figure 2**) possibly a specificity of *Drosophila* in humans and mice; AltProts are produced mainly from lncRNA (https://www.openprot.org/). Through our peptidomics workflow we showed that 52 RNA, previously described as non-coding, actually encode a protein and should be reannotated as mRNA instead of lncRNA (**Figure 2**). Regarding potential functions of the identified AltProts, protein domain, SLiMs, or subcellular localization were predicted for 235 of them (**Figures 4B**,C, **5A**) pointing toward potential functions for these small proteins. However, the lack of predicted protein domains and low number of SLiMs identified on dORFs implies that the AltProts produced from these ORFs might not be functional. Fluorescence confocal microscopy confirmed the colocalization of the *CG34150* AltProt and RefProt and showed that the *CG2650* AltProt and RefProt can colocalize under certain conditions (**Figures 5C,D**). The comparison of the abundance (using the iBAQ value as an approximation) of alternative and canonical proteins revealed that AltProts are not necessarily less abundant and might actually be the main product of several genes (**Figures 6A,B** and **Supplementary Figure S14B**). This result rules out that the AltProts identified in our study are transient and unstable products of translation. These data suggest that it might be worth reconsidering the phenotypes observed in certain mutants in *D. melanogaster* as they might be mediated by the mutation/deletion of the alternative protein rather than the canonical one. Finally, several AltProts were identified in only specific developmental stages or upon heat shock, implying that their expression is finely tuned during *D. melanogaster* development or under stress conditions (**Figures 6C,D**). These proteins might have important functions during development or heat-shock response, hence requiring further functional investigation.

DATA AVAILABILITY STATEMENT

All the mass spectrometry data have been deposited with the MassIVE repository with the dataset identifier: MSV000088656.

AUTHOR CONTRIBUTIONS

BF and SP conceived the project and supervised the research. BF wrote the manuscript. BF, CD, AM, and DK performed the experiments. BF, MD, and MC performed mass spectrometry analysis. BF, SC, CP, and CB analyzed the data. OB-S, SR, J-PC, and KL contributed to the data analyses and manuscript discussion. All authors read, edited, and approved the final manuscript.

FUNDING

This work has been supported by the Fondation ARC pour la recherche sur le cancer. BF is funded by Biotechnology and Biological Science Research Council (ref: BB/L002817/1) and a long term EMBO fellow (ALTF 1204-2015) cofounded by Marie Curie Actions (LTFCOFUND 2013, GA-2013-609409). SC is funded by a Fondation pour la Recherche Médicale fellowship (FDT202106013072). DK is funded by Biotechnology and Biological Science Research Council (ref: BB/L002817/1). The work was funded in part by grants from the Région Occitanie, European funds (Fonds Européens de Développement Régional, FEDER), Toulouse Métropole, and the French Ministry of Research with the Investissement d'Avenir Infrastructures Nationales en Biologie et Santé program (ProFI, Proteomics French Infrastructure project, ANR-10-INBS-08). We acknowledge the Centre de Calcul Intensif d'Aix-Marseille for granting access to its high-performance computing resources. This work was funded by the French ANR project BiomiPEP (ANR-16-CE12-0018-01).

ACKNOWLEDGMENTS

We would like to thank Aurelie Le Ru for her help with confocal microscopy.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2022.901351/full#supplementary-material

REFERENCES

Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H., and Winther, O. (2017). DeepLoc: Prediction of Protein Subcellular Localization Using Deep Learning. *Bioinforma. (Oxford, England)* 33, 3387–3395. doi:10.1093/bioinformatics/btx431

Almagro Armenteros, J. J., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., et al. (2019). SignalP 5.0 Improves Signal Peptide Predictions Using Deep Neural Networks. *Nat. Biotechnol.* 37, 420–423. doi:10.1038/s41587-019-0036-z

Aspden, J. L., Eyre-Walker, Y. C., Phillips, R. J., Amin, U., Mumtaz, M. A. S., Brocard, M., et al. (2014). Extensive Translation of Small Open Reading Frames Revealed by Poly-Ribo-Seq. *eLife* 3, e03528. doi:10.7554/eLife.03528

Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., et al. (2021). UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Res.* 49, D480–D489. doi:10.1093/nar/gkaa1100

Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., et al. (2021). The InterPro Protein Families and Domains Database: 20 Years on. *Nucleic Acids Res.* 49, D344–D354. doi:10.1093/nar/gkaa977

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, D403–D410. doi:10.1093/nar/gky936

Brunet, M. A., Leblanc, S., and Roucou, X. (2020). Reconsidering Proteomic Diversity with Functional Investigation of Small ORFs and Alternative ORFs. *Exp. Cell Res.* 393, 112057. doi:10.1016/j.yexcr.2020.112057

