



**HAL**  
open science

# Exploration of the functions of short open reading frames in monocytes

Sébastien A Choteau

► **To cite this version:**

Sébastien A Choteau. Exploration of the functions of short open reading frames in monocytes. Bio-Informatique, Biologie Systémique [q-bio.QM]. Aix-Marseille Université, 2022. Français. NNT : . tel-04413066

**HAL Id: tel-04413066**

**<https://hal.science/tel-04413066>**

Submitted on 23 Jan 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

# THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université  
le Vendredi 09 décembre 2022 par

**Sébastien A. CHOTEAU**

Exploration of the functions of short open reading frames in  
monocytes

**Discipline**

Biologie santé

**Spécialité**

Génomique et Bioinformatique

**École doctorale**

ED 62 - Sciences de la Vie et de la Santé

**Laboratoire/Partenaires de recherche**

TAGC - Theories and Approaches of  
Genomic Complexity

CIML - Centre d'Immunologie  
de Marseille Luminy

CENTURI - Turing Centre for Living Systems  
PhD Programme

**Composition du jury**



Sylvie RICARD-BLUM  
Université Lyon 1

Rapportrice

Yves VANDENBROUCK  
CEA

Rapporteur

Benoit BALLESTER  
TAGC

Examineur

Serge PLAZA  
LRSV

Président du jury

Christine BRUN  
TAGC

Directrice de thèse

Philippe PIERRE  
CIML

Co-Directeur de thèse

# Affidavit

I, undersigned, Sébastien A. CHOTEAU, hereby declare that the work presented in this manuscript is my own work, carried out under the scientific direction of Christine BRUN and Philippe PIERRE, in accordance with the principles of honesty, integrity and responsibility inherent to the research mission. The research work and the writing of this manuscript have been carried out in compliance with both the french national charter for Research Integrity and the Aix-Marseille University charter on the fight against plagiarism.

This work has not been submitted previously either in this country or in another country in the same or in a similar version to any other examination body.

Marseille, 01/09/2022,



Cette œuvre est mise à disposition selon les termes de la Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International.

# List of publications and participation to conferences

## List of publications related to the PhD project:

1. **Choteau SA**, Pierre P, Spinelli L, Zanzoni A, Brun C (2022). Short open reading frames-encoded peptides in human monocytes are involved in ubiquitous regulatory functions, metabolism and immunology responses. *In preparation*.
2. **Choteau SA**, Cristianini M, Maldonado K, Drets L, Boujeant M, Brun C, Spinelli L, Zanzoni A (2022). mimicINT: a workflow for microbe-host protein interaction inference. *bioRxiv*, doi: 10.1101/2022.11.04.515250.
3. Fabre B, **Choteau SA**, Dubo   C, Pichereaux C, Montigny A, Korona D, Deery MJ, Camus M, Brun C, Burlet-Schiltz O, Russell S, Combier J, Lilley KS, Plaza S (2022). In depth exploration of the alternative proteome of *Drosophila melanogaster*. *Frontiers in Cell and Developmental Biology*, doi: 10.3389/fcell.2022.901351. eCollection 2022.
4. **Choteau SA**, Wagner A, Pierre P, Spinelli L, 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*, doi: 10.1093/database/baab032. 2021:baab032.
5. Mendes A, Gigan JP, Rodriguez Rodrigues C, **Choteau SA**, Sanseau D, Barros D, Almeida C, Camosseto V, Chasson L, Paton AW, Paton JC, Arg  ello RJ, Lennon-Dum  nil A, Gatti E, Pierre P (2020). Proteostasis in dendritic cells is controlled by the PERK signaling axis independently of ATF4. *Life Science Alliance*, doi: 10.26508/lsa.202000865. 4(2):e202000865.



## **Participation to conferences and summer schools during the PhD:**

1. 2022. International Conference on Systems Biology (ICSB), Berlin - Short talk.
2. 2022. Journées Ouvertes de Bioinformatique et Mathématique (JOBIM), Rennes - Poster.
3. 2021. Cold Spring Harbor Laboratory (CSHL) Network Biology Meeting, Remote - Poster.
4. 2020. Journées Ouvertes de Bioinformatique et Mathématique (JOBIM), Remote - Poster.
5. 2020. Networks and Molecular Biology winter school, Marseille.
6. 2019. Turing Centre for Living Systems (CENTURI) summer school, Marseille.
7. 2018. European Macrophage and Dendritic cell Society (EMDS) conference, Marseille.
8. 2018-2022. Turing Centre for Living Systems (CENTURI) Scientific days, Marseille - Posters and talks.

# Abstract

The recent development of high-throughput technologies and computational methods revealed the existence of many non-canonical short open reading frames (sORFs) on most prokaryotic and eukaryotic RNAs, including presumptive non-coding RNAs. Because of their short size (< 100 codons) and the use of alternative start codons and reading frames, these ubiquitous elements have been missed for long. Functional sORF-encoded peptides (sPEPs) have been demonstrated to be involved in a wide range of biological processes, including cell physiology and proliferation, signaling, organogenesis, cell growth and death, transport, enzymatic regulation, metabolism, development, cytoskeleton organization and major histocompatibility complex class-I (MHC-I) presentation. Some of them are even taking part in disease onset (*e.g.* cancers). Nonetheless, this novel class of peptides remains poorly characterized and annotation of most sPEPs is still missing. In addition, sORFs located upstream of the canonical ORFs of mRNAs (upstream ORFs, uORFs), have been early described as *cis* regulators of the translation. By changing the efficiency of the translation initiation at the canonical ORF, uORFs participate to the translational regulatory mechanism. Indeed, some uORFs have been shown to alleviate the repression of the protein synthesis of canonical ORFs under stress. However, existing models of regulation of the translation by uORFs are still limited to a few set of genes, and the mechanisms remain cryptic for most RNAs.

This project aims to investigate the sORFs functions by (i) identifying all sORFs in human genome, (ii) exploring sPEP functions in monocytes and (iii) exploring the mechanisms of regulation of the translation by the uORFs. Human monocytes constitute a good model as they are able to express MHC molecules, whilst numerous sPEPs have been determined to be presented as self-antigens. Monocytes are playing a major role in the initiation of immune responses and derived from a bone marrow progenitor common to dendritic cells. These last have special needs regarding their translational regulation and could thus constitute an interesting model to study sORFs *cis*-regulatory functions.

To address these questions, (i) publicly available data were gathered in a repository of unique sORFs identified by complementary methods, (ii) interactions of sPEPs with canonical proteins in monocytes were predicted to identify the processes targeted by sPEPs and (iii) ribosomes' behaviours were mimicked by implementing an agent-based model to identify the most important parameters for translational regulation by uORFs.

(i) By gathering publicly available sORF data, normalizing them and summarizing redundant information, a total of 664,771 unique sORFs were identified in human. This repository allows new analyses at locus, gene, transcript and ORF levels. (ii) Our

findings suggest that sPEPs are involved in fundamental regulatory functions, both ubiquitous (protein, DNA and RNA metabolism, gene expression...) and related to specialized functions (immunological responses...). We also demonstrated that most sPEPs are preferentially interacting with annotated proteins of the same process as their cognate canonical protein. (iii) Finally, the agent-based model developed does not success yet to explain the mechanisms of translational regulation by the uORFs, but provides an adaptable tool to the scientific community for their investigation.

Keywords: short open reading frame (sORF), sORF-encoded peptide (sPEP), protein-protein interaction (PPI), translation

# Résumé

Le développement récent des technologies haut-débit et des méthodes computationnelles a révélé l'existence de nombreux petits cadres ouverts de lecture (sORFs) non canoniques sur la majorité des ARNs procaryotes et eucaryotes, y compris ceux supposés non codants. Du fait de leur petite taille (< 100 codons) et de l'usage de codons d'initiation et de cadres de lecture alternatifs, ces éléments ubiquitaires ont été négligés pendant longtemps. Il a été démontré que des peptides encodés par les sORFs (sPEPs) sont fonctionnels et impliqués dans une large gamme de processus biologiques. Ces sPEPs prennent notamment part à des activités dans la physiologie des cellules, de prolifération, signalisation, organogenèse, croissance, mort cellulaire, transport, régulation enzymatique, métabolisme, développement, organisation du cytosquelette et présentation antigénique (complexe majeur d'histocompatibilité) de classe I (MHC-I). Certains d'entre eux participent même à l'étiologie de maladies (e.g. cancers). Cependant, cette nouvelle classe de peptides demeure mal caractérisée et la majorité des sPEPs ne sont pas encore annotés. De plus, les sORFs localisés en amont des ORFs canoniques des mRNAs (appelés *upstream ORFs*, uORFs) ont été précocement décrits comme étant des éléments *cis* régulateurs de la traduction. En modifiant l'efficacité d'initiation de la traduction de l'ORF canonique, les uORFs participent à la régulation traductionnelle. En effet, certains uORFs sont capables de réduire une répression globale de la synthèse protéine des ORFs canoniques en condition de stress. Néanmoins, les modèles existants de régulation de la traduction par les uORFs sont limités à un nombre restreint de gènes et ces mécanismes demeurent cryptiques pour la majorité des ARNs.

Mon projet vise à élucider les fonctions des sORFs en (i) identifiant tous les sORFs du génome humain, (ii) explorant les fonctions des sPEPs dans les monocytes, et (iii) explorant les mécanismes de régulation de la traduction par les uORFs. Les monocytes humains constituent un modèle d'intérêt car ils sont capables d'exprimer les molécules du MHC, tandis que de nombreux sPEPs sont présentés comme antigènes du soi. Les monocytes jouent un rôle fondamental dans l'initiation de la réponse immunitaire et dérivent de progéniteurs de la moëlle osseuse communs aux cellules dendritiques. Ces dernières ont des besoins spécifiques quant à leur régulation traductionnelle et constituent donc un modèle intéressant d'étude des fonctions *cis*-régulatrices des sORFs.

Afin de répondre à ces questions, (i) des données publiées ont été recueillies dans une base de données de sORFs uniques identifiés par des méthodes complémentaires, (ii) les interactions des sPEPs avec les protéines canoniques des monocytes ont été prédites afin d'identifier les processus ciblés par les sPEPs, et (iii) le comportement des ribosomes a été reproduit par l'implémentation d'un modèle agent afin d'identifier

les paramètres les plus importants à la régulation traductionnelle par les uORFs.

(i) En recueillant les données disponibles sur les sORFs, en les normalisant, et en supprimant les entrées redondantes, un total de 664,771 sORFs uniques a été identifié chez l'humain. Ce répertoire permet de nouvelles analyses au niveau des locus, gènes, transcripts et ORFs. (ii) Nos résultats suggèrent que les sPEPs sont impliqués dans des fonctions régulatrices fondamentales, à la fois ubiquitaires (métabolisme des protéines, ADNs, ARNs, expression génique ...) et spécialisées (réponses immunitaires ...). Nous avons également démontré que la majorité des sPEPs interagissent préférentiellement avec les protéines annotées du même processus que la protéine canonique codée par leur propre transcrit. Enfin, si le modèle agent implémenté ne permet pas d'expliquer les mécanismes de régulation traductionnelle par les uORFs à l'heure actuelle, il fournit à la communauté scientifique un outil facilement adaptable pour approfondir leur étude.

Mots clés : petit cadre ouvert de lecture (sORF), peptides codés par des sORFs (sPEPs), interaction protéine-protéine, traduction

# Acknowledgments

Cette thèse et ses productions sont le fruit de plus de quatre ans de travail au sein de CENTURI, du TAGC et du CIML. Elles n'auraient pu aboutir sans le concours de nombreuses personnes. La thèse est une expérience scientifique et surtout, humaine, si intense que je me dois de conclure ce manuscrit en remerciant les nombreuses personnes qui m'ont tant apportées au cours de ces années. Je prie d'accepter mes excuses aux personnes que je n'ai pas mentionnées et qui m'ont néanmoins offert leur support scientifique / technique / moral au cours de cette expérience, ou avec qui j'ai simplement partagé de bons moments.

Je tiens à remercier en premier lieu Sylvie Ricard-Blum et Yves Vandembrouck d'avoir acceptés de relire ce manuscrit et de m'avoir aidé à sa correction et son amélioration. Je remercie également Serge Plaza d'avoir accédé à ma requête pour présider mon jury de thèse. Merci également à Benoit Ballester d'avoir consenti à évaluer mes travaux à l'occasion de ma soutenance, ainsi que pour votre apport scientifique dans son contenu.

Je souhaite exprimer ma gratitude à mes directeurs de thèse, Christine Brun et Philippe Pierre et de m'avoir accueilli dans leurs équipes et supervisé pendant ces années. Merci de m'avoir donné l'opportunité de développer ces différents projets et de m'avoir fait confiance quant à mes capacités à réaliser cette thèse en soutenant notamment mes candidatures à CENTURI et à la FRM. Merci également pour l'autonomie que vous m'avez laissée durant ces quatre années et pour le partage de la nécessaire prise de conscience de l'équilibre entre vie professionnelle et ma sphère privée.

Je tiens à adresser un grand merci à Lionel Spinelli pour son encadrement scientifique et technique et pour son soutien moral. Merci pour ta rigueur bienveillante, ta disponibilité et ton aide constante pour m'aider à trouver des solutions dans les moments difficiles et pour ta guidance dans cette thèse.

J'espère retenir les enseignements que vous avez tenté de me transmettre, et savoir en faire bon usage dans le futur et vous remercie des moments passés ensemble à l'occasion des meetings que j'ai sollicité, parfois même en dernière minute.

Merci aussi à Andreas Zanzoni d'avoir pris part à l'encadrement scientifique et technique de ma thèse, de m'avoir proposé de rejoindre les nombreux projets liés à mimicINT, émaillés de discussions enrichissantes sur les différents projets, y compris ceux spécifiques à ma thèse.

Je remercie également Pierre Milpied, Benoit Ballester et François Payre, de leur regard extérieur critique, mais bienveillant, sur mes travaux. Merci Pierre et Ben de

vous êtes toujours montrés accessibles et attentifs à mon bien-être. Merci Ben, pour tes compliments encourageants sur mes présentations orales. Ceux-ci ont su motiver "le psychopathe de la diapo" que je suis à chercher à améliorer encore ces dernières !

J'adresse un grand merci aux communautés de CENTURI, du TAGC et du CIML. Je ne pourrai citer toutes les personnes qui m'ont apporté soutien professionnel ou personnel. Néanmoins, je tiens à citer Mélina de Oliveira, Marlène Salom, Simon Legendre, Jasmina Stamenova, Alizée Guarino et Gaël Le Mehaute de m'avoir si bien accueilli, dans cette communauté naissante où j'étais si isolé au début ! Merci pour votre réactivité et votre aide dans les démarches administratives, et surtout pour votre implication en faveur de la communauté étudiante de l'Institut. Je remercie également CENTURI et la FRM et leurs comités de sélection, sans lesquels aucun financement n'aurait été possible. Merci également aux équipes du Mésocentre de m'avoir laissé tester les limites de leur cluster et des capacités de stockage individuelles du data center !

Merci à Audrey Wagner, Kévin Maldonado, Mégane Boujeant, Marceau Cristianini et Lilian Drets pour votre contribution à mes projets scientifiques, l'allègement facilitateur de mon travail, votre enthousiasme et votre curiosité et de m'avoir laissé prendre part à votre supervision. Merci à Audrey de ton intérêt pour la biologie des sORFs et pour ton incroyable travail produit dans le développement de l'interface web. J'espère que vous avez apprécié autant que moi les moments partagés et mes réponses à vos attentes dans vos projets professionnels et parfois personnels. Je vous souhaite le meilleur pour la suite !

J'adresse également ma gratitude à Bertrand Fabre et Serge Plaza pour m'avoir donné l'opportunité de collaborer avec eux dans le cadre du projet visant à caractériser les sPEPs chez la drosophile.

Merci également à Béatrice Nal-Rogier, Eva Strock, Marisa Reverendo et Renaud Vincentelli pour les projets entrepris ensemble et nos tentatives de collaboration, malheureusement interrompues précocement faute de temps, mais je vous sais gré de votre enthousiasme et de votre implication. Peut-être auront-ils une suite ensemble dans l'avenir ?

Une pensée pour Fatiha Tabet, Laurence Conraux, Philippe Legrand, Vincent Rioux, Philip Barter et Gilles Lambert qui ont su stimuler ma curiosité scientifique, m'ont soutenu dans l'idée de faire une thèse et pour les connaissances et compétences transmises si utiles durant cette période.

Je tiens à remercier Julien Gigan et Rémy Char: nous avons traversé les mêmes "phases de vie de doctorants" (parfois avec un peu de décalage). Peu confiants d'avoir tous trois une quatrième année, je suis heureux de votre soutenance quasi en même temps que moi, ayant débutés cette aventure le même jour ! Merci à vous pour les multiples discussions scientifiques et élargies, tentatives de collaborations, conseils, soirées, et pour votre soutien, en particulier durant des périodes difficiles. Je remercie aussi fortement Julie Bavais pour nos multiples échanges ; les repas passés à m'entendre douter et râler, les rires autour d'une bière et de m'avoir apporté tant de

support depuis son arrivée à Marseille. Un grand merci aussi à Marie Dessard pour tes idées de génie (aussi folles que les miennes), les moments passés ensemble et les divers échanges sur nos passions communes. Merci également à Eva Strock et Ania Baaziz pour les partages et le soutien mutuel lors de nos “craquages”. Je tiens enfin à remercier Cloé Zamit, Rosario Lavignolle, Morgane Jaeger, Romain Fenouil, Lou Galliot, Raphaël Chapuy, Elèna Brunet, Thomas Morvan, Eglantine Hector, Alexandre Bonomo, Meriem Djendli, Pauline Brochet, Pauline Andrieux, Marina Cresci ainsi que tous les autres collègues pour ces très bons moments. Je vous souhaite beaucoup de courage et de réussite pour la fin de vos thèses, la suite de vos carrières, et surtout dans vos vies !

Je ne pourrais écrire ces remerciements sans mentionner mes amis les plus proches, bien que parfois géographiquement éloignés. Un très grand merci à vous tous pour m’apporter votre appui, votre bonne humeur, votre joie de vivre et votre écoute si précieux au cours de ces dernières années. Je souhaite tout particulièrement remercier les Faidherbards et les Agros, Véro, Maxou, Lucie, Toinou, PE, Manon, Jéro, Poppy, Myriam, Zoé, Cysouche, MB, Laure, Michou, Teug, Thibault et Eva ; Chouise, Solveig, Pi-Axe, le KGB, Florence, Guillaume, Leslie, Yoze, Laetitia, Alice, Alex, Emilie et Panette pour les occasions festives ensemble : accueilli chez vous ou en visite à Marseille ! Merci également à Alix, Ambre, Chloé, Léa, Laurie, Andrea, Julien, Stéphanie, Marie, Emma, Wendy et Cécile pour tous ces épisodes partagés aux écuries, et plus particulièrement à Pascalou pour avoir su me faire progresser autant, pour son excellent accueil, son énergie débordante et ses soirées cavalières sans égal ! Merci à Emma pour nos échanges et notre soutien mutuel pendant cette période de rédaction. Un très grand merci à Marie pour son soutien, sa disponibilité, ses précieux conseils et pour m’avoir aidé à tant avancer au cours des dernières années. Enfin, une reconnaissance particulière à Tzarine et Pacific qui sont sûrement ceux qui ont eu à me (sup)porter le plus souvent, même au quotidien, dans les meilleurs moments comme dans les pires, et ce, sans jamais chercher à me dégager (littéralement !).

Pour terminer, j’adresse un immense merci à ma famille et plus particulièrement à mes parents, Brigitte et Jacques, sans qui rien n’aurait été possible. Merci pour votre confiance, votre indéfectible présence concernée et affectueuse qui m’a offerte l’opportunité de réaliser ces longues études. Une pensée vers mes grand-parents défunts qui m’ont toujours soutenu.

A toutes et à tous : notre temps partagé, nos moments rares, vos conseils aidants, votre joie de vivre, votre bonne humeur et vos rires ont été mes moteurs !



# Contents

<b>Affidavit</b>	<b>2</b>
<b>List of publications and participation to conferences</b>	<b>3</b>
<b>Abstract</b>	<b>5</b>
<b>Résumé</b>	<b>7</b>
<b>Acknowledgments</b>	<b>9</b>
<b>Contents</b>	<b>12</b>
<b>List of Figures</b>	<b>15</b>
<b>List of Tables</b>	<b>16</b>
<b>List of Acronyms</b>	<b>17</b>
<b>General overview</b>	<b>27</b>
<b>1. Introduction</b>	<b>30</b>
1.1. Short open reading frames (sORFs) are ubiquitous elements expressed in many species . . . . .	30
1.1.1. sORFs constitute a novel class of coding sequences . . . . .	30
1.1.2. sORFs and their products of translation can be identified by complementary methods . . . . .	33
1.1.3. sORFs have been gathered in publicly available repositories . . . . .	41
1.2. Short open reading frames encode functional peptides . . . . .	43
1.2.1. sORF-encoded peptide (sPEP) functions are mainly unknown . . . . .	43
1.2.2. Eukaryotic genomes should no longer be described as monocistronic . . . . .	51
1.3. Short open reading frames regulate the translation of CDSs . . . . .	52
1.3.1. eIF2 $\alpha$ factor is essential to the translation initiation . . . . .	52
1.3.2. The phosphorylation of eIF2 $\alpha$ triggers a translational arrest . . . . .	54
1.3.3. The regulation of the translation by uORFs is related to eIF2 $\alpha$ availability . . . . .	55
1.3.4. uORFs are involved in many processes and diseases as translational <i>cis</i> regulator . . . . .	59

1.4. Short open reading frame variants are conserved across species and involved in the etiology of diseases . . . . .	60
1.4.1. sORFs are conserved across species . . . . .	60
1.4.2. sORF variants have been related to diseases . . . . .	61
1.5. Many questions about short open reading frames and their functions remain unanswered . . . . .	62
<b>2. sORFs identified in human and mouse genomes have been gathered as unique entries in a database: MetamORF</b>	<b>66</b>
2.1. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis . . . . .	66
<b>3. sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins</b>	<b>99</b>
3.1. Studying protein-protein interactions (PPIs) may help characterizing proteins of unknown functions . . . . .	99
3.1.1. Study of sPEP interactions with canonical proteins should provide new insights about their functions . . . . .	99
3.1.2. Proteins functions can be predicted by studying their interactions with annotated proteins . . . . .	100
3.1.3. Short linear motifs and domains mediate PPIs . . . . .	101
3.2. A workflow was required to infer sPEP-protein interactions (sPEPRIs) from protein sequences . . . . .	102
3.3. mimicINT is of major interest to explore the human sPEP-ome . . . . .	127
<b>4. Understanding of the translational regulation by uORFs may be improved by mathematical modeling</b>	<b>160</b>
4.1. Agent-based modeling may help deciphering complex mechanisms . . . . .	160
4.1.1. ICIER, a published TASEP model, partially explains the translational regulation by a single uORF . . . . .	160
4.1.2. Many parameters and uORF features may impact the translation and should be considered in future models of translational regulation by uORFs . . . . .	163
4.1.3. Agent-based models (ABMs) have been used to solve complex questions . . . . .	166
4.1.4. Agent-based modeling allowed to implement a new model of translational regulation by the uORFs . . . . .	168
<b>5. Concluding remarks, limitations and perspectives</b>	<b>173</b>
<b>A. Article: In depth exploration of the alternative proteome of <i>Drosophila melanogaster</i></b>	<b>177</b>
<b>B. Article: Proteostasis in dendritic cells is controlled by the PERK signaling axis independently of ATF4</b>	<b>209</b>



# List of Figures

1.1. The annotation of ORFs is usually based on existing annotations, length, reading frame, transcript biotype and relative position on the transcript	32
1.2. Principle of the ribosome profiling (RiboSeq)	37
1.3. Some demonstrated and putative functions of sPEPs	44
1.4. eIF2 $\alpha$ factor is essential to the translation initiation	53
1.5. ATF4-like mechanism of regulation of the translation by uORFs	56
3.1. Overview of the mimicINT workflow	106
4.1. Principle of the ICIER model	161
4.2. Decision tree of the agent-based modeling of uORF <i>cis</i> regulatory functions informed by experimental data (ABMCisReg)	169

# List of Tables

1.1. Methods of detection and identification of the sORFs and sPEPs . . . . .	33
1.2. Computational tools for identifying sORFs and/or assessing their coding potential. . . . .	34
1.3. Publicly available repositories of sORFs and sPEPs . . . . .	42
1.4. Examples of sPEPs whose functions have been elucidated . . . . .	44

# List of Acronyms

## **3'UTR**

3' untranslated region. [31](#), [53](#), [175](#)

## **43S PIC**

43S pre-initiation complex. [53](#), [54](#), [58](#), [161](#), [164](#)

## **5TOP**

5' terminal oligopyrimidine. [175](#)

## **5'UTR**

5' untranslated region. [31](#), [53](#), [55](#), [57–62](#), [64](#), [160–162](#), [164](#), [168](#), [171](#), [175](#)

## **aa**

amino acid. [31](#), [36](#), [39](#), [43–51](#), [62](#), [66](#), [164](#), [175](#)

## **ABM**

agent-based model. [166–168](#), [170](#)

## **ABMCisReg**

agent-based modeling of uORF *cis* regulatory functions informed by experimental data. [168](#), [170–172](#), [174](#)

## **acORF**

annotated coding open reading frame. [30](#)

## **altORF**

alternative open reading frame. [31](#), [42](#), [60](#)

## **AltProt**

alternative protein. [43](#)

## **ATF4**

Activating transcription factor 4. [55](#), [57](#), [63](#), [162](#), [163](#), [174](#), [209](#)

## **ATF5**

activating transcription factor 5. [55](#)

**BAX**

Bcl-2-associated X. [47](#)

**BCL-2**

apoptosis regulator Bcl-2. [59](#)

**BDI**

Beliefs - Desires - Intentions. [167](#), [168](#)

**bp**

base pair. [47](#)

**cDNA**

complementary DNA. [103](#)

**CDS**

coding sequence. [30](#), [31](#), [33](#), [35](#), [38](#), [50–52](#), [54](#), [55](#), [57–64](#), [68](#), [160–162](#), [164](#), [165](#), [168](#), [174](#), [175](#)

**CHOP**

C/EBP homologous protein, a.k.a. DNA damage inducible transcript 3 (DDIT3). [55](#), [57](#), [58](#), [209](#)

**C-MYC**

Myc proto-oncogene protein. [59](#)

**CReP**

constitutive repressor of eIF2 $\alpha$  phosphorylation a.k.a. protein phosphatase 1 regulatory subunit 15B (PPP1R15B). [54](#), [57](#), [58](#)

**CRISPR-Cas9**

clustered regularly interspaced palindromic repeat - Cas9. [33](#), [99](#), [176](#)

**DC**

dendritic cell. [48](#), [65](#), [209](#)

**DDI**

domain-domain interaction. [101](#), [127](#)

**DMI**

domain-SLiM interaction. [101](#), [105](#), [127](#)

**DNA**

deoxyribonucleic acid. [27](#), [33](#), [34](#), [46](#)

**dORF**

downstream open reading frame. [31](#), [52](#), [177](#)

**eIF**

eukaryotic translation initiation factor. [19](#), [20](#), [57](#)

**eIF1**

eukaryotic translation initiation factor 1. [54](#)

**eIF2**

eukaryotic translation initiation factor 2. [19](#), [53](#), [54](#), [57](#), [58](#), [65](#), [174](#), [209](#)

**eIF2A**

eukaryotic translation initiation factor 2A. [58](#)

**eIF2B**

eukaryotic translation initiation factor 2B. [54](#), [57](#)

**eIF3**

eukaryotic translation initiation factor 3. [53](#), [164](#)

**eIF4F**

eukaryotic translation initiation factor 4F. [53](#), [175](#)

**eIF5**

eukaryotic translation initiation factor 5. [54](#)

**ELM**

eukaryotic linear motif. [101](#)

**EMF**

extreme multifunctional protein. [101](#)

**ER**

endoplasmic reticulum. [54](#), [209](#)

**FLOSS**

fragment length organization similarity score. [38](#)



**GADD34**

growth arrest and DNA-damage inducible 34, a.k.a. protein phosphatase 1 regulatory subunit 15A (PPP1R15A). [54](#), [55](#), [57](#), [58](#), [209](#)

**GCH1**

[guanosine triphosphate \(GTP\) cyclohydrolase 1](#). [62](#)

**GCN2**

general control non-depressible 2, a.k.a. [eukaryotic translation initiation factor 2 alpha kinase 4](#). [54](#)

**GCN4**

general control non-depressible 4. [55](#), [57](#), [163](#)

**GDP**

guanosine diphosphate. [53](#), [54](#)

**GEF**

guanosine exchange factor. [54](#), [57](#)

**GO**

Gene Ontology. [127](#), [128](#)

**GSEA**

gene set enrichment analysis. [209](#)

**GTP**

guanosine triphosphate. [20](#), [53](#), [54](#)

**HMM**

hidden Markov model. [34](#), [35](#), [41](#), [105](#)

**HMT2**

protein arginine N-methyltransferase 1. [61](#)

**HPC**

high-performance computing. [107](#), [108](#), [128](#)

**HRI**

[eukaryotic translation initiation factor 2 alpha kinase 1](#). [54](#)

**HTR3A**

5-hydroxytryptamine receptor 3A. [59](#)

**ICIER**

initiation complexes interference with elongating ribosomes. [160–163](#), [168](#)

**IFRD1**

interferon related developmental regulator 1. [58](#)

**IGFBP3**

insulin-like growth factor binding protein 3. [47](#)

**IRES**

internal ribosome entry site. [53](#), [64](#), [171](#), [175](#)

**ISR**

integrated stress response. [54](#), [55](#), [57](#), [175](#), [209](#)

**kDa**

kilodalton. [39](#)

**lncRNA**

long non-coding RNA. [39](#), [42](#), [43](#), [45–49](#), [51](#), [64](#)

**m<sup>6</sup>A**

N<sup>6</sup>-methyladenine. [53](#), [175](#)

**m<sup>7</sup>G**

m<sup>7</sup>-methylguanosine. [175](#)

**MAPK**

mitogen-activated protein kinase. [59](#)

**MAS**

multi-agents system. [170](#)

**Met-tRNA<sub>i</sub>**

methionyl-initiator tRNA. [53](#)

**MHC**

major histocompatibility complex. [58](#), [65](#)

**MHC-I**

major histocompatibility complex class-I. [27](#), [43](#), [65](#)

**MHC-II**

major histocompatibility complex class-II. [47](#), [65](#)

**miRNA**

micro RNA. [30](#), [48](#), [49](#), [51](#)

**mRNA**

messenger RNA. [27](#), [39](#), [42](#), [46](#), [51–55](#), [58](#), [60](#), [161](#), [163](#), [165](#), [171](#), [174–176](#)

**MS**

mass spectrometry. [33](#), [35](#), [39](#), [40](#), [50](#), [67](#), [103](#), [173](#), [177](#)

**MSH5**

MutS protein homolog 5. [59](#)

**mTOR**

mammalian target of rapamycin. [175](#)

**ncORF**

non-canonical open reading frame. [30](#), [31](#), [33](#), [37](#), [41](#), [51](#), [62](#), [66](#), [67](#)

**ncRNA**

non-coding RNA. [22](#), [30](#), [43](#), [46](#), [47](#), [49](#), [51](#), [65](#)

**ncRNA-ORF**

open reading frame on [non-coding RNAs \(ncRNAs\)](#). [51](#)

**NGS**

next generation sequencing. [40](#)

**NMD**

nonsense-mediated decay. [64](#), [175](#)

**nORF**

novel open reading frame. [30](#)

**nRibo-seq**

nascent [Ribo-seq](#). [39](#)

**nt**

nucleotide. [37](#), [164](#)

**oORF**

overlapping open reading frame. [31](#)

**ORF**

open reading frame. [27](#), [30](#), [31](#), [33–39](#), [41](#), [42](#), [51](#), [52](#), [57–59](#), [61](#), [63](#), [66–69](#), [163–165](#), [168](#), [174–177](#)

**P53**

cellular tumor antigen p53. [59](#)

**PERK**

protein kinase R-like endoplasmic reticulum kinase. [54](#), [209](#)

**PI3K**

phosphatidylinositol 3-kinase. [59](#)

**PKR**

protein kinase R. [54](#)

**PMVK**

phosphomevalonate kinase. [61](#)

**PPI**

protein-protein interaction. [43](#), [100–104](#), [107](#), [127](#), [176](#)

**PPIN**

protein-protein interaction network. [104](#), [107](#)

**PRS**

procedural reasoning system. [167](#)

**Ptch1**

protein patched homolog 1. [59](#)

**PTEN**

phosphatidylinositol 3,4,5-trisphosphate 3-phosphatase and dual-specificity protein phosphatase PTEN. [59](#)

**PTM**

post-translational modification. [46](#)

**RefProt**

reference proteins. [51](#), [57](#), [63](#), [100](#), [104–106](#), [127](#), [128](#)

**Ribo-seq**

ribosome profiling. [22](#), [24](#), [33–40](#), [55](#), [57](#), [65](#), [67](#), [127](#), [160](#), [163](#), [173](#), [176](#)

**RNA**

ribonucleic acid. [27](#), [30](#), [33](#), [37](#), [38](#), [43](#), [47](#), [51–53](#), [63](#), [128](#), [162](#), [165](#), [170](#), [174](#), [175](#)

**RNA-seq**

RNA sequencing. [38](#)

**RPF**

ribosome-protected fragment. [35–38](#), [170](#)

**RPKM**

reads per kilobase per million. [38](#)

**rRNA**

ribosomal RNA. [30](#), [48](#)

**RRS**

ribosome release score. [38](#)

**scRibo-seq**

single-cell [Ribo-seq](#). [38](#)

**SEP**

[sORF](#)-encoded peptide (or sEP). [43](#)

**SERCA**

Sarco/endoplasmic reticulum calcium-ATPase (adenosine triphosphatase) pump. [45](#)

**SHH**

sonic hedgehog protein. [59](#)

**SLIM**

short linear motif. [101–105](#), [107](#), [173](#), [177](#)

**smORF**

small open reading frame. [31](#)

**SNP**

single nucleotide polymorphism. [62](#)

**sORF**

short open reading frame. [24](#), [25](#), [27–29](#), [31](#), [33–41](#), [43](#), [46](#), [47](#), [50–52](#), [59–69](#), [99](#), [128](#), [168](#), [171](#), [173–176](#)

**SOX9**

transcription factor SOX-9. [59](#)

**sPEP**

sORF-encoded peptide. [27](#), [28](#), [33](#), [35](#), [39–41](#), [43](#), [44](#), [46–51](#), [57](#), [59](#), [60](#), [62–65](#), [99–107](#), [127](#), [128](#), [173–177](#)

**sPEPRI**

sPEP-RefProt interaction. [100](#), [103](#), [107](#), [127](#), [128](#), [176](#)

**sPEPRIN**

sPEP-RefProt interaction network. [51](#), [100](#), [127](#)

**SVM**

support-vector machine. [35](#), [41](#)

**TASEP**

totally asymmetric simple exclusion process. [160](#)

**TC**

ternary complex. [53–55](#), [57](#), [161](#), [174](#)

**TE**

translation efficiency. [38](#), [58](#), [59](#), [164](#), [165](#)

**TIS**

translation initiation site. [34](#), [37](#), [39](#), [42](#)

**TI-Seq**

translation initiation sequencing. [34](#), [39](#)

**TISU**

translation initiator of short 5'UTRs. [53](#)

**TPO**

thyroid peroxidase. [59](#)

**TRDD1**

T cell receptor delta diversity 1. [43](#)

**uORF**

upstream open reading frame. [27](#), [28](#), [31](#), [35](#), [36](#), [47–50](#), [52](#), [53](#), [55](#), [57–65](#), [128](#), [160–165](#), [168](#), [171](#), [174–177](#)

**UPR**

unfolded protein response. [54](#)

**UTR**

untranslated region. [61](#)

**VPS53**

vacuolar protein sorting-associated protein 53 homolog. [61](#)

**Y2H**

yeast two-hybrid system. [103](#)

# General overview

## Biological context

Advances in biology over the past decades has revealed the existence of many non-canonical **short open reading frames (sORFs)** on most prokaryotic and eukaryotic **RNAs**. For long, they have been successively missed and ignored, being part of what was designated as "junk **DNA**" at the beginning of the century (even though it would probably be more accurate to talk about "junk **RNA**" in the case of **sORFs**). Indeed, many regions of the genomes were believed to be non-functional at that time, because they were assumed not to encode for any functional protein.

However, earlier publications (based on *low-scale* experiments) already suggested the functionality of this shunned part of the genomes. These have later been supported by the advent of high-throughput technologies and computational methods. The realisation that these regions of the **DNA** (and subsequently **RNA**) could be functional brought a full new level of complexity in our understanding of biological systems. This led to the development of new topics of interest as well as new fields of biology; and we may now probably count at least as many labs interested in these parts of the genomes assumed earlier to be non coding than in the canonical encoding regions.

The non-canonical **sORFs** are characterized by their short size (< 100 codons), the use of alternative start codons and of alternative reading frames. A growing body of evidence suggests they are able to encode functional peptides (called **sORF-encoded peptides (sPEPs)**) that are taking part in a wide range of biological processes, including cell physiology and proliferation, signaling, organogenesis, cell growth and death, transport, enzymatic regulation, metabolism, development, cytoskeleton organization and **major histocompatibility complex class-I (MHC-I)** presentation. In addition, a particular class of **sORFs**, located upstream of the canonical **open reading frames (ORFs)** of **mRNAs** (**upstream open reading frames (uORFs)**) has been described as a novel regulator of the translation. Finally, **sORFs** have been shown to be involved in the etiology of many diseases, including cancers and neurodegenerative diseases for instance. However, the functions of **sORFs** and the peptides they encode, as well as the mechanisms in which they are involved, remain poorly characterized at this time.

## PhD thesis project and objectives

**sORFs** constitute thus a novel repertoire of fascinating biological entities whose roles have certainly been underestimated so far. In particular, their *large-scale* functional



characterization is still in its infancy. Hence, my PhD thesis aimed at exploring **sORFs** functions and to bring new elements of answer to the following question:

*What are the biological functions of **sORFs**?*

As **sORFs** have been proved to have both *trans* and *cis* functions, respectively as peptides and as regulators of the translation, my PhD thesis aimed to address the two following questions:

1. What are the biological functions of the **sPEPs**?
2. Could we elucidate new mechanisms of translational regulation by the **uORFs**?

In order to discuss these questions, I thus decided to:

1. Gather all **sORFs** identified in *H. sapiens* into a repository ([chapter 2](#))
2. Ascertain **sPEPs** functions in monocytes through a system approach, by predicting their interaction with canonical proteins ([chapter 3](#))
3. Explore mechanisms of translational regulation by the **uORFs** in monocytes, by using agent-based modeling to imitate the translation process ([chapter 4](#))

## Organisation of the manuscript

In this manuscript, I report several research projects with the attempt to investigate the role of **short open reading frames**.

The introduction (section 1) focuses on the biology of **short open reading frames** and reviews the current knowledge about these. It notably provides the reader the knowledge necessary to understand the biological context and it states the hypotheses and questions that are being addressed in this thesis. For the sake of clarity, literature related to other topics has been voluntarily omitted from this section.

The chapters 2, 3 and 4 bring additional elements of literature and references necessary to understand the methodology used. They aim at detailing the scientific process chosen and to provide a general overview of the methodology employed and the main findings. Discussions, scattered with personal scientific opinions, are also proposed. The chapter 2 describes an approach to gather all **sORFs** identified in *H. sapiens* and *M. musculus* into a repository, addressing thus the objective 1. The chapter 3 focuses on the exploration of **sPEP** functions in order to explore the question 1 (and its corresponding objective 2). The chapter 4 focuses on the elucidation of translational regulation by the **uORFs** (question 2 and its corresponding objective 3). Studies presented in chapters 3 and 4 are based on the repository presented in chapter 2. Hence,

whilst the chapter 2 has to be read first to understand the source of data used in the other chapters, the chapters 3 and 4 may be read independently.

The appendices present publications related to collaborative projects to which I contributed.

Finally, I voluntarily tried to be as concise as possible and developments about side topics have been reduced to the points necessary to understand this thesis.

## Scientific environment

My PhD thesis took place in the frame of the *Turing Centre for Living Systems* (CENTURI) international PhD program. This last aims at supporting interdisciplinary projects with the willing to decipher the complexity of living systems. My PhD thesis has been led as a collaboration between the *Theories and Approaches of Genomic Complexity* (TAGC) and the *Centre d'Immunologie de Marseille-Luminy* (CIML), in order to combine the strengths of Christine BRUN's (TAGC) and Philippe PIERRE's (CIML) labs respectively in the fields of network biology and dendritic cell biology. My PhD was funded by CENTURI for 38 months and the *Fondation pour la Recherche Médicale* (FRM) for 12 additional months.

In addition to the biological questions that I tried to address during my thesis, my work aimed at bringing new data, methods and tools that may be easily used for other projects and more generally accessible to the scientific community. Mentions of my work related to other scientific projects are briefly made in this manuscript, but details about these have been voluntarily omitted as this is outside the scope of the study of the sORFs.

# 1. Introduction

*Deciphering the genetic information encoded in RNA molecules is one of the biggest challenges in current biology*

Vitorino et al. (2021)

## 1.1. Short open reading frames (sORFs) are ubiquitous elements expressed in many species

### 1.1.1. sORFs constitute a novel class of coding sequences

Genes sequences were initially defined as containing **ORFs**, which are coding sequences eventually translated into functional proteins. A historical arbitrary threshold of 100 codons was early used to define canonical **ORFs** (usually designated as **coding sequences (CDSs)**, **annotated coding open reading frame (acORFs)** or main **ORFs**) under the assumption that **ORFs** of fewer than this size have no coding potential [8, 36, 94, 133, 147]. In this manuscript, **coding sequence (CDS)** will be used to systematically refer to the canonical coding sequence of a **ribonucleic acid (RNA)**. **ORFs** shorter than 100 codons were thus discarded in most gene annotation programs until the beginning of the 2010s, with the notion that they had no coding potential [92, 106, 147, 161]. This threshold was historically set by reasoning on the size distribution of random **ORFs** generated by modelling an equivalent random genome to determine the **ORF** length distribution [133].

Nonetheless, the development of high-throughput technologies and computational methods during the past decades revealed the existence of many **non-canonical open reading frames (ncORFs)** (or **novel open reading frames (nORFs)**) outside of annotated protein-coding **ORFs** on most **RNAs**, including presumptive **ncRNAs**, **microRNAs (miRNAs)** and even **ribosomal RNAs (rRNAs)** [25, 93, 102, 105, 147, 164]. Because of their short size and the use of alternative start codons (*i.e.* other than AUG; CUG, GUG and UUG being the most frequently used alternative start codons [46, 72, 131, 164]) and alternative reading frames [8, 80, 92, 109, 132, 147], **ncORFs** have been missed for a long time, and proteins they encode were commonly discarded from proteomics datasets. Indeed, a minimal length of 100 codons, the use of an AUG as start codon as well as the presence of a single **ORF** per transcript were the criterion commonly

1. Introduction – 1.1. Short open reading frames (sORFs) are ubiquitous elements expressed in many species

used so far to identify the CDS in order to decrease the false positives during the assignment of protein-coding ORFs [94, 113]. However, ncORFs have been detected in most prokaryotes (bacteria...) and eukaryotes (yeast, invertebrates, mammals...) [8, 113] and it has been demonstrated that some of these ubiquitous elements are conserved across species, although they seem to be less conserved than canonical protein-coding ORFs.

Despite a growing effort in identifying, characterizing and classifying these novel ORFs during the past years [36], there is still no clear consensus regarding their definition or the nomenclature to use [97]. However, a common nomenclature tends to emerge and to be used by the vast majority of the scientific community to annotate the non-canonical open reading frames. This nomenclature mainly relies on the annotation of the ORFs according to three main criteria: their length, relative position on the transcript and their frame (Fig. 1.1). Several efforts have been developed in that sense and a correspondence of main authors in the field has been recently published to propose a standardized annotation of translated ORFs [91].