Brunet, M. A., Lucier, J.-F., Levesque, M., Leblanc, S., Jacques, J.-F., Al-Saedi, H. R. H., et al. (2021). OpenProt 2021: Deeper Functional Annotation of the Coding Potential of Eukaryotic Genomes. *Nucleic Acids Res.* 49, D380–D388. doi:10.1093/nar/gkaa1036

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

Cassidy, L., Kaulich, P. T., Maaß, S., Bartel, J., Becher, D., and Tholey, A. (2021). Bottom-up and Top-Down Proteomic Approaches for the Identification, Characterization, and Quantification of the Low Molecular Weight Proteome with Focus on Short Open Reading Frame-Encoded Peptides. *Proteomics* 21, 2100008. doi:10.1002/pmic.202100008

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

Chen, Y., Li, D., Fan, W., Zheng, X., Zhou, Y., Ye, H., et al. (2020b). PsORF: a Database of Small ORFs in Plants. *Plant Biotechnol. J.* 18, 2158–2160. doi:10.1111/pbi.13389

Choteau, S. A., Wagner, A., Pierre, P., Spinelli, L., and Brun, C. (2021). MetamORF: A Repository of Unique Short Open Reading Frames Identified by Both Experimental and Computational Approaches for Gene and Metagene Analyses. *Database* 2021, baab032. doi:10.1093/database/baab032

Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N., and Mann, M. (2014). Accurate Proteome-Wide Label-Free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Mol. Cell. Proteomics* 13, 2513–2526. doi:10.1074/mcp.M113.031591

Delcourt, V., Brunelle, M., Roy, A. V., Jacques, J.-F., Salzet, M., Fournier, I., et al. (2018). The Protein Coded by a Short Open Reading Frame, Not by the Annotated Coding Sequence, Is the Main Gene Product of the Dual-Coding Gene MIEF1. *Mol. Cell. Proteomics* 17, 2402–2411. doi:10.1074/mcp.RA118.000593

Edwards, R. J., Paulsen, K., Aguilar Gomez, C. M., and Pérez-Bercoff, Å. (2020). Computational Prediction of Disordered Protein Motifs Using SLiMSuite. *Methods Mol. Biol.* 2141, 37–72. doi:10.1007/978-1-0716-0524-0_3

Fabre, B., Combier, J.-P., and Plaza, S. (2021). Recent Advances in Mass Spectrometry-Based Peptidomics Workflows to Identify Short-Open-Reading-Frame-Encoded Peptides and Explore Their Functions. *Curr. Opin. Chem. Biol.* 60, 122–130. doi:10.1016/j.cbpa.2020.12.002

Fabre, B., Korona, D., Groen, A., Vowinckel, J., Gatto, L., Deery, M. J., et al. (2016a). Analysis of *Drosophila M* Proteome Dynamics during Embryonic Development by a Combination of Label-Free Proteomics Approaches. *Proteomics* 16, 2068–2080. doi:10.1002/pmic.201500482

Fabre, B., Korona, D., Lees, J. G., Lazar, I., Livneh, I., Brunet, M., et al. (2019). Comparison of *Drosophila M* Embryo and Adult Proteome by SWATH-MS Reveals Differential Regulation of Protein Synthesis, Degradation Machinery, and Metabolism Modules. *J. Proteome Res.* 18, 2525–2534. doi:10.1021/acs.jproteome.9b00076

Fabre, B., Korona, D., Mata, C. I., Parsons, H. T., Deery, M. J., Hertog, M. L. A. T. M., et al. (2017). Spectral Libraries for SWATH-MS Assays for *Drosophila M* and Solanum Lycopersicum. *Proteomics* 17, 1700216. doi:10.1002/pmic.201700216

Fabre, B., Korona, D., Nightingale, D. J. H., Russell, S., and Lilley, K. S. (2016b). SWATH-MS Data of *Drosophila M* Proteome Dynamics during Embryogenesis. *Data Brief* 9, 771–775. doi:10.1016/j.dib.2016.10.009

Fabre, B., Korona, D., Nightingale, D. J. H., Russell, S., and Lilley, K. S. (2016c). SWATH-MS Dataset of Heat-Shock Treated *Drosophila M* Embryos. *Data Brief* 9, 991–995. doi:10.1016/j.dib.2016.11.028

Fabre, B., Lambour, T., Bouyssié, D., Menneteau, T., Monsarrat, B., Burlet-Schiltz, O., et al. (2014). Comparison of Label-Free Quantification Methods for the Determination of Protein Complexes Subunits Stoichiometry. *EuPA Open Proteom.* 4, 82–86. doi:10.1016/j.euprot.2014.06.001