By convention, and mainly for historical reason, ORFs shorter than 100 codons are usually designated as sORFs (or small open reading frames (smORFs)), whilst ORFs located in an alternative frame (*i.e.* other than the canonical one) are usually referred to as alternative open reading frames (altORFs) [8, 36, 80, 117, 139, 147]. One of the most commonly used nomenclature relies on the location of the ORF on its transcript, relatively to the CDS [81]: ORFs having their start codon located in the 5' untranslated region (5'UTR) and their stop codon located upstream of those of the CDS are designated as uORFs (as they are located upstream of the CDS), ORFs located in the 3' untranslated region (3'UTR) are referred to as downstream open reading frames (dORFs) [36, 57, 74, 109] and ORFs overlapping with the CDS are designated as overlapping open reading frames (oORFs) [57, 84]. It should be noticed that while most sORFs contain one single exon, some of them contain introns [81], an important feature that is found in all categories of sORFs and most of the time overlooked in the literature.

It should be stressed out that all classes of ncORFs relatively to their location on the transcript cannot be distinguished one from another other than through information about their location relatively to the CDS. Because this information relies exclusively on existing annotation of canonical ORFs, and no clear distinction can be *a priori* made between ncORFs and the canonical ones based on biological features [48], the biological relevance of making such distinctions between the ORFs may be argued. However, the only manner to eventually identify such distinguishing features, properties or functions between annotated CDSs and newly discovered ORFs is to study the last one as a distinct biological class of coding sequences.

In the scope of my thesis, I decided to focus exclusively on sORFs, which constitute the largest class of ncORFs and is more likely to present specific features than long ncORFs. As a consequence, discussions related to ncORFs longer than 100 amino acids

1. Introduction – 1.1. Short open reading frames (sORFs) are ubiquitous elements expressed in many species

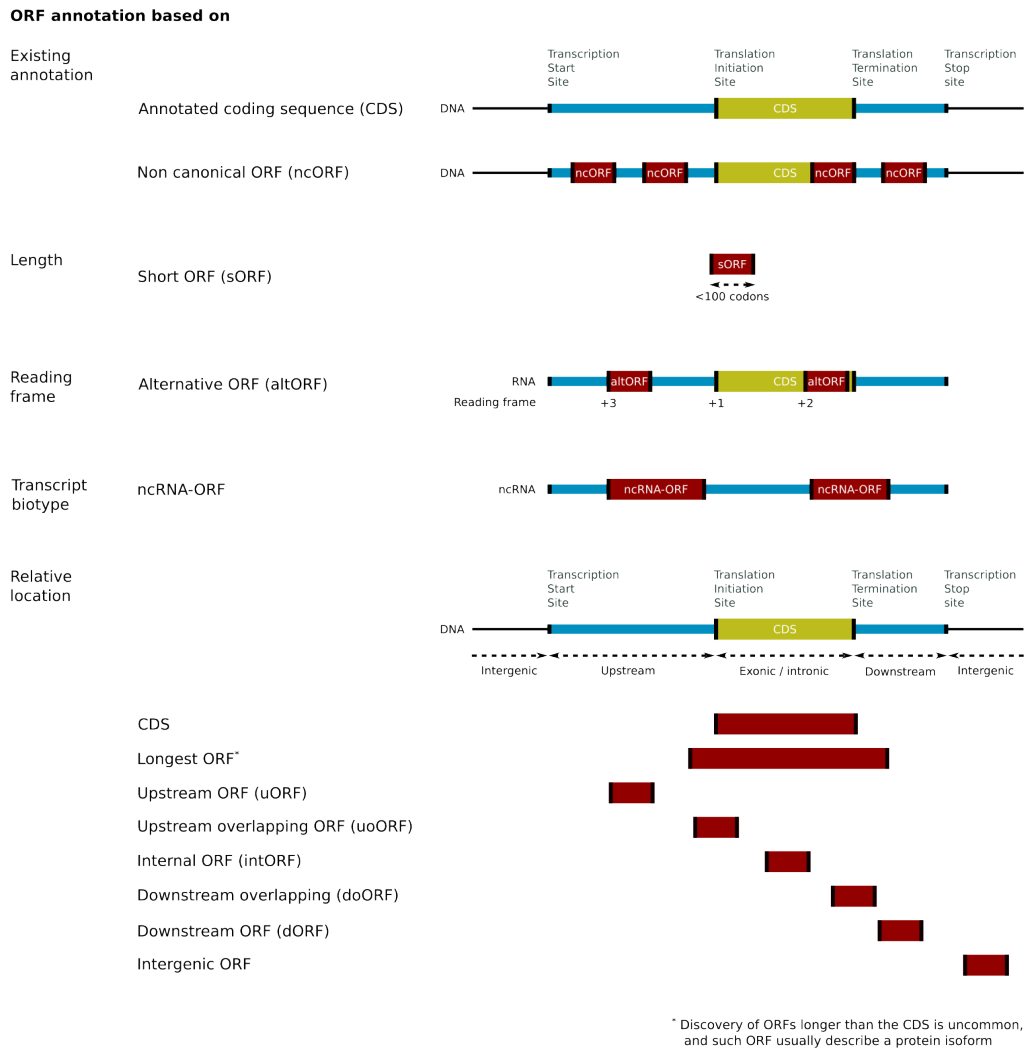


Figure 1.1.: The annotation of ORFs is usually based on existing annotations, length, reading frame, transcript biotype and relative position on the transcript.

1. Introduction – 1.1. Short open reading frames (sORFs) are ubiquitous elements expressed in many species

(aas) and the discovery of protein isoforms will not be developed in this manuscript.

### 1.1.2. sORFs and their products of translation can be identified by complementary methods

Over the past decades, various methods have emerged and been successfully used to identify and study the sORFs, in particular as traditional biochemical techniques failed for identifying them [142]. These methods rely either on the identification of the sORFs themselves, at the genomic (DNA) or transcriptomic level (RNA), or of their products of translation. I review briefly here the methods most commonly used to identify ncORFs. These includes computational approaches, sequencing-based methods, mass-spectrometry based proteomics as well as proteogenomics (Table 1.1).

Table 1.1.: Methods of detection and identification of the sORFs and sPEPs.

Method	Advantages	Drawbacks	Optimal application
Computational approaches	No requirement of experimental data, large-scale and comprehensive prediction of sORFs, conservation analyses	Predictions regarding sORF functionality requires additional data; requires <i>a priori</i> hypothesis for identification (start codon sequences, bias in sORF codon usage etc.)	Large-scale and comprehensive discovery of sORFs
Ribosome profiling (Ribo-seq) and polysomal profiling	Discovery of new ORFs, alternative start codons, translation efficiency and rates, translational pause sites	Requires specialised equipment, labour intensive, expensive, complex bioinformatics analysis, elucidation of sORF functionality requires additional data	Large-scale discovery of sORFs, estimation of translation efficiency
Mass spectrometry (MS)-based proteomics	Direct identification and quantification of sPEPs	Requires specialised equipment, labour intensive, requires novel protocols, numerous sPEPs are missing from the spectral databases, elucidation of sORF functionality requires additional data or non-classical proteomics experiments	Identification of sORFs at the peptide level
Proteogenomics	Direct identification and quantification of sPEPs, allows identifying more sPEPs than proteomics alone as spectra are matched to a database generated <i>in silico</i> from genomic or transcriptomic data	Requires specialised equipment, labour intensive, complex bioinformatics analysis, elucidation of sORF functionality requires additional data	Large-scale discovery of sPEPs, re-processing of existing proteomics datasets
Low-scale biochemical experiments (Western-blot, immunoprecipitations, CRISPR-Cas9 etc.)	Direct identification and functional characterization of sPEP, no complex bioinformatics analysis	Requires <i>a priori</i> knowledge about the sORF or sPEP of interest and available material (e.g. antibodies), labour intensive	In-depth characterization of sPEPs previously identified by other methods

#### 1.1.2.1. Computational approaches allow systematic searches for sORFs

Computational identification of sORFs is challenging because of the difficulty to distinguish sORFs from chance in-frame start and stop codons [113]. Nevertheless, computational approaches that aim at identifying sORFs which are distinct from established well-known CDSs have been successfully developed [8]. These approaches

1. Introduction – 1.1. Short open reading frames (sORFs) are ubiquitous elements expressed in many species

encompass several methods usually based either on sequence analysis or comparative genomics [81, 82, 94, 147]. The table 1.2 reviews the tools most commonly used for identifying sORFs or assessing their coding potential.

Table 1.2.: **Computational tools for identifying sORFs and/or assessing their coding potential.**

Tool	Method	Input	Description	Ref.
<b>sORFinder</b>	Comparative genomics	Sequences	Identifies sORFs with high-coding potential based on the nucleotide composition bias (hexamer composition bias) among coding sequence and the functional constraint at the amino acid level through evaluation of synonymous and non-synonymous substitution rates. Potential coding sORFs are tested for functionality by searching for homologs.	[50]
<b>MiPeptid</b>	Sequence analysis	Sequences	Uses a logistic regression based on hexamer features to predict whether a peptide is encoded by an sORF from its sequence.	[166]
<b>PhyloCSF</b>	Comparative genomics	Sequences	Analyzes a multispecies nucleotide sequence alignment to determine whether an ORF is likely to represent a conserved protein-coding region, based on substitution rates of synonymous to nonsynonymous ratios.	[77]
<b>PhastCons</b>	Comparative genomics	Sequences	Predicts conserved elements in multiple alignment sequences taking into account the probability of nucleotide substitutions at each site in a genome and how this probability changes from one site to the next, using an <a href="#">hidden Markov model (HMM)</a> .	[124]
<b>CRITICA</b>	Comparative genomics and Sequence analysis	Sequences	Coding region identification tool invoking comparative analysis (CRITICA) is a prediction algorithm for identifying protein-coding sequences in DNA. It combines a comparative analysis of homologous sequences based on synonymous variations between sequences with noncomparative methods based on the analysis of codon bias usage.	[12]
<b>Ribo-TISH</b>	Ribo-seq data analysis	Ribo-seq and TI-Seq data	Detects and quantitatively compares <a href="#">translation initiation sites</a> across conditions from <a href="#">translation initiation sequencing (TI-Seq)</a> data and predicts sORFs from <a href="#">ribosome profiling (Ribo-seq)</a> data. It is an unsupervised method that does not rely on prior knowledge of the ORF annotation and allows predicting <i>de novo</i> ORFs.	[164]

1. Introduction – 1.1. Short open reading frames (sORFs) are ubiquitous elements expressed in many species

Tool	Method	Input	Description	Ref.
<b>RiboTaper</b>	Ribo-seq data analysis	Ribo-seq data	Permits <i>de novo</i> prediction of ORFs from Ribo-seq data, on the basis of the characteristic three-nucleotide periodicity of the data. It is built upon the multi-taper unsupervised method developed in the signal-processing field.	[27]
<b>ORF-RATER</b>	Ribo-seq data analysis	Ribo-seq data	ORF regression algorithm for translational evaluation of ribosome-protected fragments (RPFs) (ORF-RATER) predicts ORFs from Ribo-seq data. It uses a linear regression method that assumes that translated ORFs display a pattern of ribosome occupancy that mimics that of annotated CDS.	[46]
<b>RiboHMM</b>	Ribo-seq data analysis	Ribo-seq data	Predicts translated ORFs from Ribo-seq data, using a HMM method.	[107]
<b>RiboORF</b>	Ribo-seq data analysis	Ribo-seq data	Analyzes Ribo-seq data to identify translated ORFs that combines alignment of ribosomal A-sites, trinucleotide periodicity and uniformity across codon, using a support-vector machine (SVM).	[60]
<b>PRICE</b>	Ribo-seq data analysis	Ribo-seq data	Probabilistic inference of codon activities by an expectation – maximization algorithm (PRICE) models experimental noise to accurately resolve overlapping sORFs and non-canonical translation initiation.	[44]
<b>uORF-seqr</b>	Comparative genomics and Ribo-seq	Ribo-seq data	Identifies uORFs based on comparative genomics to study AUG and non-AUG uORFs. The algorithm uses regression to select and weight the features that correlate with uORF detection in biological replicate Ribo-seq datasets.	[131]
<b>uPEPPERoni</b>	Comparative genomics	Sequence	Was an online tool (no longer available) for the identification of putative functional sPEPs, based on the detection of conserved uORFs in eukaryotic transcripts.	[127]
<b>SPECtre</b>	Ribo-seq data analysis	Ribo-seq data.	Uses Ribo-seq data to model the trinucleotide periodicity of ribosomal occupancy using a classifier based on spectral coherence.	[34]
<b>PROTEOFORMER</b>	Proteogenomics	Ribo-seq and MS data	Is a pipeline that allows performing proteogenomics by automatically processing of MS and Ribo-seq data. It builds a reference database by identifying ORFs from Ribo-seq data that can thus be compiled for MS-based identification.	[37, 145]



1. Introduction – 1.1. Short open reading frames (sORFs) are ubiquitous elements expressed in many species

Tool	Method	Input	Description	Ref.
<b>ORFik</b>	Ribo-seq data integration and visualisation	Ribo-seq data	Is a Bioconductor R package supporting standard translation analysis, including read mapping for Ribo-seq, trimming, P-site shifting and ribosomal occupancy quantification. It can notably quantify translation initiation through scanning efficiency and ribosome recruitment. It also incorporates tools for visualisation.	[136]
<b>RiboNT</b>	Ribo-seq data analysis	Ribo-seq data	Is a noise-tolerant ORF predictor that can utilize RPFs with poor periodicity. It uses the RPF periodicity as well as the codon usage to identify translated ORFs. This tools has been proved to be more efficient with RPF-sparsed data due to low-level of translation.	[130]
<b>uORF-Tools</b>	Ribo-seq data analysis	Ribo-seq data	Is a pipeline enabling the identification of differentially translated uORFs from Ribo-seq data, using the Ribo-TISH tool for the identification of uORFs.	[120]

One of the simplest and most common approach consists in detecting all start codons and their downstream stop codon in all three possible reading frames [84]. Because this method is susceptible to generate many ORFs, when several ORFs are sharing a stop codon, this is usually the longest one which is retained (*i.e.* the one with the most upstream start codon) [81, 84]. However, whilst such computational methods relying on the detection of long continuous coding potential has shown to be quite efficient for long ORFs, the inflated false discovery of sORFs requires to integrate additional parameters or to consider other methods of detection, in particular when considering alternative start codons [48, 81, 82]. As an example, McGillivray *et al.* [84] identified nearly 1.3 million uORFs by scanning the human genome for uORFs beginning with an ATG or a single nucleotide variant of ATG and associated with protein coding genes. Because sORFs are not encoded by typical genes containing classical gene structure elements, it was believed that expressed sORFs have a biased use of nucleotides and codons, which has been used for the development of initial computational strategies [8, 36, 106, 147].

A more recent set of computational methods relies on phylogenetic conservation analyses and cross-species comparison to identify conserved sORFs. The rationale for conservation analysis is related to the fact that sORFs that lack cross-species conservation are less likely to be functional or to encode functional peptides<sup>1</sup>. In addition, evolutionary conservation has shown to be a strong indicator of functionality for canonical proteins. However, non-conserved sORFs should not be dismissed as species-specific sORFs may also be biologically relevant and susceptible to evolve more quickly [8, 82, 102, 117]. As highlighted by Makarewich and Olson [82], cross-species comparisons are powerful techniques because most genes are subject to

<sup>1</sup>The term peptide refers to proteins of approximately less than 50 aa

1. Introduction – 1.1. Short open reading frames (sORFs) are ubiquitous elements expressed in many species

evolutionary pressure to maintain sequence conservation and display a prevalence of synonymous codon substitutions versus nonsynonymous substitution.

Finally, many of existing bioinformatics strategies are combining pure computational approaches with the integration of experimental data, as described hereafter.

**1.1.2.2. Ribosome profiling (RiboSeq) allows detection of translated sORFs**

**Ribo-seq** is a method introduced for the first time in 2009 by Ingolia *et al.* [58] and that aims at identifying **ncORFs** with their exact position and to estimate their level of translation [8, 14, 31, 36, 57]. This approach evolved from the polysomal profiling method developed in the 1960s that is based on sucrose-gradient separation of translated mRNAs from untranslated ones [31, 66]. **Ribo-seq** takes a ribosome-centric perspective in order to provide a high-resolution quantitative profile of the translation across the transcriptome [57]. To do so, it aims at stabilizing the interactions between **RNA** molecules and their translating ribosomes (Fig. 1.2), providing then the ability to map the position of ribosomes in all **RNAs** [147].

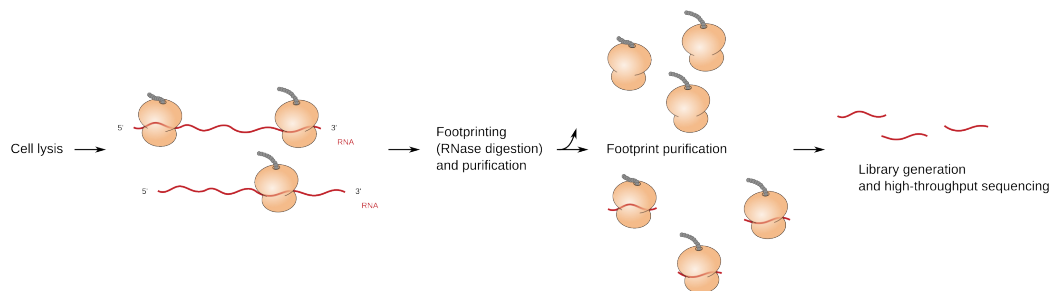


Figure 1.2.: **Principle of the ribosome profiling (RiboSeq).**

The first step consists in stalling the elongating ribosome on the transcript and/or the initiating ribosome at the **translation initiation sites (TISs)**, using respectively cycloheximide and harringtonine, two chemical inhibitors of the translation. The non-protected **RNA** fragments are then digested using a RNase treatment in order to collect **RPFs**. **RPFs** usually size 20-32 **nucleotides (nts)** and are massively sequenced to be mapped on the transcriptome. The density of ribosome footprint is finally exploited to infer the level of translation of each **ORF** [57, 92, 106, 147].

Because **Ribo-seq** allows large-scale systematic detection of translated **ORFs** at codon resolution, it is considered as the current gold standard method for identifying novel functional **sORFs** at a specific time point and in a particular context [31, 147]. In addition, it provides direct evidence for translation, regardless of the size of **ORFs**

## 1. Introduction – 1.1. Short open reading frames (sORFs) are ubiquitous elements expressed in many species

and whether they use canonical or alternative start codons [142]. **Ribo-seq** is usually performed in combination with bulk **RNA sequencing (RNA-seq)** analysis in order to estimate the **translation efficiency**<sup>2</sup> of the various **ORFs** and **CDSs** [31]. Indeed, **Ribo-seq** data provide measurements of protein synthesis that reflect both the translational status of an **ORF** and the underlying abundance of its **RNA**. Hence, normalization to **RNA-seq** data is required to be able to distinguish translational and transcriptional regulations [57]. In addition, one prerequisite for translational studies with **Ribo-seq** is to preserve the translational status of the cell as well as the integrity of its ribosomes, and the protocols need to be carefully designed to avoid any displacement of ribosomes during the preparation of samples [31].

If this recent approach has proved to be particularly efficient and has been extensively used during the past decade, Schott *et al.* [122] point out the fact that analyzing **Ribo-seq** data is complex, as ribosome density is affected both by passive and active changes [14]. Distinct computational strategies have thus been developed to distinguish actual translation events from technical noise. These last are usually based on **RPF** abundance, length and trinucleotide periodicity (since ribosomes move three nucleotides at a time during elongation), positioning within a transcript as well as responsiveness to translation inhibitors [25]. Most of these strategies allocate the translated P-sites (peptidyl sites) or A-sites (aminoacyl sites)<sup>3</sup> according to the positions and offsets of **RPFs** (*i.e.* the number of nucleotides separating the start of the A-site codon from the 5' end of the **RPF**) in order to determine the translated frame for a given sequence [130]. The **translation efficiency** is commonly computed. This metric is defined as the ratio of **RPKM** in **Ribo-seq** over **RPKM** in **RNA-seq** across the **ORF**. It is used to assign a level of translation to an **ORF** [92, 106]. **ORFscore** is an alternative method that has been developed to quantify the level of translation of the **ORF**. It exploits the three codon periodicity of the distribution of **RPKM** relative to the predicted **ORF** [16, 106]. **Fragment length organization similarity score (FLOSS)** is a method that enables to identify the true ribosome footprint bioinformatically based on the magnitude of disagreement between the **RPF**-length distribution of **CDSs** and those of the **sORFs** [94]. To the difference of most computational tools that focus on initiation sites, **ribosome release score (RRS)** is a method that has been designed to detect the termination of translation at stop codons [82].

Various modifications of the initial protocol of **Ribo-seq** as it has been described by Ingolia *et al.* [58] led to the development of slightly different methods. In 2021, VanInsberghe *et al.* notably developed **single-cell Ribo-seq (scRibo-seq)** [141] bringing the

---

<sup>2</sup>The translation (or translational) efficiency (TE) is a common metrics introduced with **ribosome profiling (Ribo-seq)** experiments that aims to quantify the level of expression of an **ORF**. It is defined as the ratio of ribosome footprint abundance over normalized mRNA abundance across the **ORF** (*i.e.*  $TE = \frac{RPKM \text{ in Ribo-seq}}{RPKM \text{ in RNA-seq}}$ ) [31, 88, 92, 106, 109]. It should be noticed that TE score is not a reliable estimator at low expression levels [10].

<sup>3</sup>The A-site of a ribosome is the binding site for the aminoacyl-transfer **RNA**, which assure the translation of the codon into an amino acid

## 1. Introduction – 1.1. Short open reading frames (sORFs) are ubiquitous elements expressed in many species

**Ribo-seq** to a new level that allows the identification of **sORFs** and the measurement of their translation at the single-cell resolution. Another example is the **nascent Ribo-seq (nRibo-seq)**, published by Schott *et al.* in 2021 [122], and that allows monitoring the latency between the appearance of nascent **messenger RNAs (mRNAs)** and their association with ribosomes. Poly-Ribo-seq is another derivative of **Ribo-seq** described by Aspden *et al.* in 2014 [10]. It aims at studying the **sORFs** bound by multiple ribosomes under the assumption that this makes them more likely to be actually translated. However, this method tends to preferentially enrich for actively translated single **sORF-containing mRNAs** because ribosomes have been reported to reach densities of 1 ribosome every 80 nucleotides, which makes the shortest **ORFs** less likely to harbor multiple ribosomes. Hence, it makes it less reliable when it comes to the identification of multiple **sORFs** on the same transcript. Finally, **TI-Seq** has been presented by Zhang *et al.* in 2017 [164] and focus exclusively on the **translation initiation sites**. It aims to discover and quantify the translation initiation events, in particular at alternative **TISs**. Many other slight variations of **Ribo-seq** have been published since the original publication and applied to many species of all kingdoms. Alternative or complementary antibiotic treatments (such as lactimidomycin or puromycin) have also been proposed to enrich specifically for initiation or termination complexes [48, 57].

Despite being the gold standard method for translational studies, **Ribo-seq** suffers from a substantial level of noise [44] and it is arguable that ribosome occupancy on **ORFs** lead to the synthesis of a functional protein. In addition, it should be highlighted that **Ribo-seq** does not provide any information regarding the stability or the eventual functionality of the products of translation [5, 8, 36, 48, 82, 133], and complementary experiments are necessarily required to address these issues.

### 1.1.2.3. Mass spectrometry (MS) based proteomics allows detection of sORF-encoded peptides (sPEPs)

Detection by **MS** remains so far the best method for identifying proteins, but the short length of **sPEPs** raises numerous technical limitations, in both isolating, identifying and detecting the peptides [48, 82, 147]. Indeed, standard protein extraction protocols usually exclude proteins smaller than 10 **kDa** [147], which requires to develop protocols specifically for short size proteins and peptides. Such protocols evolved during the past years and allowed successful direct detection of **sPEPs** [36]. However Tharakan and Sawa [133] report that the detection of peptides as short as 5 **aa** by **MS** is still impossible.

In addition, a huge number of peptides are also generated from canonical proteins by proteosomal degradation, which makes harder the specific recovery and detection of functional **sPEPs** [102]. It has also been highlighted that only a small fraction of **sORFs** produces peptides in sufficient abundance for detection [54] and that there is a rapid turnover of **sPEPs** [46]. While less than 1% of **long non-coding RNAs (lncRNAs)**-encoded peptides are currently evaluated by proteomics [147], 40 % of **lncRNAs** were

1. Introduction – 1.1. Short open reading frames (sORFs) are ubiquitous elements expressed in many species

estimated to contain translated sORFs [164], suggesting that MS-based proteomics methods still struggle at detecting sPEPs.

Hence proteomics is of great interest for identifying sPEPs and studying their functions, in particular as it provides a proof of the existence of stable peptides. However, it should not be the method of predilection when it comes to systematic identification of sORFs, in particular when one is willing to explore the functions of the coding sequence itself, independently of its translational product.

#### 1.1.2.4. Proteogenomics combines advantages of transcriptomics and proteomics approaches to identify sPEPs

The presence of many unidentified ion peaks in mass spectrometry data and the lack of comprehensive databases [8, 132] led to the recent development of proteogenomics approaches, that have become more and more popular during the past few years [82, 133]. As described by Laumon *et al.* [72], proteogenomics "leverages on next generation sequencing (NGS) to perform genomically informed proteomics". In other words, proteogenomics combines proteomics with genomics or transcriptomics [144]. It consists in searching the mass spectral data against a database generated *in silico* and that contains the conceptual translation of all six reading frames (three forwards and three reverse) of the genome or transcriptome assembly, or generated from the translome identified by Ribo-seq experiments or any of the computational methods described above [8, 106, 113, 144].

Using proteogenomics allowed thus to recover many sPEPs from existing MS data, that were missed because most databases are missing information about sPEPs. However, the central problem with these spectral databases generated based on transcriptomics or genomics data is their extremely large size, which tends to decrease the search sensitivity [133]. Proteogenomics approaches thus struggle with the distinction between true and false peptide-to-spectrum matches as the database sizes enlarge. Using Ribo-seq data allows narrowing down the search space to the translome and to partially solve this issue [144].

Hence, while ribosome profiling (Ribo-seq) allows detecting many more sORFs by far, proteomics provides a direct evidence of the accumulation of sPEPs at a meaningful level [48], providing thus an additional insight about their functionality. Using different discovery methods and combining data from experiments run in various cellular contexts is expected to provide distinct but often complementary information and should be encouraged [48, 57]. Gray *et al.* [48] even consider that two lines of evidences are necessary for annotation of sPEPs: either independent identification by conservation analysis and ribosome profiling, or by ribosome profiling and proteomics. Based on this principle, many methods have been developed by combining experimental and computational approaches to identify new sORFs and estimate

1. Introduction – 1.1. Short open reading frames (sORFs) are ubiquitous elements expressed in many species

their level of translation. Some of them take notably advantage of Bayes' classifiers, HMM, SVMs, random forest-based and logistic regression-based classifiers, as well as experimental data, to identify sORFs susceptible to be translated but for which experimental evidences at the peptide level are still missing [8, 81, 84, 106] (Table 1.2).

Because of their short size, it has been often argued that sORFs can quite easily appear randomly in the genomes, arising questions regarding the relevance to study such ORFs. Altogether, these methods have greatly improved the identification of sORFs [48] and our knowledge regarding these new biological entities and should continue to provide proofs of their functionality in the future.

### 1.1.3. sORFs have been gathered in publicly available repositories

Given the growing body of evidence regarding the functional importance of the ncORFs and their sPEPs, there is an urgent need in gathering information about these novel ORFs and making it accessible to the scientific community. Efforts have already been made in this direction and the table 1.3 reviews publicly available resources that gather information about sORFs and/or sPEPs.

It should be noticed that sORFs.org [94, 95] and OpenProt [23, 24] are currently considered as the two main and most comprehensive repositories of sORFs experimentally identified. In 2018, publicly available data were scattered across different databases, datasets were aligned on different genome builds as well as differently annotated and formatted and no uniform nomenclature was used to describe the sORFs. This called for an uniformed, easily-accessible resource where each sORF would be individually described, a first task I tried to address during my thesis. However, as stressed out by Neville et al. [93], there is by no means a comprehensive catalog of sORFs as more of them are sure to be discovered in the future, and it is likely that several databases will remain necessary as they are build on different, but complementary, paradigms.

# 1. Introduction – 1.1. Short open reading frames (sORFs) are ubiquitous elements expressed in many species

Table 1.3.: **Publicly available repositories of sORFs and sPEPs** (adapted from Choteau et al. (2021) [33]).

Database	Database type	Type of data	Species	Redundant entries?	Total number of entries	Original data vs. reprocessed data	Gene information	Reference
sORFs.org	Re-processing	Ribosome profiling	<i>H. sapiens</i> , <i>M. musculus</i> , <i>D. melanogaster</i> , <i>R. norvegicus</i> , <i>C. elegans</i> , <i>D. rerio</i>	Yes	4,377,423 ORF to transcript associations	Reprocessed	Not provided	[94, 95]
OpenProt	Re-processing	Computational predictions, Ribosome profiling, Mass spectrometry, Proteogenomics	<i>H. sapiens</i> , <i>M. musculus</i> , <i>D. melanogaster</i> , <i>R. norvegicus</i> , <i>P. troglodytes</i> , <i>C. elegans</i> , <i>D. rerio</i> , <i>B. taurus</i> , <i>S. cerevisiae</i> , <i>O. aries</i>	No	2,677,505 ORFs	Reprocessed	Provided	[23, 24]
RPFdb	Re-processing	Ribosome profiling	33 species, including <i>H. sapiens</i> , <i>M. musculus</i> , <i>S. cerevisiae</i> , <i>E. coli</i> , <i>C. elegans</i> , <i>D. rerio</i> , <i>A. thaliana</i> , <i>R. norvegicus</i> , <i>S. pombe</i>	Yes	Unknown	Reprocessed	Provided	[152]
smPROT	Re-processing	Ribosome profiling, Mass spectrometry, Literature mining	<i>H. sapiens</i> , <i>M. musculus</i> , <i>S. cerevisiae</i> , <i>E. coli</i> , <i>C. elegans</i> , <i>D. rerio</i> , <i>R. norvegicus</i>	Yes	255,010 ORFs	Reprocessed	Provided	[61]
TISdb	Original data processing	Ribosome profiling, (global translation initiation sequencing)	<i>H. sapiens</i> , <i>M. musculus</i>	No	16,964 TISs	Original	Provided	[151]
RiboSeqDB	Tool for data analysis	Ribosome profiling	<i>H. sapiens</i> , <i>M. musculus</i>	Yes	Unknown	Reprocessed	Provided	[123]
PTTDB	Original data processing	Proteogenomics	<i>H. sapiens</i> , <i>M. musculus</i> , <i>P. alecto</i> , <i>A. aegypti</i>	Yes	117,578 ORFs	Original	Provided	[114]
uORFdb	Text mining	Manual curation of the literature	All species	Yes	1,023 ORFs	Manual curation	Provided	[155]
TranslatomeDB	Re-processing	Ribosome profiling	23 species, including <i>H. sapiens</i> , <i>M. musculus</i> , <i>E. coli</i> , <i>C. elegans</i> , <i>D. rerio</i> , <i>A. thaliana</i> , <i>R. norvegicus</i>	Yes	Unknown	Reprocessed	Not provided	[78]
nORFs.org	Post-processing	Ribosome profiling, Mass spectrometry, Proteogenomics	<i>H. sapiens</i>	No	194,407 ORFs	Reprocessed	Provided	[93]
HalfORF	Computational prediction	Computational predictions	<i>H. sapiens</i>	No	17,096 aORFs, 31,422 miRNAs	-	Unknown	[140]
GWIPS-viz	Genome browser for Ribosome profiling, Seq data	Ribosome profiling	29 species, including <i>H. sapiens</i> , <i>M. musculus</i> , <i>R. norvegicus</i> , <i>D. rerio</i> , <i>D. melanogaster</i> , <i>C. elegans</i> , <i>S. pombe</i> , <i>S. cerevisiae</i> , <i>E. coli</i> , <i>A. thaliana</i>	NA	Unknown	Reprocessed	Yes	[86]
TransLuc	Post-processing	Computational predictions, Ribosome profiling, Mass spectrometry	<i>H. sapiens</i> , <i>M. musculus</i> , <i>R. norvegicus</i>	No	583,840 ORFs, 33,094 lncRNAs	Reprocessed	No	[79]



## 1.2. Short open reading frames encode functional peptides

### 1.2.1. sORF-encoded peptide (sPEP) functions are mainly unknown

A growing body of evidence demonstrated that sORFs may encode functional peptides that have been overlooked for a long time because of their short size [36, 102, 133, 142, 147]. They are designated as sORF-encoded peptides (sPEPs) (SEPs), alternative proteins (AltProts), micropeptides, microproteins or miniproteins [5, 52, 54, 74, 81, 82, 94, 97, 102, 106, 109, 147, 160]. The first description of such peptides in eukaryotes has been made during the 1990s and the first sPEPs were mainly identified by serendipity through investigation of ncRNAs [102]. The recent discovery that lncRNAs are able to encode functional peptides has drawn an increasing attention for the past few years [79] and emphasizes the functional potential of this unexplored class of peptides [113]. To the difference of peptides generated by cleavage of a long precursor (such as hormones or neurotransmitters), sPEPs are, by definition, *de novo* products of translation [52, 80, 82, 113]. It is not yet clear if there will be any lower size limit, but one can argue that there is no reason to consider a lower limit [8, 48], in particular regarding the fact that peptides as short as two amino acids are known to be functional (e.g. 2 aa TRDD1 or the 4 aa phagocytosis-stimulating peptide).

Numerous studies support the importance of sPEPs in many functions, notably in fundamental cellular and physiological processes and even their involvement in the etiology of some diseases [79]. It has been suggested that they constitute a new pool of cancer-related peptides that could become potential therapeutic targets in the future [36, 109, 113, 133] or be used as biomarkers for diagnosis [109, 147]. Functional sPEPs have already been discovered in many species, in both prokaryotes (bacteria) and eukaryotes (yeast, invertebrates, mammals, plants...) [36, 82, 113]. sPEPs have notably been demonstrated to be involved in cell proliferation, signaling, cell physiology, organogenesis, growth, cell death, transport, enzymatic regulation, metabolism and development, cytoskeleton organization, major histocompatibility complex class-I (MHC-I) presentation and control of RNA polymerase [25, 36, 44, 52, 72, 109, 133, 147], but also to be involved in protein-protein interactions, ribosomal complexes or be receptors' ligands and to have antibacterial properties [10, 25, 36, 97, 147, 161].

Due to their short size, sPEPs have also been hypothesized to easily fit into binding pockets of other proteins, which makes them candidate regulators of protein-protein interactions and enzymatic activity (as it has been demonstrated for Pri-peptides in drosophila for instance, see Table 1.4) and susceptible to take part in larger protein assemblies [25]. In addition, Hazarika et al. [52] point out the fact that protein-peptide interactions involve smaller interfaces than canonical protein-protein interactions (PPIs), which make these interactions usually of weaker affinity and transient. This can thus bring to fast changes in interactions when they are broken under sudden



## 1. Introduction – 1.2. Short open reading frames encode functional peptides

cellular perturbations, and lead *in fine* to faster cellular responses. Because they often present hydrophobic features, it has also been suggested that sPEPs are likely to interact with membranes [97, 102]. In addition canonical peptides are known to be toxins, transmembrane peptides, enzymatic interactors, modulators of enzymatic activity, transporter regulators, regulators of the transcription, secreted peptides, hormones or receptors [10, 25, 48], so we may reasonably hypothesized that some of these functions are likely to be fulfilled by sPEPs. Aspden *et al.* [10] also emphasize that sPEPs are more susceptible to interact with canonical proteins as regulators, as their size limits their structural capabilities. The Figure 1.3 presents a graphical summary of sPEPs putative and demonstrated molecular functions and the Table 1.4 presents some sPEPs which functions have been elucidated, mainly based on low-scale experiments.

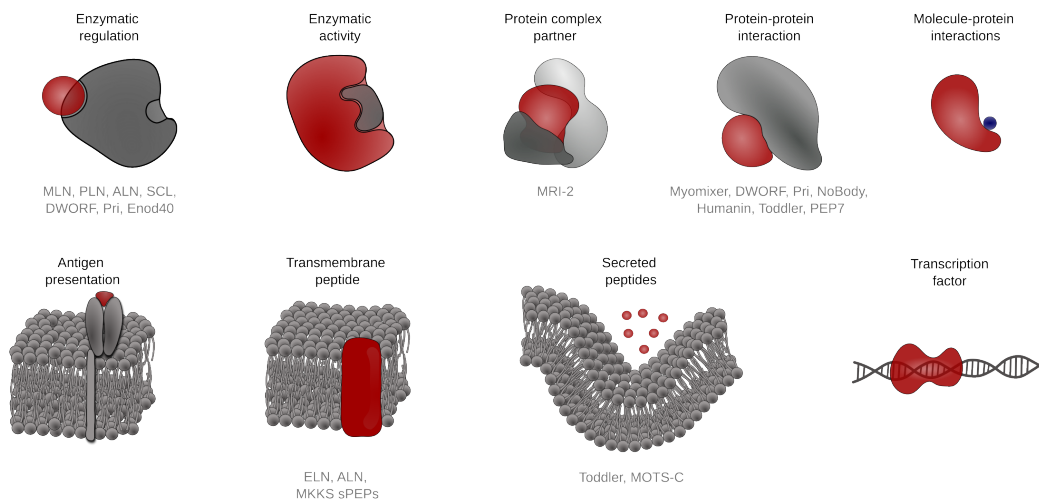


Figure 1.3.: **Some demonstrated and putative functions of sPEPs.** Examples of characterized sPEPs are presented in the Table 1.4.

Table 1.4.: **Examples of sPEPs whose functions have been elucidated.** NB: Contrary to canonical proteins, most sPEPs do not have a symbol, alias, name or unique identifier.

sPEP	Size (aa)	Organism(s)	Functions	Ref.
Myomixer	84	<i>M. musculus</i>	a.k.a. microprotein inducer of fusion (minion) or Myomerger. Myomixer has a muscle-related function and is unique to skeletal muscle. It links to Myomaker, a fusogenic membrane protein that controls the cell fusion and muscle formation.	[102, 147]

## 1. Introduction – 1.2. Short open reading frames encode functional peptides

sPEP	Size (aa)	Organism(s)	Functions	Ref.
Myoregulin (MLN)	46	Mammals	Encoded by an <a href="#">lncRNA</a> . MLN is an inhibitor of the <a href="#">Sarco/endoplasmic reticulum calcium-ATPase (adenosine triphosphatase) pump (SERCA)</a> pump in muscle. The <a href="#">SERCA</a> pump is encoded by three genes in vertebrate and belongs to a family of P-type ATPases. It ensures the transport of calcium, notably in various skeletal muscles. MLN presents structural similarities with PLN and SLN. <a href="#">SERCA</a> (Sarco/endoplasmic reticulum calcium-ATPase (adenosine triphosphatase) pump) ensure the uptake of calcium in the sarcoplasmic reticulum. The calcium is involved in many cellular processes, such as cell motility, fertilization, platelet cell activation, cardiac hypertrophy, vascular tone, neuronal transmission, synaptic plasticity and muscle contraction. The calcium release from sarco/endoplasmic reticulum (s/ER) is performed by various distinct channels and a passive leak whilst the reuptake of calcium into the S/ER is exclusively performed through <a href="#">SERCA</a> pump. In vertebrates, <a href="#">SERCA</a> is encoded by three genes (SERCA1 to SERCA3) transcribed into multiple splice isoforms expressed in distinct cell types [4, 147]	[4, 25, 81, 102, 113, 133, 147, 159, 161]
Phospholamban 52 (PLN)	52	Mammals	PLN is an inhibitor of the <a href="#">SERCA</a> pump. It is specifically expressed in the atria and ventricles of the heart and bladder in mouse embryos. It competitively binds to <a href="#">SERCA</a> with SLN, ALN and ELN.	[4, 25, 102, 161]
Sarcolipin (SLN)	31	Mammals	SLN is an inhibitor of the <a href="#">SERCA</a> pump. It is specifically expressed in the atria of the heart and embryonic slow-type skeletal muscles in mouse. It competitively binds to <a href="#">SERCA</a> with PLN, ALN and ELN.	[4, 25, 102]
Endoregulin (ELN)	65	Mammals	a.k.a. 1110017F19Rik/SMIM6. ELN is a transmembrane peptide conserved in mammals and binding to the <a href="#">SERCA</a> pump. It competitively binds to <a href="#">SERCA</a> with PLN and SLN.	[4, 133]
Another-regulin (ALN)	58	Mammals	a.k.a. 1810037I17Rik. ALN is a transmembrane peptide conserved in mammals and binding to the <a href="#">SERCA</a> pump. It competitively binds to <a href="#">SERCA</a> with PLN and SLN.	[4, 133]
Sarcolamban (SCL)	28	<i>D. melanogaster</i>	SCL is an inhibitor of the <a href="#">SERCA</a> pump in invertebrates, found in the cardiac and somatic muscle.	[4, 81, 106, 133]
Dwarf ORF (DWORF)	34	<i>D. melanogaster</i>	DWORF is encoded by a <a href="#">lncRNA</a> . It binds <a href="#">SERCA</a> in muscle and increases its activity by displacing inhibitory proteins (MLN, SLN and PLN). It has been found to be suppressed in human heart with ischemia.	[4, 102, 147]

## 1. Introduction – 1.2. Short open reading frames encode functional peptides

sPEP	Size (aa)	Organism(s)	Functions	Ref.
Tarsal-less (tal) / polished rice (pri) sORFs	11 & 32	<i>D. melanogaster</i>	The <i>pri</i> transcript is a <a href="#">lncRNA</a> that encodes four functionally redundant peptides (3 sPEPs of 11 aa, 1 sPEP of 32 aa) involved in fly embryonic development. These peptides are among the best-characterized sPEPs. Pri (a.k.a. tarsalless) is required for the development of adult legs. Its beetle homolog millepatetes (Mlpt) is also involved in the embryonic segmentation and leg specification. In <i>Drosophila</i> , epidermal cells normally differentiate extensions called trichomes. During the terminal differentiation of embryonic epidermal cells, Pri triggers the remodeling of apical cell shapes. It has been shown to mediate the <a href="#">post-translational modification (PTM)</a> of the transcription factor Shavenbaby (Svb). Svb is a large protein (1,351 aa) expressed in epidermal cells that acts as a repressor under its complete form. It orchestrates the expression of various cellular effectors responsible for the reorganization of the cytoskeleton, extracellular matrix and membrane domains, resulting in the repression of trichome formation. Pri induces N-terminal truncation of Svb by interacting with the E3 ubiquitin ligase Ubr3 and Svb, resulting in proteasomal degradation. This results into the removal of the repressor region and thus into a shorter activator (906 aa) triggering trichome formation. In addition, Pri is controlled by periodic pulse of steroid hormones and thus related to a systemic temporal control of developmental transitions.	[8, 80, 81, 102, 106, 113, 147, 160, 161]
NoBody	68	Mammals	NoBody is a well-conserved sPEP encoded by LINC01420/LOC550643 in mammals. It interacts with EDC4 that facilitates <a href="#">mRNA</a> decapping and promotes their decay. NoBody has been found to co-localize with P-body and other <a href="#">mRNA</a> decapping/decay factors.	[102, 133, 147]
Toddler	58	Zebrafish	a.k.a. ELABELA or Apela (ELA). Toddler is encoded by a <a href="#">ncRNA</a> , evolutionary conserved and expressed in the extracellular compartment. It is a ligand for the Apelin receptor that functions as a motogen ( <i>i.e.</i> stimulates the cell motility) by promoting the cell migration and gastrulation motion.	[81, 102, 106, 133, 147, 161]
Polar granule component (Pgc) sPEP	71	<i>D. melanogaster</i>	Pgc is encoded by a <a href="#">lncRNA</a> specific to germ cells and required for their development.	[8, 81, 102, 161]
MRI-2	69	Mammals	MRI-2 is involved in <a href="#">DNA</a> repair process. It activates the nonhomologous end joining (NHEJ) factor by interacting with Ku, a protein that binds to <a href="#">DNA</a> ends.	[102, 106, 113, 147, 160]