Fesenko, I., Shabalina, S. A., Mamaeva, A., Knyazev, A., Glushkevich, A., Lyapina, I., et al. (2021). A Vast Pool of Lineage-Specific Microproteins Encoded by Long Non-Coding RNAs in Plants. *Nucleic Acids Res.* 49, 10328–10346. doi:10.1093/nar/gkab816

Geladaki, A., Kočevar Britovšek, N., Breckels, L. M., Smith, T. S., Vennard, O. L., Mulvey, C. M., et al. (2019). Combining LOPIT with Differential Ultracentrifugation for High-Resolution Spatial Proteomics. *Nat. Commun.* 10, 1–15. doi:10.1038/s41467-018-08191-w

Hazarika, R. R., De Coninck, B., Yamamoto, L. R., Martin, L. R., Cammue, B. P. A., and Van Noort, V. (2017). ARA-PEPs: A Repository of Putative SORF-Encoded Peptides in *Arabidopsis T. BMC Bioinforma.* 18, 37. doi:10.1186/s12859-016-1458-y

Hraber, P., O'Maille, P. E., Silberfarb, A., Davis-Anderson, K., Generous, N., McMahon, B. H., et al. (2020). Resources to Discover and Use Short Linear Motifs in Viral Proteins. *Trends Biotechnol.* 38, 113–127. doi:10.1016/j.tibtech.2019.07.004

Immarigeon, C., Frei, Y., Delbare, S. Y. N., Gligorov, D., Machado Almeida, P., Grey, J., et al. (2021). Identification of a Micropeptide and Multiple Secondary Cell Genes that Modulate Drosophila Male Reproductive Success. *Proc. Natl. Acad. Sci. U.S.A.* 118 (15), e2001897118. doi:10.1073/pnas.2001897118

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: Genome-Scale Protein Function Classification. *Bioinformatics* 30, 1236–1240. doi:10.1093/bioinformatics/btu031

Krey, J. F., Wilmarth, P. A., Shin, J.-B., Klimek, J., Sherman, N. E., Jeffery, E. D., et al. (2014). Accurate Label-Free Protein Quantitation with High- and Low-Resolution Mass Spectrometers. *J. Proteome Res.* 13, 1034–1044. doi:10.1021/pr401017h

Krogh, A., Larsson, B., Von Heijne, G., and Sonnhammer, E. L. L. (2001). Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes. *J. Mol. Biol.* 305, 567–580. doi:10.1006/jmbi.2000.4315

Kumar, M., Gouw, M., Michael, S., Sámano-Sánchez, H., Pancsa, R., Glavina, J., et al. (2020). ELM-the Eukaryotic Linear Motif Resource in 2020. *Nucleic Acids Res.* 48, D296–D306. doi:10.1093/nar/gkz1030

Lautenbacher, L., Samaras, P., Muller, J., Grafberger, A., Shraideh, M., Rank, J., et al. (2021). ProteomicsDB: toward a FAIR Open-Source Resource for Life-Science Research. *Nucleic Acids Res.* 50, D1541–D1552. doi:10.1093/nar/gkab1026

Li, Y., Zhou, H., Chen, X., Zheng, Y., Kang, Q., Hao, D., et al. (2021). SmProt: A Reliable Repository with Comprehensive Annotation of Small Proteins Identified from Ribosome Profiling. *Genomics Proteomics Bioinforma.* 19, 602–610. doi:10.1016/j.gpb.2021.09.002

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, 3967–3975. doi:10.1021/acs.analchem.6b00191

Magny, E. G., Pueyo, J. I., Pearl, F. M. G., Cespedes, M. A., Niven, J. E., Bishop, S. A., et al. (2013). Conserved Regulation of Cardiac Calcium Uptake by Peptides Encoded in Small Open Reading Frames. *Science* 341, 1116–1120. doi:10.1126/science.1238802

Mata, C. I., Fabre, B., Hertog, M. L. A. T. M., Parsons, H. T., Deery, M. J., Lilley, K. S., et al. (2017). In-Depth Characterization of the Tomato Fruit Pericarp Proteome. *Proteomics* 17, 1600406. doi:10.1002/pmic.201600406

Menneteau, T., Fabre, B., Garrigues, L., Stella, A., Zivkovic, D., Roux-Dalvai, F., et al. (2019). Mass Spectrometry-Based Absolute Quantification of 20S Proteasome Status for Controlled *Ex-Vivo* Expansion of Human Adipose-Derived Mesenchymal Stromal/Stem Cells. *Mol. Cell. Proteomics* 18, 744–759. doi:10.1074/mcp.RA118.000958

Mészáros, B., Erdős, G., and Dosztányi, Z. (2018). IUPred2A: Context-Dependent Prediction of Protein Disorder as a Function of Redox State and Protein Binding. *Nucleic Acids Res.* 46, W329–W337. doi:10.1093/nar/gky384