## 1. Introduction – 1.2. Short open reading frames encode functional peptides

sPEP	Size (aa)	Organism(s)	Functions	Ref.
PEP7	7	Mammals	PEP7 is encoded by an <a href="#">uORF</a> of the angiotensin type 1a receptor (AT1aR) gene. Binding of angiotensin II to the angiotensin receptor activates both, G protein-coupled and non-G protein-coupled pathways to control fluid and electrolyte homeostasis. PEP7 selectively impedes the non-G protein-coupled signaling, without affecting the classical G protein-coupled signaling pathway.	[109]
Alternative-MiD51	70	Mammals	Alternative-MiD51 is encoded by an <a href="#">uORF</a> of the mitochondrial elongation factor 1 (MIEF1), a mitochondrial receptor of Drp1. Alternative-MiD51 is a mitochondria fission factor and upstream regulator of MiD51 translation.	[109]
P155	17	Mammals	P155 is encoded by <a href="#">ncRNA</a> MIR155HG. It modulates <a href="#">major histocompatibility complex class-II (MHC-II)</a> -mediated antigen presentation and T cell priming and acts as a suppressor of inflammatory diseases.	[147]
Humanin	24	<i>H. sapiens</i>	Humanin is encoded by a 75 <a href="#">bp sORF</a> localized on mitochondrial 16S <a href="#">RNA</a> . Humanin interacts with a tripartite cytokine receptor and presents anti-apoptotic functions by inhibiting <a href="#">Bcl-2-associated X (BAX)</a> protein and <a href="#">insulin-like growth factor binding protein 3 (IGFBP3)</a> .	[80, 102, 113, 133]
HOXB-AS3 <a href="#">sPEP</a>	53	<i>H. sapiens</i>	<a href="#">lncRNA</a> HOXB-AS3-encoded <a href="#">sPEP</a> (unnamed). This <a href="#">sPEP</a> suppresses tumorigenesis by PKM alternative splicing regulation and colon cancer cell metabolic reprogramming.	[147]
AGD3 <a href="#">sPEP</a>	63	<i>H. sapiens</i>	AGD3 (a.k.a. TUF)-encoded <a href="#">sPEP</a> (unnamed). This <a href="#">sPEP</a> is active in human stem cell differentiation.	[147]
Yin Yang 1 (YY1)-binding micropeptide (YY1BM)	21 <a href="#">aa</a>	<i>H. sapiens</i>	YY1BM is encoded by <a href="#">ncRNA</a> LINC00278. It is involved in esophageal squamous cell carcinoma progression by inhibiting the interaction between YY1 and the androgen receptor, making it more adaptive to nutrient deprivation.	[147]
Micropeptide inhibiting actin cytoskeleton (MIAC)	51	<i>H. sapiens</i>	MIAC is a <a href="#">sPEP</a> inhibiting actin cytoskeleton. It is an key player in cancer progression and low MIAC expression has been correlated with poor overall survival of head and neck squamous cell carcinoma patients. It has also been significantly associated with the progression of five other tumors.	[147]
CIP2A-BP	52	<i>H. sapiens</i>	CIP2A-BP is encoded by <a href="#">lncRNA</a> LINC00665. It is a prognostic marker of breast cancer and has been suggested to constitute a novel therapeutic target.	[147]
NR3C1 <a href="#">sPEP</a>	93	<i>H. sapiens</i>	NR3C1 <a href="#">sPEP</a> is localized at the cell membrane and regulates expression of the glucocorticoid receptor through interaction with unknown cellular factors.	[8]

## 1. Introduction – 1.2. Short open reading frames encode functional peptides

sPEP	Size (aa)	Organism(s)	Functions	Ref.
ASS1 sPEP	44	<i>H. sapiens</i>	This sPEP (unnamed) inhibits the expression of ASS1.	[8]
EPHX1 sPEPs	17 & 26	<i>H. sapiens</i>	Two sPEPs (unnamed) are encoded by EPHX1 uORFs. They inhibit the translation of EPHX1 through interactions with the translational machinery.	[8]
MKKS sPEPs	50	<i>H. sapiens</i>	MKKS sPEPs (unnamed) localizes to the mitochondrial membrane.	[109]
AltPrP sPEP	73	<i>H. sapiens</i>	AltPrP sPEP (unnamed) is co-expressed from the prion protein transcript in brain, primary neurons and peripheral blood mononuclear cells. It has been localized to the mitochondria.	[8]
AltATXN1 sPEP	185	<i>H. sapiens</i>	AltATXN1 sPEP (unnamed) is expressed in the cerebellum and interacts with the ATXN1 protein in the nucleus.	[8]
AltMRVII sPEP	134	<i>H. sapiens</i>	AltMRVII sPEP (unnamed) colocalizes to the nucleus and interacts with BRCA1.	[8]
Mitochondrial open reading frame of the 12S rRNA-c (MOTS-C)	16	<i>H. sapiens</i>	MOTS-C is encoded by the 12s rRNA gene in the mitochondrial genome. It is conserved between 14 mammalian species and presents an antidiabetic activity. It regulates the cellular metabolism through changes in the methionine-folate cycle and an increase in AMPK activity. MOTS-C acts as a mitochondrially derived hormone and play a systemic role in the metabolic homeostasis of skeletal muscles and fat tissues.	[102, 113, 133]
C17ORF91	57	<i>H. sapiens</i>	C17ORF91 is encoded by lncRNA MIR22HG (pri-miRNA-22) induced in response viral infection. Its functions are still unknown.	[105]
miPEP200a and miPEP200b	187 & 54	<i>H. sapiens</i>	miPEP200a and miPEP200b are respectively encoded by the pri-miRNAs of miR-200a and miR-200b. They have been shown to inhibit prostate cancer cells migration by downregulating vimentin expression, however the mechanism needs to be elucidated.	[105]
miPEP155	17	<i>H. sapiens</i>	miPEP155 is encoded by lncRNA MIR155HG (pri-miRNA-155), an important regulator of hematopoiesis, inflammation, immunity and tumorigenesis. miPEP155 suppresses autoimmune inflammation by regulating antigen transportation and presentation by antigen-presenting cells (notably in dendritic cells). miPEP155 is conserved in primates.	[105]

## 1. Introduction – 1.2. Short open reading frames encode functional peptides

sPEP	Size (aa)	Organism(s)	Functions	Ref.
miPEP133	133	<i>H. sapiens</i>	miPEP133 is encoded by MIR34ahg (pri-miRNA 34a). It is expressed in various normal tissues and down-regulated in cancer cell lines and tumors and has been reported to be a tumor suppressor when over-expressed. It localizes to mitochondria and enhances p53 transcriptional activity.	[105]
Mm47	47	<i>M. musculus</i>	Mm7 is encoded by ncRNA 1810058I24Rik and localized in the mitochondrion and seems to be involved in the responses of NLRP3 inflammasomes.	[147]
Tal1 sPEP	NA	<i>M. musculus</i>	Tal1 uORF translation results in the synthesis of truncated TAL1 isoforms that favor erythroid lineage choice.	[109]
SCL	28 & 29	<i>D. melanogaster</i>	ncRNA pncr003:2L encodes two sPEPs known as SCL and that control the calcium transport. They are also known to regulate heart muscle contraction in drosophila.	[8, 113, 147, 161]
Small peptide of amino acid response (SPAR)	90	<i>D. melanogaster</i> , <i>H. sapiens</i>	SPAR is encoded by lncRNA LINC00961 and localized in late endosomes and lysosomes. It controls the activation of mTORC1 and facilitates muscle regeneration.	[102, 133, 147]
Hemotin	88	<i>D. melanogaster</i>	Hemotin is expressed in macrophages. It regulates endosomal maturation during phagocytosis. Weak sequence conservation was found in vertebrates for this peptide, but tertiary structures similarities were found with Stannin, a 88 aa peptide which is involved in organometallic toxicity.	[102]
Early nodulin 40 gene (Enod40) sPEPs	12 & 24	Plants	Enod40 encodes two sPEPs that interact with enzyme synthesizing sucrose during the organogenesis of root nodules.	[147, 161]
DLV1	51	Plants	DLV1 is involved in the regulation of organogenesis.	[147]
Zm908p11	97	Plants	Zm908p11 facilitates the pollen development.	[147]
Zm401p10	89	Plants	Zm401p10 facilitates the pollen development.	[147]
AtCDC26	65	Plants	AtCDC26 regulates the accumulation complex/cyclosome (APC/C) target proteins during the plant anaphase. It is involved in the control of cell division, growth and embryo development.	[97]
POLARIS sPEPs	36	<i>A. thaliana</i>	POLARIS encodes a sPEP (unnamed) that affects the vascular patterning of leaves and root growth.	[147, 161]
PLS	36	<i>A. thaliana</i>	PLS modulates root growth and leaf vascular patterning.	[8]
ROT4	53	<i>A. thaliana</i> , maize	ROT4 controls polar cell proliferation in lateral organs and leaf morphogenesis.	[8, 147]
ROT18	25	<i>A. thaliana</i>	ROT18 regulates the programmed cell death.	[147]

## 1. Introduction – 1.2. Short open reading frames encode functional peptides

sPEP	Size (aa)	Organism(s)	Functions	Ref.
Brk	76	<i>A. thaliana</i> , maize	Brk has been shown to play a role in leaf morphogenesis.	[81, 147]
SAMDC sPEP	52	<i>A. thaliana</i>	The expression of SAMDC CDS is regulated by polyamines binding to the nascent uORF-encoded peptide (unnamed).	[8]
CPA1	25	<i>S. cerevisiae</i>	CPA1 reduces the expression of the CDS through ribosomal stalling.	[8]
SgrT	43	Bacteria	SgrT is involved in glucose metabolism through interaction with glucose transporter PtsG.	[113]
spoVM	26	Bacteria	spoVM is an essential for endospore formation.	[161]

Despite the characterization of some sPEPs and an intensification of the studies about these peptides during the past years [48], very few have been identified by mass spectrometry [132] and their functional annotation remains a major challenge, as the functions of most of them is still unknown. The fact that an entire class of peptides has been missed so far implies that one has been missing a whole level of regulation, critical structural components, as well as unique mechanisms of action [48]. In addition, because the relative abundance of sPEPs changes during stress [48, 88], it is legitimate to wonder if they play a role in the homeostasis of the cell. The functional characterization of novel sPEPs is thus challenging, as they are likely to significantly increase the proteome<sup>4</sup> of many species [10] (and sometimes designated as being part of the "dark proteome" [97]).

However, one of the first major issue remains to know if the product of a translated sORF has a function in the cell [48]. Since all large-scale techniques are susceptible to generate false positive, the fact that sPEPs could be artifact of the techniques used for their discovery is still discussed [133]. Nonetheless, the growing body of evidence about functional sPEPs shed the light on their actual existence at peptide level, their stability and of their functionality, including if the number of sPEPs in human and model species is still controversial and varied by orders of magnitudes between studies [133]. If the use of labels or tags is usually relevant for exploring functions of unknown proteins [133], it could be especially tricky when it comes to sPEPs, as such sequences are sometimes of a similar size or even longer than the sPEP of interest itself. Thus, the size and biochemical properties of the tag itself (such as hydrophobicity or charge) must be considered when analyzing data from such experiments [147].

---

<sup>4</sup>The proteome of a cell is defined as the full set of proteins and peptides actually expressed in the cell. When referred to at species or organism level, the proteome actually refers to the full set of proteins (and peptides) that can be synthesized from the genome of this species or organism. By definition, sORF-encoded peptide (sPEP) should thus be considered as being part of the proteome as soon as there are evidence of their existence at the peptide level. The proteome may sometimes be distinguished from the peptidome, which contains only the shortest proteins (*i.e.* the peptides).



## 1. Introduction – 1.2. Short open reading frames encode functional peptides

To my knowledge, large-scale functional characterization and annotation of **sPEPs** (including the shortest ones) are still missing and interactions of these novel peptides with the canonical proteins (designated hereafter as **reference proteins (RefProts)**) remain mostly uncharacterized. Because elucidation of the role of unknown proteins have been successfully performed by studying their interactions with proteins of known functions in the past, I suggest to use similar approaches to explore and discover the functions and biological processes in which **sPEPs** are implicated. In addition, Gray *et al.* [48] emphasizes that **sPEPs** are typically missing from large-scale protein localization and interaction studies, and predicting the first large-scale **sPEP-RefProt interaction network** will tackle this issue.

### 1.2.2. Eukaryotic genomes should no longer be described as monocistronic

I must first stress out the fact that the definition of **ncORFs** reveals a dual issue of biology and semantic. In particular, transcripts lacking canonical **ORFs** longer than 100 **aa** were early defined as **ncRNAs** (including **lncRNAs**, **miRNAs** etc.) [102, 133], whilst the recent detection of **sPEPs** encoded by **sORFs** on such transcripts questions the actual relevance of referring to them as **non-coding RNAs**. Hence, calling such transcript "non-coding" does not make sense anymore and they should be reclassified as **mRNAs** [8, 81, 159], in particular regarding the fact that they share features with **mRNAs**, such as capping and polyadenylation [36]. However, we still need to use the term of **ncRNA** to point at the fact that **sORFs** were identified on **RNAs** that were believed to be non-coding. As a consequence, the **ncRNA-ORF** term will be used in the rest of this manuscript to refer to **ORFs** encoded by **ncRNAs**.

Gray *et al.* [48] highlighted that, in general, the main defining and unifying characteristics of **sPEPs** is their short size and the fact that they commonly have been missed, a definition that is thus currently not related to any biological feature or property of these peptides.

Finally, eukaryotic genomes have been described for long as monocistronic, a dogma which is mainly accepted by the scientific community [54]. As a growing body of evidence suggests that many transcripts harbors several **ncORFs** or both **ncORFs** and canonical **ORF(s)**, this deeply questions the relevance of referring to such genomes as monocistronic. Because the existence of several **sORFs** encoding functional **sPEPs** on the same **RNA** has been proved in eukaryotes (including in mammals), we may already describe a polycistronic organization of eukaryote **mRNAs** [46, 88, 92, 109, 113]. In addition, some **sPEPs** interact directly with the protein encoded by the **CDS** of their transcript. Moro *et al.* [88] hypothesized that the co-expression from the same **mRNA** could facilitate the functionalisation of **sPEPs** and their integration in cellular pathways related to the main protein product. This hypothesis is supported by Samandi *et al.* [117] who even suggest that this coordinated transcriptional regulation could be similar to prokaryotic operons, and Andreev *et al.* [7] who stressed out that



## 1. Introduction – 1.3. Short open reading frames regulate the translation of CDSs

being located on the same RNA for two functionally related proteins may represent an advantage for the coordination of their expression. Based on these evidences, a growing community of scientists suggests not to describe any longer eukaryotic genomes as monocistronic [90].

### 1.3. Short open reading frames regulate the translation of CDSs

An inverse correlation between the number of uORFs within the transcript and the efficiency of CDS translation has been highlighted quite early, and the presence of uORFs correlates with reduced steady-state levels of transcripts [25]. This suggested a possible role of uORFs in the regulation of the translation as well as the stability of transcripts. Translational control is an essential step of gene expression that is tightly controlled and can change the final protein abundance more rapidly than through the synthesis of new transcripts. Indeed, the translation and ribosome occupancy can change in a matter of seconds whereas the synthesis of new RNAs occurs over many minutes [57, 164].

A growing body of evidence also demonstrated that during a global translational arrest, some stress-resistant CDSs are preferentially translated [7]. It has been observed that the mRNAs encoding for these CDSs possess translated uORFs and several models of regulation of the translation by the uORFs have been proposed since then [5, 6]. Under certain cellular contexts, uORFs engage initiating ribosomes and may reduce initiation of the translation at the CDSs through various mechanisms, including ribosome stalling, altering the ribosome's capacity to initiate, terminate and reinitiate translation [109]. To understand these various mechanisms, it is important to describe first the mechanisms of initiation of the translation, the most important rate-limiting step at which the translation is regulated and controlling post-transcriptional gene expression [6, 35, 74, 120, 164].

Because they were discovered first to be involved in the regulation of the translation, uORFs were considered for long as the sole class with such regulatory function among sORFs, even if some cases of translational regulation by dORFs or internal ORFs have been recently reported [97]. Hence, the exploration of the translational regulatory functions of sORFs in the scope of this thesis has been restricted to uORFs, including if some recent studies demonstrated that other classes of sORFs may also be key regulatory elements of the translation.

#### 1.3.1. eIF2 $\alpha$ factor is essential to the translation initiation

The first step necessary for the translation is the association of ribosomes to the transcript (Fig. 1.4). Translation can happen in a cap-dependent or alternative cap-independent process, meaning that the ribosome can start scanning the transcript

## 1. Introduction – 1.3. Short open reading frames regulate the translation of CDSs

from its 5'UTR extremity or fix it directly at a latter position, notably through the involvement of **internal ribosome entry sites (IRESs)** or **translation initiator of short 5'UTRss (TISUs)** and scanning-free translation initiation [11, 54, 102, 109, 157]. Cap-independent translation is facilitated by conserved **N<sup>6</sup>-methyladenine (m<sup>6</sup>A)** modifications in the 5'UTR which promote direct binding of a translation factor (**eIF3**) to a 5'UTR m<sup>6</sup>A that recruits itself the translation machinery. RNA methylation can be affected by various stimuli, including heat shock responses [109]. Nonetheless, the role of **uORFs** in the regulation of the translation has been quasi-exclusively studied in the context of cap-dependent translation so far. Hence I decided to focus on cap-dependent translation mechanisms and only those will be discussed in this manuscript.

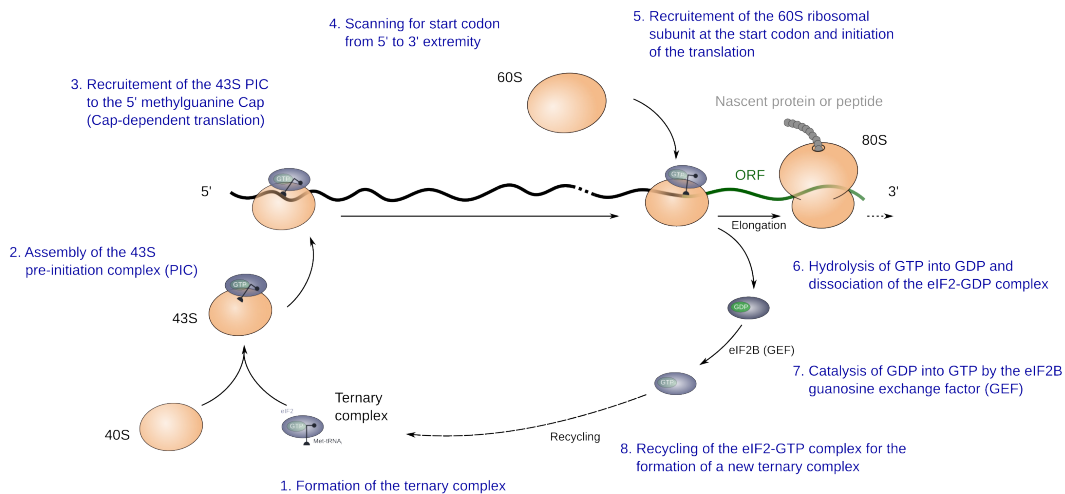


Figure 1.4.: **eIF2 $\alpha$  factor is essential to the translation initiation.**

Before the initiation of the translation, the **eIF2 $\alpha$**  factor assembles with a **guanosine triphosphate (GTP)** and a **methionyl-initiator tRNA (Met-tRNA<sub>i</sub>)** to form the **ternary complex**<sup>5</sup> [6, 7, 99, 102, 109]. The **ternary complex** then associates with the 40S ribosome subunit and several small initiation factors to assemble as the **43S pre-initiation complex (43S PIC)** [6, 7, 54, 99, 102, 109]. The **43S PIC** is finally recruited to the 5' methylguanine Cap of mRNA in a process helped by the cap-binding complex **eIF4F** [5, 54, 74, 99, 109]. Once fixated to the transcript, the **43S PIC** scans the mRNA from the 5'UTR extremity to its 3'UTR end until it recognizes a start codon [6, 54, 102, 109]. When a start codon is met, the **Met-tRNA<sub>i</sub>** anticodon is bound to the start codon (canonically an AUG), which triggers the recruitment of the 40S ribosomal subunit with translation initiation factors and results in the hydrolysis of **GTP** into a **guanosine**

<sup>5</sup>The ternary complex is a complex necessary to the initiation of the translation. It is constituted by the association of the eIF2 $\alpha$  factor with GTP and Met-tRNA<sub>i</sub> [99, 109].

## 1. Introduction – 1.3. Short open reading frames regulate the translation of CDSs

diphosphate (GDP) under the help of eIF5 [54, 74, 99]. This results in the dissociation of the factor eIF1 and the eIF2-GDP complex from the 40S ribosomal complex and produces a stable 48S pre-initiation complex [7, 54, 99]. This is immediately followed by joining of the large 60S ribosome subunit to produce an 80S initiation complex ready to begin the protein synthesis [54]. Finally, the GDP is catalyzed into a GTP by eIF2B, a guanosine exchange factor (GEF). This restores the eIF2 $\alpha$  factor to its active form, that can be used another time to form a new ternary complex [7, 99, 104, 132].

### 1.3.2. The phosphorylation of eIF2 $\alpha$ triggers a translational arrest

As previously described, eIF2 $\alpha$  is an essential factor for the initiation of the translation. This factor can be phosphorylated and its phosphorylation severely impairs the global level of protein synthesis and results in a translational arrest for most CDSs [7, 88]. Phosphorylated eIF2 $\alpha$  is a competitive inhibitor of eIF2B that blocks the conversion of GDP into GTP. This prevents the formation of the 43S PIC and results *in fine* into a global inhibition of the protein synthesis [7, 35, 54, 99, 104, 121]. Such phosphorylation of eIF2 $\alpha$  may be triggered by the integrated stress response (ISR), a conserved eukaryotic stress response to many stimuli, including viral and bacterial infections, growth factor deprivation, some cytokines, ribotoxic stress, stress granules sensing, heparin, amino acid deprivation, UV light sensing, heme deprivation, (arsenite-induced) oxidative stress, heat shock, osmotic stress, 26S proteasome inhibition, nitric oxide and endoplasmic reticulum (ER) stress [35, 99, 121, 132]. The ISR can also be activated by the unfolded protein response (UPR), a signaling pathway triggered by an ER stress during the sensing of misfolded proteins or a viral infection and known to promote inflammation [35, 73, 150]. The ISR can even be persistently activated in some pathological conditions, such as in mice with traumatic brain injury [104]. It is activated by the sensing of the stress by one of four kinases (GCN2, PKR, HRI or PERK) presenting different regulatory domains and that phosphorylate eIF2 $\alpha$  on Ser51. It is terminated by the dephosphorylation of eIF2 $\alpha$  under the action of either CReP or GADD34, two proteins respectively constitutively expressed in unstressed cells and specifically expressed in the later stage of ISR [7, 35, 99, 121, 132]. The ISR can be solved by cell survival (short-lived ISR) or cell death when the homeostasis cannot be restored (prolonged ISR), but the precise mechanism of switch between pro-survival and pro-death signaling is still largely misunderstood [35, 99]. The diminution of the global protein synthesis ensures a diminution of the translation of viral mRNAs in the case of an infection and diminishes the need of amino acids required for protein synthesis in the case of amino acid depletion [99]. Whilst the phosphorylation of eIF2 $\alpha$  has been shown to have protecting effects against metabolic and oxidative stress [99], an aberrant level of phosphorylation related to the malfunction of eIF2 $\alpha$  kinases has been shown to play a role in various pathologies [35]. Hence, the phosphorylation of eIF2 $\alpha$  is associated with a global repression of the translation but some

## 1. Introduction – 1.3. Short open reading frames regulate the translation of CDSs

particular CDSs, such as those encoding for [Activating transcription factor 4 \(ATF4\)](#)<sup>6</sup>, [ATF5](#), [CHOP](#) or [GADD34](#), are preferentially translated during the [ISR](#) [7, 99, 132]. It has been observed that the most stress-resistant CDSs possess [uORFs](#) and several studies demonstrated the importance of these [uORFs](#) in this stress resistance [5, 7, 102].

### 1.3.3. The regulation of the translation by uORFs is related to eIF2 $\alpha$ availability

Translation of the yeast transcription factor [general control non-depressible 4 \(GCN4\)](#) and its mammalian homolog [ATF4](#) is strongly increased under amino acid starvation. It has been early described to be regulated by [uORFs](#) and remains one of the best-studied example for regulatory [uORFs](#) [6, 25, 36, 102, 131, 132].

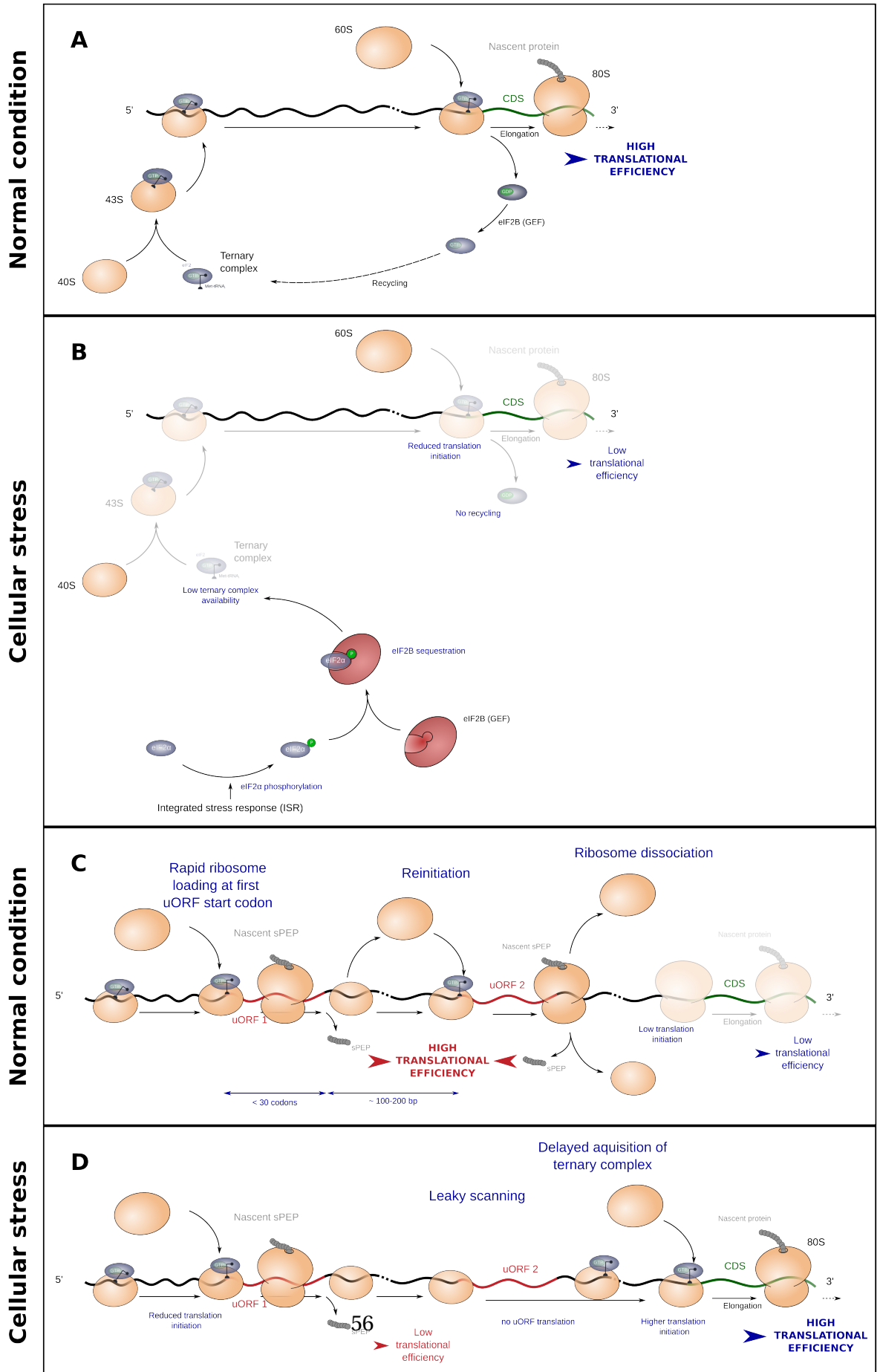
The 5'UTR of *S. cerevisiae* [GCN4](#) contains four [uORFs](#) that repress the translation of the CDS under normal conditions (Fig. 1.5). The presence of [uORF1](#) alone reduces the translation of [GCN4](#) by nearly 50%, mainly because after the translation of [uORF1](#), only half of the 40S ribosomes will remain attached to the [mRNA](#) and be available to reinitiate translation at the next start codon. In a similar way, [uORF2](#) enables reinitiation, likely as a backup mechanism to capture scanning ribosomes that would have failed to initiate at [uORF1](#). The translation of [uORF3](#) and [uORF4](#) favors the release of translating ribosomes, further preventing them from reaching the CDS start codon. Therefore, under normal conditions, the translation of [GCN4 uORFs](#) represses the synthesis of [GCN4](#). In summary, under normal conditions the two [uORFs](#) the closest to the CDS repress the CDS translation and will be translated but the two first are permissive for downstream translation. However, during a cellular stress triggering a decrease of [ternary complex](#) availability (such as the [ISR](#)), the association of the 40S ribosome with the [ternary complex](#) is delayed and the ribosomes that reinitiate after the translation of [uORF1](#) are more likely to bypass the three other [uORFs](#) ([uORF2](#), [uORF3](#) and [uORF4](#)), promoting efficient translation of [GCN4 CDS](#) [6, 25, 88, 102, 109]. It should be noticed that recent [Ribo-seq](#) experiments identified additional non-AUG initiating [uORFs](#) on the transcript of [GCN4](#), which nevertheless appear dispensable for the translational control of the CDS [102].

---

<sup>6</sup>Activating transcription factor 4. ATF4 is a leucine zipper (bZIP) transcription factor of the ATF/CREB family. ATF4 is able to form many homodimers as well as heterodimers with other bZIP transcription factors, including [CHOP](#). It is known to regulate more than 400 genes important in cell survival and cell death. It has been reported to have functions in the regulation of obesity, glucose homostasis, energy expenditure, neural plasticity, lipid metabolism, thermoregulation, muscle weakness in aging, memory formation, autophagy, amino acid transfer and biosynthesis [99]. ATF4 is activated during the integrated stress response (ISR) [73, 99, 104, 109, 150] and notably promotes the transcription of [CHOP](#) and [GADD34](#) [35].

1. Introduction – 1.3. Short open reading frames regulate the translation of CDSs

mRNA no harboring uORFs



Stress-resistant protein-encoding mRNA (ATF4-like mechanism)

## 1. Introduction – 1.3. Short open reading frames regulate the translation of CDSs

### Figure 1.5.: ATF4-like mechanism of regulation of the translation by uORFs. (A)

Considering a protein encoded from a canonical ORF and expressed under normal condition, and under the assumption of the canonical model of translation (*i.e.* without uORF regulation), translation initiation occurs normally (Fig. 1.4) and results in the protein synthesis. (B) However, under a cellular stress, such as the integrated stress response, eIF2 $\alpha$  is phosphorylated and sequestered by the eukaryotic translation initiation factor 2B (eIF2B) GEF, reducing the global availability of ternary complex, which results in turn in a global repression of the translation. (C) In the archetypal model of ATF4, the transcript harbors two functional uORFs in its 5'UTR. Translation of uORF1 occurs and is followed by a reinitiation, a mechanism by which the 40S subunit resumes scanning despite lacking the ternary complex (reinitiation) and is able to initiate again the translation at the uORF2, resulting thus in the synthesis of sPEPs. However, the ribosome dissociates at the end of the translation of uORF2 and no reinitiation is possible, which lead to a reduced synthesis of the RefProt. (D) Under stress, after the translation of uOR1, the 40S subunit resumes scanning despite lacking the ternary complex but, as the distance scanned by the 40S before acquiring the ternary complex depends on ternary complex availability, reinitiation occurs at the CDS start codon. GEF: guanosine exchange factor, ISR: integrated stress response.

To the difference of GCN4, ATF4 harbors only two short uORFs, but the mechanism described is quite similar. Under normal conditions, uORF1 is efficiently translated, a fraction of 40S ribosome remain attached to the transcript and the scanning resumes after, allowing reinitiation at uORF2. In such cases, the translation of a downstream element is only made possible by the re-acquisition of the other initiation factors before the 40S reaches the start codon [74]. Because uORF2 overlaps the CDS, there are less scanning ribosomes reaching the start codon of this last and the translation of uORF2 represses the expression of ATF4 CDS. Under stress conditions, reduced ternary complex availability together with leaky scanning (*i.e.* the ability of the ribosome to read-through uORF without initiating) allow bypassing the inhibition by uORF2. A fraction of re-scanning ribosome re-binds the ternary complex after leaky scanning the inhibitory uORF2 and initiates the translation at the CDS [7, 54, 109].

Because uORFs have been identified in more than half of mammalian transcripts, they may actually constitute key players in the regulation of the translation [28, 32, 54, 74, 84, 102, 109, 132]. A growing number of ribosome profiling studies identified uORF-mediated regulation in several species (such as for FIL1, UBI4 and eIF-5 in yeast [88] or CHOP, GADD34, CREP in mammals [7, 99, 113, 132]), but molecular mechanisms of regulation that differ from the one described for ATF4 are rare. Indeed, mechanisms to explain the translational regulation by the uORFs are limited only to a few set of CDSs and it is important to mention that even the archetypal example of the regulation of ATF4 expression is still debated. Some recent ribosome profiling data showed increased translation of ATF4 uORF2 upon stress. This contradicts the admitted model of leaky scanning that assumes that uORF2 translation is bypassed. As the model of regulation of ATF4 by uORFs is one of the best characterized, this raises questions about the accuracy of this last and highlight the importance to pursue our efforts in unraveling the mechanisms of translational regulation by uORFs [109].



### 1. Introduction – 1.3. Short open reading frames regulate the translation of CDSs

In addition, this model involves at least two **uORFs**, whilst examples of modulation of the translation by one single **uORF** have been described, suggesting that other mechanisms of regulation of the translation during cellular stress exist [5]. As an example, a single **uORF** seems to be sufficient to ensure the translation of the **CDSs** of **GADD34**, **CREP**, **CHOP** or **IFRD1** under stress [6].

Despite being commonly considered as inhibitors of the translation under normal condition, evidence that particular **uORFs** do not affect or enhance translation of downstream **ORFs** or that **uORFs** translation is up-regulated under stress exists also for a relatively small number of genes [54, 88]. Chew et al. [32] demonstrated that vertebrate **uORFs** tend to have features associated with a weak repressiveness, including if they were able to show an association between **uORFs** and the diminution of **CDS** translation in a "dose-dependent" manner (more **uORFs** in transcripts was correlated with reduced translation of **CDSs**). They demonstrated that **uORFs** are generally modestly repressive towards downstream **CDS** translation and that various sequence features modulate the **uORF** repressiveness and they showed that such features are broadly conserved among vertebrates.

Finally, whilst most stress-resistant **mRNAs** possess **uORFs**, only a small fraction of **uORF**-containing **mRNAs** are actually stress-resistant, demonstrating that the sole presence of an **uORF** on the transcript is not sufficient to ensure stress-resistance of the **CDS** [5, 6]. In addition, Gerashchenko et al. showed that increased ribosome occupancy at the **5'UTR** does not affect the **translation efficiency** of the downstream **CDS** [88]. Hence, the translation of most **uORFs** is unlikely to systematically affect the translation of downstream **CDSs** [88]. Some studies did not observe an overall positive or negative correlation between the translation of **uORFs** and that of their **CDS** as could be expected under the archetypal model of regulation [10].

It should be mentioned that alternative mechanisms of translation that do not require **eIF2 $\alpha$**  exist also [54, 109] and may explain a part of the variability observed. In particular, Leucyl-tRNA<sup>Leu</sup> can be engaged by a scanning **43S PIC** in a manner requiring the non-canonical initiation factor **eIF2A**, but not **eIF2**, a mechanism that has notably been shown to occur in the synthesis of antigenic precursors for loading on **major histocompatibility complex (MHC)** molecules [54].

Hence, there is an important need for new models to explain the *cis* regulatory functions of **uORFs** and to identify the features actually important to such regulation. Andreev et al [5, 6] proposed a stochastic and a deterministic model to explain the regulation of the translation by a single **uORF**. I suggest to implement this mathematical model with new experimental data to improve our understanding of the mechanisms of regulation of the translation by **uORFs**.

### 1.3.4. uORFs are involved in many processes and diseases as translational *cis* regulator

Interestingly, stress response genes have been shown to be significantly enriched in the group of genes up-regulated at the translation level and uORFs may play a key role in the rapid activation of such genes [88]. Some uORFs have been shown to be involved in development, learning and memory, cardiovascular diseases, neurodegenerative disorders and cancers [54, 102] and many diseases have already been reported to be associated with a dysregulation of the translation [102, 163, 164], notably because alteration of uORFs could result in aberrant protein levels and subsequently in diseases [102, 156]. uORFs have notably been shown to control the translation of components of integral developmental signaling pathways, such as SHH, WNT, PI3K, MAPK as well as many pluripotency factors, including the homeobox protein NANOG and C-MYC [109]. By the way, the list of mutations associated with human diseases that increase or decrease the influence of uORFs on the translation of the CDS is growing over time [54].

As an example, it has been shown that the G185A<sup>7</sup> mutation in the 5'UTR of SOX9 in human creates a novel ORF of 62 codons encoding for a functional sPEP. This novel uORF reduces the translation efficiency of wildtype transcription factor SOX-9 (SOX9) and causes campomelic dysplasia, a rare semi-lethal developmental disorder characterized by a distinctive pattern of abnormal skeletal features [109, 148].

The C178T<sup>8</sup> mutation in the sORF of a serotonin receptor gene HTR3A is responsible of a decrease in the repressive activity of the uORF, leading to an increase of HTR3A protein levels, which has been demonstrated to be associated with bipolar disorder and depression [109]. Also, disruption of a protein patched homolog 1 (Ptch1) uORF leads to decreased overall hedgehog signaling activity and disrupts neurogenesis [109].

Other mutations have been reported to disrupt an existing uORF of the thrombopoietin (TPO) gene, resulting in an increased translational efficiency of the CDS. This causes hereditary thrombocytosis, *i.e.* an increased number of platelets in the peripheral blood, along with a increased thrombosis risk [109].

Even expression of tumor suppressors and oncogenes have been shown to be under the regulation of uORFs, such as C-MYC, BCL-2, MSH5, PTEN, P53 [109].

---

<sup>7</sup>A guanosine is replaced by an adenosine at position 185

<sup>8</sup>A cytidine is replaced by an thymidine at position 178, resulting in a change of a proline for a serine



## 1.4. Short open reading frame variants are conserved across species and involved in the etiology of diseases

### 1.4.1. sORFs are conserved across species

The conservation of an sORF and its surrounding genes is usually considered as a clue that this sORF and/or its product may have a function [48, 113]. However, as previously discussed by Couso and Patraquim [36], the true conservation and homology of sORFs is difficult to establish, in particular because short sequences tend to have lower conservation score than longer canonical proteins and because the probability of short sequences to get a low conservation score by chance is higher. They report that uORFs show low average conservation and their amino acid usage is different from random values, but is slightly different from canonical proteins. In addition, Lee et al. [74] demonstrated a weak selection to maintain amino acid identity in sPEP encoded by uORFs compared to canonical proteins. By comparing human 5'UTRs with shuffled transcriptomes, Samandi et al. [117] observed that the density of altORFs observed in the mRNA leader sequence was much lower than in the shuffled transcriptome, supporting evidence that negative selection eliminates uORFs.

However, Mackowiak et al. [81] demonstrated in 2015 that predicted sORFs show stronger conservation signatures than those identified in previous studies and are sometimes conserved over large evolutionary distances. They showed that sORFs are often widely conserved at the sequence level, with uORFs being the most conserved class of sORFs between species [81]. Other recent findings suggest that some sPEPs are evolutionary conserved [25, 38, 60, 139] and that genomic positions with the potential to produce new uORFs are strongly conserved across vertebrates [74]. Lee et al. [74] showed also that uORF start codons are frequently conserved across species. Some other studies, including those performed by Chew et al. [32], demonstrated also that the regulatory effect related to the presence of the uORF is conserved across vertebrates, in particular regarding their role in repressing CDS translation [25, 28, 56, 61], which suggests that the regulatory effect of uORFs may be more conserved than their sequences themselves. Zhang et al. [163] showed that start codons of uORFs with Kozak contexts, particularly the translated ones, tend to be maintained by functional constraints during evolution.

Despite a growing body of evidence of the conservation of sORFs across species, it is to note that this does not necessarily mean that the sPEPs are functional, as the coding region can be constrained for optimizing exclusively *cis* regulatory functions [163].

Finally, it has been proposed several models for the generation of sORFs over time: they may have been generated (i) from existing protein-coding sequence, where sORFs emerge as fragments of longer protein-coding genes and only a small fragment of

## 1. Introduction – 1.4. Short open reading frame variants are conserved across species and involved in the etiology of diseases

the original ORF remains intact; (ii) from a small region of a larger protein (*e.g.* a transmembrane domain) that has been duplicated and acquired a separate function; or (iii) because of a *de novo* formation from previously non-coding sequences. It is likely that some sORFs have emerged from the three evolutionary mechanisms proposed here. As highlighted by Couso and Patraquim [36, 48], if sORFs may appear at random, their nucleotide sequences are subjected to selection, whatever this selection is related to a coding or a non-coding function and needs to be ascertained [48]. Zhang *et al.* [163] proposed that the majority of newly formed uORFs are deleterious and quickly removed from the population, whilst a smaller fraction are beneficial and rapidly fixed in population under positive selection.

### 1.4.2. sORF variants have been related to diseases

As discussed earlier, a growing body of evidence demonstrates that changes in sORF functionality are linked to diseases [32, 129, 132]. Sequence variations in human 5'UTR have notably been associated with variations in gene expression and mutations in uORFs are now known to contribute to diseases [32]. Mackowiak *et al.* [81] observed that predicted sORFs mainly permit synonymous more than non-synonymous sequence variations when comparing within or between species, suggesting the importance of the peptide sequence. However, using such ratio of nonsynonymous over synonymous substitutions can be tricky for sORFs as the number of possible changes is low. Interestingly, McGillivray *et al.* [84] demonstrated that despite CUG being the most prevalent start codon in usage frequency, it is altered relatively frequently by natural human variants, whilst AUG is relatively conserved among uORFs. It has also been demonstrated that translated uORF stop codons were significantly depleted of UAAs compared to background UTRs distributions, suggesting that weaker stop codons are preferentially used by uORFs [74].

Lee *et al.* [74] demonstrated that uORF variants introducing new stop codons or strengthening existing stop codons are under strong negative selection comparable to protein-coding missense variants. This result suggests that upstream stop codon variants may functionally disrupt the protein expression [74]. In the same manner, variants destroying stop codons in translated uORFs are under strong negative selection too, as resultant translational read-through can decrease the start codon recognition and the translation initiation at the CDS [74]. They also showed that genetic variants creating new uORFs are rare, and suggested they are also subjected to strong negative selection due to their capacity to cause pathogenic loss-of-function of associated protein [74].

Heterozygous and homozygous individuals carrying 5'UTR stop codons and stop-strengthening variants have been described and associated with pathological conditions [74]. In particular stop-strengthening variants in *PMVK* have been associated to an increased risk of Type 1 diabetes; in *VPS53* to a protective effect against anxiety disorders; and in *HMT2* to cardiac and movement disorders, including congenital

## 1. Introduction – 1.5. Many questions about short open reading frames and their functions remain unanswered

anomalies of the great vessels, abnormal involuntary movements, abnormality of gait, Mobitz II atrioventricular block, and arrhythmia [74].

As another example, *GCH1* has been associated with familial Dopa-responsive dystonia<sup>9</sup>. Single nucleotide polymorphisms (SNPs) have been identified in its 5'UTR at the position +14 (changing a cytidine for a thymidin) that generates an upstream initiation codon ATG and consequently an uORF. This new uORF represses the translation of the CDS and allows instead for the translation of an aberrant, cytotoxic, 73 aa sPEP [109]. Zhang et al. [163] report also that mutations generating polymorphic uORFs are usually deleterious and selected against in humans.