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., et al. (2021). Pfam: The Protein Families Database in 2021. *Nucleic Acids Res.* 49, D412–D419. doi:10.1093/nar/gkaa913

Montigny, A., Tavormina, P., Duboe, C., San Clémente, H., Aguilar, M., Valenti, P., et al. (2021). Drosophila Primary microRNA-8 Encodes a microRNA-Encoded Peptide Acting in Parallel of miR-8. *Genome Biol.* 22, 1–21. doi:10.1186/s13059-021-02345-8

Mudge, J. M, Ruiz-Orera, J., Prensner, J. R., Brunet, M. A., Gonzalez, J. M., Magrane, M., et al. (2021). A Community-Driven Roadmap to Advance Research on Translated Open Reading Frames Detected by Ribo-Seq. bioRxiv. doi:10.1101/2021.06.10.447896

Müller, J. B., Geyer, P. E., Colaço, A. R., Treit, P. V., Strauss, M. T., Oroshi, M., et al. (2020). The Proteome Landscape of the Kingdoms of Life. *Nature* 582, 592–596. doi:10.1038/s41586-020-2402-x

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, D497–D502. doi:10.1093/nar/gkx1130

Omenn, G. S., Lane, L., Overall, C. M., Paik, Y.-K., Cristea, I. M., Corrales, F. J., et al. (2021). Progress Identifying and Analyzing the Human Proteome: 2021 Metrics from the HUPO Human Proteome Project. *J. Proteome Res.* 20, 5227–5240. doi:10.1021/acs.jproteome.1c00590

Ouspenskaia, T., Law, T., Clauser, K. R., Klaeger, S., Sarkizova, S., Aguet, F., et al. (2021). Unannotated Proteins Expand the MHC-I-Restricted Immunopeptidome in Cancer. *Nat. Biotechnol.* 40, 209–217. doi:10.1038/s41587-021-01021-3

Patraquim, P., Mumtaz, M. A. S., Pueyo, J. I., Aspden, J. L., and Couso, J.-P. (2020). Developmental Regulation of Canonical and Small ORF Translation from mRNAs. *Genome Biol.* 21, 128. doi:10.1186/s13059-020-02011-5

Payros, D., Alonso, H., Malaga, W., Volle, A., Mazères, S., Déjean, S., et al. (2021). Rv0180c Contributes to *Mycobacterium Tuberculosis* Cell Shape and to Infectivity in Mice and Macrophages. *PLoS Pathog.* 17, e1010020. doi:10.1371/journal.ppat.1010020

Plaza, S., Menschaert, G., and Payre, F. (2017). In Search of Lost Small Peptides. *Annu. Rev. Cell Dev. Biol.* 33, 391–416. doi:10.1146/annurev-cellbio-100616-060516

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., et al. (2021). The STRING Database in 2021: Customizable Protein-Protein Networks, and Functional Characterization of User-Uploaded Gene/Measurement Sets. *Nucleic Acids Res.* 49, D605–D612. doi:10.1093/nar/gkaa1074

Wan, C., Borgeson, B., Phanse, S., Tu, F., Drew, K., Clark, G., et al. (2015). Panorama of Ancient Metazoan Macromolecular Complexes. *Nature* 525, 339–344. doi:10.1038/nature14877

Wang, Z., Pan, N., Yan, J., Wan, J., and Wan, C. (2022). Systematic Identification of Microproteins during the Development of *Drosophila Melanogaster*. *J. Proteome Res.* 21, 1114–1123. doi:10.1021/acs.jproteome.2c00004

Wessels, H.-H., Imami, K., Baltz, A. G., Kolinski, M., Beldovskaya, A., Selbach, M., et al. (2016). The mRNA-Bound Proteome of the Early Fly Embryo. *Genome Res.* 26, 1000–1009. doi:10.1101/gr.200386.115

Wright, B. W., Yi, Z., Weissman, J. S., and Chen, J. (2021). The Dark Proteome: Translation from Noncanonical Open Reading Frames. *Trends Cell Biol.* 32, 243–258. doi:10.1016/j.tcb.2021.10.010

Wu, Q., Wright, M., Gogol, M. M., Bradford, W. D., Zhang, N., and Bazzini, A. A. (2020). Translation of Small Downstream ORFs Enhances Translation of Canonical Main Open Reading Frames. *EMBO J.* 39, 1–13. doi:10.15252/embj.2020104763

Zanet, J., Benrabah, E., Li, T., Pélissier-Monier, A., Chanut-Delalande, H., Ronsin, B., et al. (2015). Pri sORF Peptides Induce Selective Proteasome-Mediated Protein Processing. *Science* 349, 1356–1358. doi:10.1126/science.aac5677

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.