Neville et al. [93] demonstrated that sORFs show large heritability enrichments characteristics of CDSs. Additionally, uORFs that overlap their CDSs showed larger heritability enrichments than those which do not, suggesting a possible functional importance on heridity [93]. They also highlighted that disease mutations that appear benign to canonical proteins may be highly deleterious to sORFs and hypothesized that numerous variants in disease mutation database could potentially have sORF-related mechanism of pathogenicity (stop-lost, stop-gained, frameshift mutations) [93]. Hence, they identified potential disease-causing variants in ncORFs, in particular cancer-associated genes with mutation with benign consequences in CDS but with deleterious consequences in the sORF [93].

### 1.5. Many questions about short open reading frames and their functions remain unanswered

Growing efforts have been made by the scientific community during the past decades to better understand the functions of sORFs and their peptides. It is now clear that they are prevalently translated in eukaryotic cells and that uORFs can be key players of the translation regulation [163]. However, this novel class of molecules challenge our current understanding of genetics and there are still many questions that remain unanswered or even unexplored so far (Box 1).

---

<sup>9</sup>Dopa-responsive dystonia was described for the first time in 1972. Dystonic syndromes with L-dopa responsiveness are very heterogenous and its clinical presentation is still debated. Nonetheless, it has a classic presentation of childhood or adolescent-onset dystonia, mild parkinsonism, marked diurnal fluctuations, improvement with sleep or rest, and a dramatic and sustained response to low doses of L-dopa without motor fluctuations or dyskinesias [76]

1. Introduction – 1.5. Many questions about short open reading frames and their functions remain unanswered

Box 1.: **Some unanswered questions about sORFs.**

Some of the questions that remain opened in the field of sORF biology:

- Are all the sORFs actually functional (either as *cis* or *trans* regulatory elements)?
- What is the therapeutic potential of targeting sORFs or their peptides, regarding their apparent involvement in many processes and diseases?
- Could we edit existing sORFs or introduced artificial ones in the genome in the context of novel gene therapies?
- Could we use synthetic peptides as drugs to re-establish disturbed homeostasis related to sPEP loss-of-function in certain diseases?
- What is the place of the sORFs in the homeostasis of the cell?
- Are *cis* and *trans* regulatory functions of the sORFs related one with each other (for all/some transcripts)?
- Are there sPEPs that interact with their own transcript? Is this common?
- Are there sPEPs that interact with their own RefProt? Is this common?
- Are there "extreme multifunctional sPEPs", *i.e.* sPEPs whose multiple functions are very dissimilar to one another?
- How long-lived are sPEPs?
- What are the mechanisms of regulation of the translation by the uORFs? In which way do they differ of the ATF4-like mechanism?
- What are the features that make some uORF-harboring RNAs resistant to stress and some others not?
- Are there sub-classes of uORFs that have distinct effects on the translation?
- How is the uORF translation regulated?
- In the case of nested ORFs, does the translation of the longer ORF(s) affects the translation of the embedded shorter ones? or vice-versa?
- What makes an uORF more likely to repress or increase the CDS translation under stress condition?

1. Introduction – 1.5. Many questions about short open reading frames and their functions remain unanswered

- Could we consider the use of antisense oligonucleotides to specifically target some **uORFs** in order to enhance or reduce levels of disease-relevant proteins?
- Does early translation termination in **uORFs** facilitate active **nonsense-mediated decay (NMD)**?
- What is the importance of **uORFs** in the regulation of protein levels during changes in the cellular identity along development trajectories?
- Can the interactions of scanning and elongating ribosomes trailing along the **uORFs** explain the up-regulation or down-regulation of downstream expression under certain conditions?
- Can the translation of **uORFs** result in structural changes of the transcript? Or reveal new **IRESs**?
- Are there specific sequences or stop codons in the **uORFs** that are susceptible to stall ribosomes during the elongation or termination?
- Does the presence of **uORFs** in the **5'UTR** facilitate ribosomal frameshifts?
- How do **uORFs** and their translation impact the stability of transcripts?
- How is the heterogeneity of ribosomal loading and translational initiation mechanisms exploited in a cell to control **uORF** and **CDS** translation?
- Do (putative) non-transcribed intergenic **sORFs** have functions?
- Are **sORFs** and **CDSs** showing an additive effect in heritability?
- Are there operon-like systems in eukaryotes? In particular on presumptive **lncRNAs** that harbor several coding-**sORFs**?
- Are **uORFs** transcriptional *cis* regulators?
- Are there alternative transcriptions that favor the encoding and expression of some **sORFs**?

In the frame of my thesis project, I decided to explore the following questions:

1. Can data about the **sORFs** identified in *H. sapiens* be gathered as unique entries into a publicly accessible repository?
2. What are the biological functions of the **sPEPs**?
3. Could we elucidate new mechanisms of translational regulation by the **uORFs**?

## 1. Introduction – 1.5. Many questions about short open reading frames and their functions remain unanswered

As highlighted by Aspden et al. [10], *the putative function of sORFs and their encoded peptides is a separate issue from their translation, just as the transcription of thousand of apparently ncRNAs is an accepted fact separated from their, as yet, not fully understand function.* Fields et al. [46] also state that the translation may be functionally important independently of the sequence of the encoded peptide. Hence I propose here to explore the questions (2) and (3) as complementary but separate issues, despite their obvious entanglement. In addition, these questions have been addressed using computational approaches on monocytes.

Human monocytes are an heterogeneous population of innate immune cells of the mononuclear phagocyte system that may differentiate into macrophages and play a major role in the initiation of immune responses. They are able to express molecules of the **major histocompatibility complex class-I (MHC-I)** and **MHC-II**, which make them of particular interest as numerous **sPEPs** have been determined to be able to fixate the **MHC-I** and be presented as self-antigens with high predicted binding affinities [25, 35, 44, 64, 72, 96], a process which is suspected to depend mainly on translation rates rather than overall peptide abundance [44]. In addition, because the presentation of peptides by **MHC** molecules is largely independent of the amino acid sequence, and many **sPEPs** may not need proteosomal degradation before entering the **MHC-I** presentation pathway, a certain fraction of **sPEPs** is likely to be involved in immunological functions [25, 72]. Recent estimates suggest that **MHC-I** alone could display up to 120,000 peptides on each cell surface [132] and that more than half of **MHC-I**-associated peptides may result from out-of-frame translation [72]. They are specialized cells that are able to activate immunology responses, such as the production of cytokines under the sensing of a pathogen [111]. Monocytes are derived from a bone marrow progenitor common to **dendritic cells**. These last are known to display extremely high levels of **eIF2 $\alpha$**  phosphorylation both *in vivo* and *in vitro*, suggesting they should have special needs regarding their translation regulation. Finally, since much of the functional data is based on conditions that knock-out whole transcripts, validating **uORF** functions is difficult [10]. Hence, it may be more efficient for the moment to base studies on specific cell contexts. It is important to note that publicly available data (notably from **Ribo-seq** experiments) identifying **sORFs** in monocytes were available in 2018 and can be exploited to explore the questions previously raised. Altogether, monocytes constitute a good model to study the functions of the **sORFs** and their **sPEPs** and the **chapter 3** and **chapter 4** focus on their roles in monocytes.

In order to discuss the questions previously stated, I thus decided to:

1. Gather all **sORFs** identified in *H. sapiens* into a repository (**chapter 2**)
2. Ascertain **sPEPs** functions in monocytes through a system approach, by predicting their interaction with canonical proteins (**chapter 3**)
3. Explore mechanisms of translational regulation by the **uORFs** in monocytes, by using agent-based modeling to imitate the translation process (**chapter 4**)

## 2. sORFs identified in human and mouse genomes have been gathered as unique entries in a database: MetamORF

### 2.1. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis

Study of the sORFs has started to be a hot topic since a few years. Whatever we are considering experimental or computational tools for their study, any approach that aims at better characterizing them first requires to identify them properly. Similar reasons supported the development of repositories of canonical (yet uncharacterized) proteins a few decades ago, with the willing to gather at the same place all the current knowledge about them, in a way that makes this information easily accessible to end-user biologists (including those without advanced computational skills). This requirement, along with the emergence of computational sciences and of the internet, accelerated the development of bioinformatics and computational biology in the late 1990s. A few years later, well-known databases, such as SWISS-PROT, NCBI and Ensembl databases were released. Those databases now constitute crucial resources that are known by any biologist and used on a daily basis by both experimental and computational biologists.

However, for many reasons explained earlier, sORFs have been missed for long. Because we still lack many information about them and because of the ongoing discussion in the scientific community regarding the relevance to consider them as actual functional elements, these biological entities are still missing from the most commonly used repositories. This led to the development of novel resources, that aim at exclusively gathering information about ncORFs or sORFs. So far, the two main active repositories that fall into that category are sORFs.org [94, 95] and OpenProt [23, 24]. These publicly accessible resources present a mine of information waiting to be exploited, as they gather computational and experimental evidence of the existence of sORFs. However, it is noteworthy that OpenProt is missing all sORFs shorter than 30 aa, a class of ORFs I am particularly interested in the frame of my thesis regarding the growing body of evidence that suggest their functionality in the literature. On



2. *sORFs identified in human and mouse genomes have been gathered as unique entries in a database: MetamORF – 2.1. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis*

the other side, when sORFs.org is not missing such short ORFs, it contains a lot of redundancies and does not provide the opportunity to easily look for sORFs identified on a particular transcript or gene. Such missing feature makes their analysis more complicated and would be of great interest for biologists, especially for those without advanced computational skills. In addition, it is unfortunate that despite the important number of datasets that have been reprocessed and integrated in sORFs.org, the redundancy cannot be easily exploited by the users to identify at a glance the sORFs identified by many datasets. Finally, as discussed earlier, no clear consensus regarding the nomenclature to use to annotate sORFs was designed in 2018 (and it is noteworthy that this is still not the case in 2022, despite growing efforts of our community to address this particular issue). Hence, comparing sORFs coming from various studies or resources can quickly become terrible as different authors used sometimes the same annotation with distinct definitions, sometimes even incompatible (*e.g.* an "altORF" may refer exclusively to ORFs using an alternative reading frames according to some authors whilst it may be used to designate at any ncORF by some others).

Because we were willing to get a resource of unique human and mouse sORFs, as comprehensive as possible, with proof of existence from complementary methods, and providing homogenized information (coordinates on the same genome annotation, homogenized nomenclature of sORFs etc.), I proposed to take advantage of existing resources to address these particular issue. I was also willing to post-process the entries integrated in order to provide some novel information (such as Kozak contexts) more easily accessible to the users. I notably took advantage of the strengths of sORFs.org, which already reprocessed lots of data from Ribo-seq experiments and integrated data from 73 original studies at that time. After careful evaluation of 18 data sources [6, 32, 44–46, 51, 61, 72, 75, 78, 81, 84, 95, 110, 117, 123, 152, 155] (either single datasets from original publication or repositories of reprocessed data) of sORFs in *H. sapiens* and/or *M. musculus*, 6 of them [44, 61, 72, 81, 95, 117] were matching the selection criterion I defined (*e.g.* absolute sORF coordinates on the genome) and were retained for inclusion in a new database, called MetamORF. The remaining 12 data sources were discarded because they were not providing the absolute genomic coordinates of the ORF start and stop codons [6, 32, 45, 110, 123, 155], because they were already fully included in another data source [46], because they did not allow export of the full database [78, 152] or because they were missing crucial information regarding the ORF splicing [51, 75, 84]. The six data sources selected had the quality to provide crucial information to characterize the ORFs (such as absolute genomic coordinates and a clear identifier for the transcript(s) harboring them) and to be derived from either computational prediction or experimental evidences (Ribo-seq, MS-based proteomics or proteogenomics).

The first step for building MetamORF consisted in integrating these data in an uniform format, an apparently easy task that in reality raised lots of issues because of the diversity of formats in which the data were available, and mainly because of the lack of well-documented metadata. This difficulty highlight the absolute necessity



2. *sORFs identified in human and mouse genomes have been gathered as unique entries in a database: MetamORF – 2.1. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis*

for our community to agree on rules for storing and sharing biological data, and I personally acknowledge and strongly encourage initiatives such as the definition of FAIR (Findability, Accessibility, Interoperability, and Reusability of digital assets) principles or the Dublin core initiative. As the amount of data generated by biologists becomes more and more important over years, the respect of such principles will become an absolute necessity to ensure the reliability, traceability and usability of data in a near future.

Because genome annotations have evolved a lot over time, and data about **sORFs** have been published over several years, those were mapped on several genome annotations (notably GRCh37 and GRCh38 releases for human **sORFs**). Hence, my next task was to bring all **sORF** coordinates on the same genome. Despite the availability of tools to perform such coordinate homogenization (called *liftover*), this operation required additional steps to ensure the consistency of data (as an example, an **ORF** having start and end positions of its exons mixed together because of inconsistency in the conversion of genomics coordinates was discarded).

Once data had been inserted into a query-able database and normalized by bringing genomics coordinates to the same annotation, the next step (definitively one of the most important for MetamORF) consisted in merging all redundant entries into single entries. This actually constitutes an originality of our database, as the aggregation of redundant entries into single ones allows *in fine* to provide to the user the number of computational and experimental proofs of existence for each **sORF**.

Finally, data were post-processed to include some novel information to MetamORF. In particular relative coordinates of **sORFs** on their transcripts were computed and exploited to annotate the **ORFs** based on a novel nomenclature where I tried to take into account the most frequently used annotations at that time. Because original data sources were not providing cell types as understandable terms, I also reprocessed this information in order to get cell types that could be easily understood by the end-users, and connected those terms to existing ontologies (Cell Ontology [40], Cell Line Ontology [118], BRENDA Tissue Ontology [29], Human Cell Atlas Ontology [154], Foundational Model of Anatomy Ontology [47], Ontology for Biomedical Investigations [13], NCI Thesaurus OBO Edition [126], Experimental Factor Ontology [83], BioAssay Ontology [1] and Ontology for MIRNA Target [55]). Finally, by looking for regular expressions, I computed Kozak contexts [53, 69] based on local sequences near to the start codons. If the role of this context was clearly demonstrated for canonical **CDSs** by many times from its initial discovery by Marilyn Kozak (1986) [69], its importance in the particular case of **sORFs** is still largely debated. As a consequence, I assumed that having such information in a database of **sORFs** would be of primary interest for people willing to tackle this still-debated question.

The last step of this work was to build an user-friendly web interface, a common task for database developers, but that brought many difficulties because of the complex nature and the huge amount of the biological data manipulated there. This interface allows notably to navigate easily between the **ORFs**, transcripts and genes, to get the most important information (number of computational and experimental evidences,

2. *sORFs identified in human and mouse genomes have been gathered as unique entries in a database: MetamORF – 2.1. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis*

data sources, absolute and relative coordinates, sequences, Kozak contexts etc.), to export them at convenient formats (notably FASTA and BED formats) and to visualize the **ORFs** on the genome using the UCSC genome browser [65]. The development of the web interface has been performed in close collaboration with A. Wagner, a technology degree graduate student (*DUT*) I had the opportunity to co-supervise (using agile methods).

MetamORF finally registers 1,162,675 unique **sORFs**, identified in *H. sapiens* and *M. musculus* and derived from the processing of 5,445,846 original entries. It is to note that despite their apparent simplicity, all the steps described above require computational skills, many hours of processing and eventually access to advanced computational facilities. In addition, even for computational biologists, it can be inconvenient to perform such many tasks prior to start studying **sORFs**. By developing MetamORF, I think and hope it is providing an easy-to-use resource to the scientific community. It is for sure that no database can be fully comprehensive, and MetamORF will definitely not escape this rule. However, it offers complementary information with existing databases (notably **sORFs.org** and **OpenProt**), and is the first database to provide at the same time homogenized information about **sORFs** (without size limit) identified by both computational and experimental methods and accessible at both the **ORF**, transcript and gene levels. Clearly, MetamORF does not address by itself any of the biological questions previously raised; however, it provides scientists the necessary data to address such questions. To that extent, it is important to note that MetamORF data can be exploited for large-scale studies (which I did during my thesis and present in the following sections), or even for low-scale studies, by helping scientists to check known **sORFs** on their gene(s) of interest for instance. Finally, we may hope that future releases of MetamORF will integrate more data, in particular about new species, and continue to make links with existing resources such as **sORFs.org**. This update constitutes an important task for maintaining such repository and will probably require to perform in a few years a similar work to the one I did there. I think that maintaining relationships between resources is of primary importance, and should always be considered and supported during the development of databases. In our case, MetamORF did not intend to overstep existing databases, but instead to provide complementary information that were missing or not easily accessible in resources that were available at the time I started this project.

**Choteau SA**, Wagner A, Pierre P, Spinelli L, 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*, 10.1093/database/baab032, 2021:baab032.

2. sORFs identified in human and mouse genomes have been gathered as unique entries in a database: MetamORF – 2.1. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis



Database, 2021, 1–12  
doi:10.1093/database/baab032  
Original article



Original article

## MetamORF: a repository of unique short open reading frames identified by both experimental and computational approaches for gene and metagene analyses

Sebastien A. Choteau <sup>1,2</sup>, Audrey Wagner<sup>1</sup>, Philippe Pierre<sup>2,3,4</sup>,  
Lionel Spinelli<sup>1,2</sup> and Christine Brun <sup>1,5,\*</sup>

<sup>1</sup>Aix-Marseille University, INSERM, TAGC, Turing Centre for Living Systems, 163 Avenue de Luminy, Marseille 13009, France, <sup>2</sup>Aix-Marseille University, INSERM, CNRS, CIML, Turing Centre for Living Systems, 163 Avenue de Luminy, Marseille 13009, France, <sup>3</sup>Department of Medical Sciences, Institute for Research in Biomedicine (iBiMED) and Ilidio Pinho Foundation, University of Aveiro, Aveiro 3810-193, Portugal, <sup>4</sup>Shanghai Institute of Immunology, School of Medicine, Shanghai Jiao Tong University, Shanghai, China and <sup>5</sup>CNRS, 31 Chemin Joseph Aiguier, Marseille 13009, France

\*Corresponding author: Tel: +33 4 91 82 87 12; Fax: +33 4 91 82 87 01; Email: [christine-g.brun@inserm.fr](mailto:christine-g.brun@inserm.fr)

Citation details: Choteau, S.A., Wagner, A., Pierre, P. *et al.* MetamORF: a repository of unique short open reading frames identified by both experimental and computational approaches for gene and metagene analyses. *Database* (2021) Vol. 2021: article ID baab032; doi:10.1093/database/baab032

Received 23 December 2020; Revised 8 April 2021; Accepted 17 May 2021

### Abstract

The development of high-throughput technologies revealed the existence of non-canonical short open reading frames (sORFs) on most eukaryotic ribonucleic acids. They are ubiquitous genetic elements conserved across species and suspected to be involved in numerous cellular processes. MetamORF (<https://metamorf.hb.univ-amu.fr/>) aims to provide a repository of unique sORFs identified in the human and mouse genomes with both experimental and computational approaches. By gathering publicly available sORF data, normalizing them and summarizing redundant information, we were able to identify a total of 1 162 675 unique sORFs. Despite the usual characterization of ORFs as short, upstream or downstream, there is currently no clear consensus regarding the definition of these categories. Thus, the data have been reprocessed using a normalized nomenclature. MetamORF enables new analyses at locus, gene, transcript and ORF levels, which should offer the possibility to address new questions regarding sORF functions in the future. The repository is available through an user-friendly web interface, allowing easy browsing, visualization, filtering over multiple criteria and export possibilities. sORFs can be searched starting from a gene, a transcript and an ORF ID, looking in a genome area or browsing the whole repository for a species. The database content has also been made available through track hubs at UCSC Genome Browser. Finally, we demonstrated

© The Author(s) 2021. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Page 1 of 12

(page number not for citation purposes)

Downloaded from <https://academic.oup.com/database/article/doi/10.1093/database/baab032/6307706> by INSERM user on 19 July 2022

## 2. sORFs identified in human and mouse genomes have been gathered as unique entries in a database: MetamORF – 2.1. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis

an enrichment of genes harboring upstream ORFs among genes expressed in response to reticular stress.

**Database URL:** <https://metamorf.hb.univ-amu.fr/>

### Introduction

Short open reading frames (sORFs) are usually defined as sequences delimited by a start codon and a stop codon and potentially translatable into proteins of <100 amino acids (1–8). They are present in all classes of transcripts [including presumptive long non-coding ribonucleic acids (lncRNAs)] and have been identified in most eukaryotic RNAs (2, 5, 8–15). In addition, their sequence often begins with a non-canonical start codon (8). Consequently, they have long been overlooked, and interest in their possible regulatory functions has only raised recently with the advent of the ribosome profiling method that strongly suggests their translation (1, 3, 5, 6, 16–22).

Several sORF categories have been defined according to their location on RNAs (Figure 1). For instance, upstream ORFs (uORFs) are located in the 5' untranslated regions (5' UTRs) of messenger RNAs (mRNAs) and have been defined as sORFs whose start codon precedes the main coding sequence (CDS; 6, 8, 17, 18, 23). They are conserved across species (5, 6, 11, 21, 24), but less conserved than canonical protein-coding ORFs (25). To date, uORFs have been essentially reported as gene-expression *cis*-regulatory elements that regulate the efficiency of translation initiation of the main CDS, notably alleviating the repression of translation during cellular stress (13, 17, 18, 20, 23, 26). Moreover, the discovery of uORF-encoded peptides, and more generally sORF-encoded peptides, led to the assumption that they may also play functional roles in *trans* (2–4, 7, 9, 10, 18, 24, 27–30), for instance as ligands of major histocompatibility complex molecules (12, 22, 23). Very interestingly, uORF-encoded peptides have also been shown to form protein complexes with the protein encoded by the main CDS of the same mRNA (31), and it has been suggested that polycistronic sequences may exist in eukaryotes (24, 31). Furthermore, given the increasing evidence on the regulatory functions of peptides encoded by sORFs located within mRNAs, introns of pre-mRNAs, lncRNAs and primary transcripts of microRNAs or ribosomal RNAs (2, 8–15, 26), there is an urgent need to study sORFs (i) individually and (ii) at the whole proteome scale. Indeed, the latter should reveal important features of sORFs, thus enabling the characterization and the identification of their functions. However, the fact that (i) the publicly available data are scattered across different databases and (ii) datasets are aligned on different genome builds, differently annotated and formatted, calls for an uniformed resource

where each sORF is individually described. With this in mind, we have built a resource database of publicly available sORFs identified in the human and mouse genomes, by gathering information from computational predictions and Ribo-seq and proteomic experiments. The curation of data, their homogenization in order to merge the redundant information into unique entries, the completion and computation of missing information (e.g. sequences and Kozak contexts) and the re-annotation of sORF classes represent the added value of this database. Notably, this enables the analyses at locus, gene, transcript and ORF levels. In this work, we propose (i) a pipeline to regularly update the content of the database in a reproducible manner, (ii) a database content that can be fully downloaded for custom computational analyses and (iii) an user-friendly web interface to ease data access to biologists.

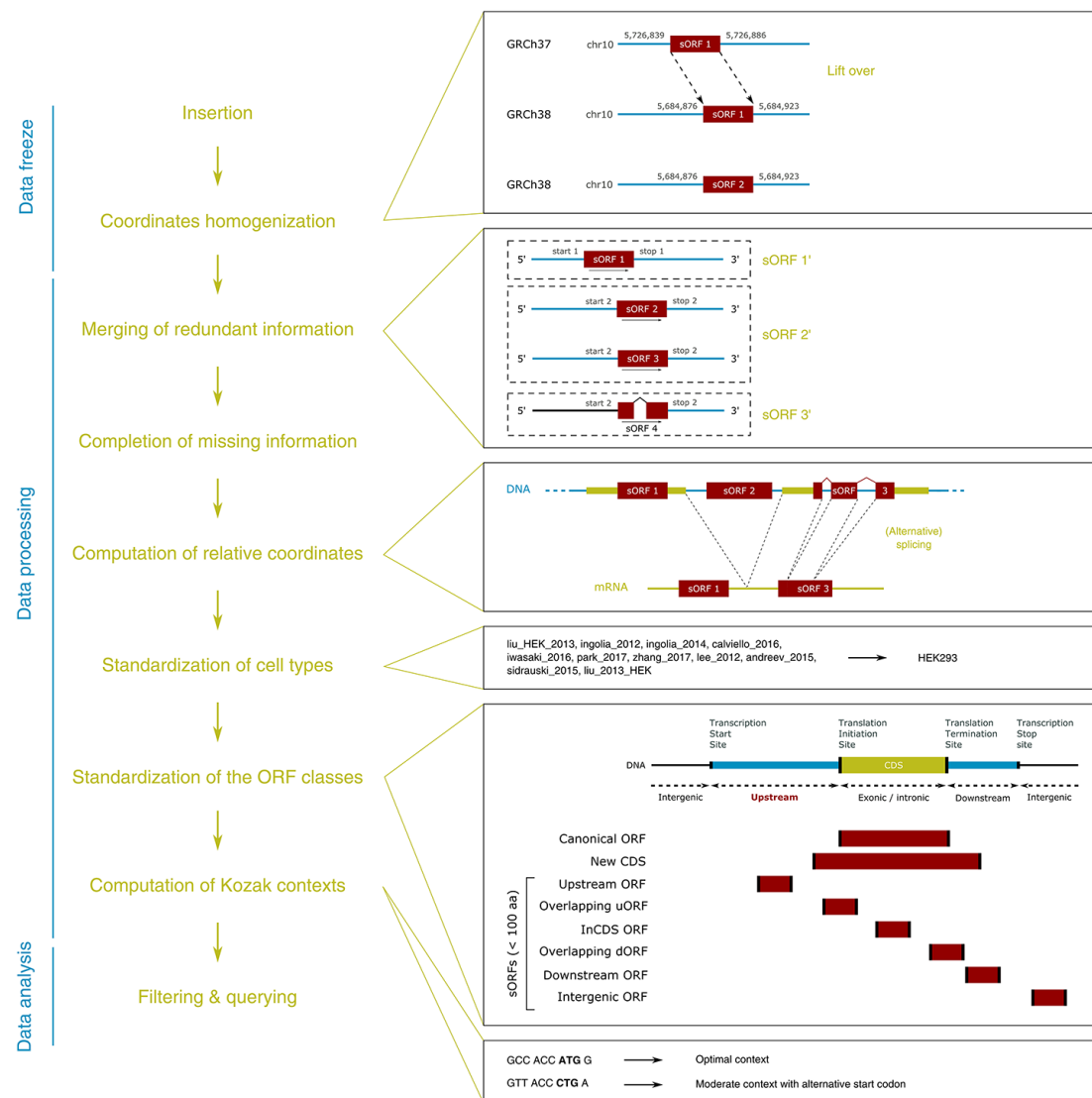
### Material and methods

#### MetamORF pipeline and database development

#### Inclusion criteria for publicly available sORF-related data

A total of 18 data sources, either *Homo. sapiens* and *Mus. musculus* original datasets or re-processed publicly available sORFs repositories, have been considered for inclusion in our database (Supplementary Table S1) (5, 7, 11, 12, 14, 15, 17–22, 32–37). These data sources provide results from computational predictions, Ribo-seq experiment analyses and mass spectrometry (proteomics/proteogenomics) analyses. The data sources not providing the absolute genomic coordinates of the ORF start and stop codons (5, 17, 20, 32–34) or fully included in another data source considered here (21) have been discarded. Databases that did not allow export of their content in a single file or automating the download of all the files from their website have also been discarded (19, 35). Despite their short size, it has been noticed that sORFs can be spliced. Theoretical lengths of the ORFs have been computed as the distance between the start and stop codons, eventually removing the intron length(s) when information about ORF splicing was provided. Due to splicing, the theoretical length and the one reported by the data source may be different. Data sources harboring this difference for >95% of their entries were discarded as this indicates the splicing information was missing (10). Finally, data sources for which we were not able to perform this assessment as they were not providing

2. sORFs identified in human and mouse genomes have been gathered as unique entries in a database: *MetamORF – 2.1*. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis



**Figure 1.** MetamORF pipeline. This figure represents the workflow used to build MetamORF. First, the data from the sources selected have been inserted into the database, and the absolute genomic coordinates have been homogenized from their original annotation version to the most recent version (GRCh38 or GRCm38). Then, the redundant information, i.e. the entries describing the same ORFs (same start, stop and splicing), has been merged, allowing to get one single and unique entries for each ORF detected on the human and mouse genomes. The missing information (sequences and transcript biotypes) has been downloaded from Ensembl, and the ORF relative coordinates have been computed. Finally, the cell types and ORF classes have been normalized, and the Kozak contexts have been computed using the sequences flanking the start codons.

information regarding (i) the splicing of the ORF and (ii) ORF length (15, 36) have not been included as well. Hence, the database has been made by collecting data from six distinct sources (Table 1), including either original datasets (Table 1 and Supplementary Table S2) (11, 12, 14, 18, 22) or reprocessed data (37), and discarding 12 of them (Supplementary Table S1). Notably, we have included data from sORFs.org (37), considered as

the main and most comprehensive repository of sORFs identified by genome-wide translation profiling (Ribo-seq), that currently integrates re-processed data from 73 original publications.

For each of these sources, a set of features essential to properly characterize the sORFs, related to their location, length, sequences, environmental signatures and cell types (i.e. cell lines, tissues or organs) in which they are expressed,



## 2. sORFs identified in human and mouse genomes have been gathered as unique entries in a database: MetamORF – 2.1. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis

**Table 1.** Information about the data sources used to build MetamORF

Publication	DOI
Mackowiak <i>et al.</i> , 2015, <i>Genome Biol.</i> (11)	10.1186/s13059-015-0742-x
Erhard <i>et al.</i> , 2018, <i>Nat. Methods</i> (22)	10.1038/nmeth.4631
Johnstone <i>et al.</i> , 2016, <i>EMBO J.</i> (18)	10.15252/embj.201592759
Laumont <i>et al.</i> , 2016, <i>Nat. Commun.</i> (12)	10.1038/ncomms10238
Samandi <i>et al.</i> , 2017, <i>eLife</i> (14)	10.7554/eLife.27860
Olexiouk <i>et al.</i> , 2018, <i>Nucleic Acids Res.</i> (37)	10.1093/nar/gkx1130

See Supplementary Table S1 for more information about these data sources.

have been collected (see Table 2 for a full list of features considered for inclusion). When it was not provided by the source, the symbol of the gene related to the sORF was recovered using the transcript identifier (ID, if provided) or searching for the gene(s) or ncRNA(s) overlapping with the sORF coordinates in the original annotation version by querying Ensembl databases (38) in their appropriate versions (v74, 75, 76, 80, 90) with pyensembl (v1.8.5, <https://github.com/openvax/pyensembl>). In addition to these features, information regarding the transcript(s) harboring the ORFs have been collected from the data sources when available. This is of particular interest as some ORF features, such as the ORF class, may depend on the transcript they are located in (e.g. an ORF may be located in the 5' UTR of a transcript and be overlapping with the CDS of another transcript). Finally, 3 379 219 and 2 066 627 entries from these six data sources have been collected and inserted in MetamORF for *H. sapiens* and *M. musculus*, respectively (Table 3).

### Homogenization of genomic coordinates

As the data sources were providing genomic coordinates from different genome annotation versions (e.g. GRCh38 and GRCh37), all the genomic coordinates registered in our database have been lifted over the latest annotation version (GRCh38 for *H. sapiens* and GRCm38 for *M. musculus*) using pyliftover (v0.4, <https://pypi.org/project/pyliftover>). The liftover has been considered as failed for an entry if (i) at least one of the coordinates (i.e. start, stop or one of the start or end exon coordinates) was located on a strand different from all the others or (ii) the chromosome of the position changed during the liftover or (iii) the distance (in nucleotides) between the sORF start and stop codons has changed after the liftover. All the entries for which the liftover failed were removed from the database. Based on the previous assumptions, the liftover failed for 709 ORFs

(377 failed due to the last criteria) in *H. sapiens* and for none of the *M. musculus* entries (Table 3). The choice of such stringent criteria has been strengthened by the fact that these entries (i) only represent <0.05% of the entries for *H. sapiens* and (ii) are more susceptible to be unreliable entries.

### Merge of redundant information

As our database aims to provide a repository of unique identified sORFs of the human and mouse genomes, all the redundant entries describing the same sORFs have been merged. In a first step, we identified all the sORF entries for which all the identification features were provided (chromosome, strand, start position, stop position, splicing status and splicing coordinates). sORFs sharing the same feature values were merged. In a second step, we identified all the remaining entries with only partial identification features provided: the chromosome as well as either (i) both the strand and the start positions or (ii) both the strand and the stop positions or (iii) both the start and the stop positions. Those entries were merged to the best matching fully described entries identified in the first step. If no matching fully described entry was found, then the entries were removed. In order to keep track of the number of times a same sORF has been described in the original data sources, the initial number of entries merged together was registered for each sORF.

During this merging, information regarding the transcripts that harbor the sORFs has been registered too. Hence, when several sORFs were merged into one single entry in MetamORF, the resulting new sORF entry was registered as harbored by all the distinct transcripts related to the original entries. After this removal of redundant information, we were finally able to identify 664 771 and 497 904 unique sORFs for *H. sapiens* and *M. musculus*, respectively (Table 3).

It should be noticed that all unique sORF entries generated at this stage have been kept, including the ones describing ORFs longer than 100 amino acids. Entries describing such ORFs may be either coming from data sources that (i) did not remove the ORFs longer than 100 amino acids or (ii) used a higher threshold or (iii) described the ORF as unspliced while it is actually susceptible to be spliced (and thus has a shorter sequence on the transcript than the one expected).

### Completion of missing information and computation of relative coordinates

In the original data sources, the only information provided (when provided) on the transcripts was the transcript ID. Detailed information was retrieved from Ensembl

2. sORFs identified in human and mouse genomes have been gathered as unique entries in a database: *MetamORF – 2.1*. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis

**Table 2.** Features allowing to characterize the sORFs

Family	Feature	Details
Location	Chromosome	The chromosome or scaffold on which the ORF is located
	Strand	The strand of the sORF
	ORF start	The absolute genomic coordinates of the start codon (position of the first nucleotide)
	ORF stop	The absolute genomic coordinates of the stop codon (position of the third nucleotide)
	Splicing status	Is the sORF spliced?
	Splicing coordinates	The coordinates of the start and end of each exon constituting the sORF
	Transcript	The name or ID of the transcript(s) related to the sORF (eventually with transcript strand, start and end positions and transcript biotype)
	Gene	The name, symbol, alias or ID of the gene(s) related to the sORF (when not intergenic)
Lengths	Length	The length of the sORF (in nucleotides)
	Putative sPEP length	The length of the (putative) sORF-encoded peptide in amino acids
Category	Category	The category to which the sORF belongs (e.g. upstream or downstream)
Sequence signature	Start codon sequence	The nucleic sequence of the sORF start codon
	Nucleic sequence	The nucleic sequence of the sORF
	Amino acid sequence	The amino acid sequence of the (putative) sORF-encoded peptide
Environmental signature	Kozak context	Does a Kozak context has been identified for the sORF start codon?
Conservation	PhyloCSF score	The PhyloCSF score computed for the sORF
	PhastCons score	The PhastCons score computed for the sORF
Coding potential assessment	FLOSS class and score	The FLOSS class and score computed for the sORF
	ORF score	The ORF score computed for the sORF
Biological context	Cell context	The cellular context in which the sORF has been identified or detected

**Table 3.** MetamORF most important statistics

Feature		<i>H. sapiens</i>	<i>M. musculus</i>
Original data sources	ORFs	1 344 978	1 249 176
	Transcripts	101 597	85 653
	Predicted ORFs for which the transcript is unknown	181 122	213 301
	ORFs detected by Ribo-seq for which the transcript is unknown	79 422	8546
	ORFs detected by MS for which the transcript is unknown	54	0
	ORF to transcript associations	3 379 219	2 066 627
	ORFs predicted	202 309	222 705
	ORFs identified by ribosome profiling	1 142 669	1 026 471
	ORFs identified by MS	166	0
ORFs for which the homogenization of genomic coordinates failed	709	0	
MetamORF database	ORFs	664 771	497 904
	Transcripts	90 406	63 147
	Predicted ORFs for which the transcript is unknown	13 440	14 327
	ORFs detected by Ribo-seq for which the transcript is unknown	71 158	2
	ORFs detected by MS for which the transcript is unknown	48	0
	ORF for which the transcripts are unknown	83 403	14 329
	ORF to transcript associations	729 793	696 785
	ORFs predicted	17 027	14 500
	ORFs identified by ribosome profiling	664 771	497 904
ORFs identified by MS	147	0	
Genes harboring at least 1 sORF		23 767	15 869
ORFs having at least one class annotation (short, upstream)		630 953	497 904

MS: mass spectrometry.

## 2. sORFs identified in human and mouse genomes have been gathered as unique entries in a database: MetamORF – 2.1. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis

databases (v90) through their REST API and inserted in our database: (i) the transcript biotype, (ii) the transcript start and end genomic coordinates, (iii) the codon of the canonical CDS (for protein-coding transcripts only) start and stop genomic coordinates and (iv) the full nucleic sequence. In addition, the sequence flanking the start codon (20) has been recovered. As the sORF nucleic and amino acid sequences were not systematically provided by the data sources, these were downloaded from the Ensembl databases using their genomic coordinates.

Moreover, when the transcript ID was available, sORF start and stop relative coordinates have been computed on each of their transcript using AnnotationHub (v2.18.0; 39) and ensemblDb (v2.10.2, <https://bioconductor.org/package/release/bioc/html/ensemldb.html>) R packages (R v3.6.0).

### Standardization of the cell types and ORF classes

#### Cell types

Original data sources do not use a common thesaurus or ontology to name the cell types (e.g. ‘HFF’ and ‘Human Foreskin Fibroblast’) or use non-biological meaning names (e.g. sORFs.org (37) provides the name of the original publication as a cell type). In order to provide a uniform informative naming, we manually recovered the name of the cell line, tissue or organ used in these datasets and defined a unique name to be used in our database for each cell line, tissue or organ, trying to use the most commonly used nomenclature for cell lines (Supplementary Table S3). In addition, in order to ensure interoperability with other biological resources, we recovered the matching ontology terms from the following ontologies when feasible: the Cell Ontology (40), the Cell Line Ontology (41), the BRENDA Tissue Ontology (42), the Human Cell Atlas Ontology (43), the Foundational Model of the Anatomy Ontology (44), the Ontology for Biomedical Investigations (45), the NCI Thesaurus OBO Edition (46), the Experimental Factor Ontology (47), the BioAssay Ontology (48) and the Ontology for MIRNA Target (49), using the Ontology Lookup Service (EBI) (50) (Supplementary Table S4).

#### ORF classification

Despite the use of a common nomenclature by the wide majority of the scientific community to annotate the open reading frames, based on their size and relative position on their transcript (e.g. short, upstream, downstream and overlapping), no clear consensus about the definitions of these categories nor their names has been defined so far (25). In order to homogenize this information in MetamORF, we created a new annotation of the ORFs using the ORF length, transcript biotype, relative positions and reading frame information when available (see Supplementary Methods). In this annotation, a threshold of 100 amino

acids has been used to define the ‘short ORFs’, as this value is the most commonly used for historical reasons (2, 4, 6, 8, 24).

#### Computation of the Kozak contexts

The Kozak motif and context have been regarded as the optimal sequence context to initiate translation in all eukaryotes. We have thus assessed the Kozak context for each sORF, using the criteria defined by Hernández *et al.* (51). Briefly, for each ORF to transcript association, the Kozak context was computed looking for regular expression characterizing an optimal, strong, moderate or weak Kozak context (Supplementary Tables S5 and S6). Kozak-alike contexts were also computed for non-ATG initiated sORFs looking for the same patterns with flexibility regarding nucleotides at +1 to +4 positions.

#### MetamORF software and languages

The pipeline used to build MetamORF has been developed using Python (v2.7) with SQLAlchemy ORM (sqlalchemy.org, v1.3.5). The database has been handled using MySQL (mysql.com, v8.0.16). Docker (docker.com, v18.09.3) and Singularity (singularity.lbl.gov, v2.5.1) environments have been used in order to ensure reproducibility and to facilitate deployment on high-performance clusters.

The MetamORF web interface has been developed using the Laravel (laravel.com, v7.14.1) framework with PHP (v7.3.0), JavaScript 9, HTML 5 and CSS 3. The NGINX (v1.17.10) web server and PHP server (v7.3.0) were deployed with Docker (docker.com, v18.09.3) and Docker-compose (v1.24.0) to ensure stability.

#### Enrichment analysis

##### Gene lists

The list of genes harboring at least one uORF has been collected from MetamORF as a list of Ensembl identifiers using a SQL query.

The list of ATF4 and CHOP targets identified by ChIP-seq comes from Han *et al.* (52) (available as supplementary data on the editor’s website). Genes congruently and translationally upregulated under endoplasmic reticular (ER) stress have been provided by Guan *et al.* (53) (upon request). As these lists of genes were provided as gene symbols, matching Ensembl IDs have been recovered using the g:Convert tool available on the gProfiler web interface (54).

The universe contains all protein-coding genes annotated at least once in Gene Ontology (55, 56) (downloaded from the g:Profiler web interface on 3 November 2020).

##### Statistics

After discarding genes absent in the universe from the lists, the enrichment analysis was performed using



## 2. sORFs identified in human and mouse genomes have been gathered as unique entries in a database: MetamORF – 2.1. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis

an hypergeometric test with R 3.6.0 (<https://www.r-project.org/>). A Benjamini–Hochberg correction has been applied to allow for multiple comparisons, and a False Discovery Rate (FDR) threshold of 0.05 has been considered as significant.

### Database content, accessibility and web interface

#### A new repository of short ORF-related data

MetamORF describes 664 771 and 497 904 unique ORFs in the human and mouse genomes, respectively, providing at least the information necessary to locate the ORF on the genome, its sequence and the gene it is located on (excepted for intergenic ORFs). Extensive information related to the transcripts is provided for 614 997 (~93%) and 497 904 (100%) sORFs for the human and mouse genomes, respectively. These features allowed us to classify 630 953 (~95%) human ORFs and 497 904 (100%) mouse ORFs in at least one class (Table 3, Figure 2, Supplementary Figure S1). Interestingly, it should be noticed that a large proportion (36% and 52% for *H. sapiens* and *M. musculus*, respectively) of ORFs are using an alternative frame to the main CDS. In addition, nearly 23% of the ORFs are located on non-coding RNAs for both species.

#### User-friendly web interface and genome tracks

To provide users with a clear, fast and easy-to-use database, MetamORF can be queried through a user-friendly web interface at <https://metamorf.hb.univ-amu.fr>. A tutorial as well as a documentation page are available online. Briefly, the users may search for sORFs contained in the database starting with a gene symbol (symbol, alias, ID), a transcript ID (ID, name) and an ORF ID or screening a particular genomic area. The data are made accessible through four types of pages: (i) a ‘gene’ page (Figure 3) to allow visualizing information related to all transcripts and sORFs on a gene, (ii) a ‘transcript’ page to allow browsing information related to a transcript gene and all its sORFs, (iii) an ‘ORF’ page to allow fetching information related to all transcripts and gene that harbor the chosen ORF and finally (iv) a ‘locus’ page to allow getting information related to all sORFs located in a particular locus. In addition, the user may also browse across all sORFs related to a species or detected in a particular cell type. It is possible to navigate from one to another page easily to get extensive information about a sORF, a gene or a transcript (Supplementary Figure S2).

In each page, the results can be filtered on (i) the identification method (computational prediction, ribosome

profiling or mass spectrometry), (ii) the start codon, (iii) the Kozak context (as previously defined), (iv) the genomic length (defined as the sum of lengths of each exon constituting the ORF), (v) the transcript biotype (according to the Ensembl definitions), (vi) the ORF annotation (as previously defined) and (vii) the cell type (Supplementary Tables S3 and S4).

All results can be exported in an easily parsable format (comma-separated values file, CSV), as well as in FASTA or BED format.

On ORF, transcript and locus pages, a link allowing the user to easily visualize all the ORFs localized in a particular area on the UCSC Genome Browser (57) is proposed. We also implemented genome track hubs, to allow using UCSC Genome Browser advanced options, such as filtering on ORF categories, transcript biotypes, cell types and transcript IDs.

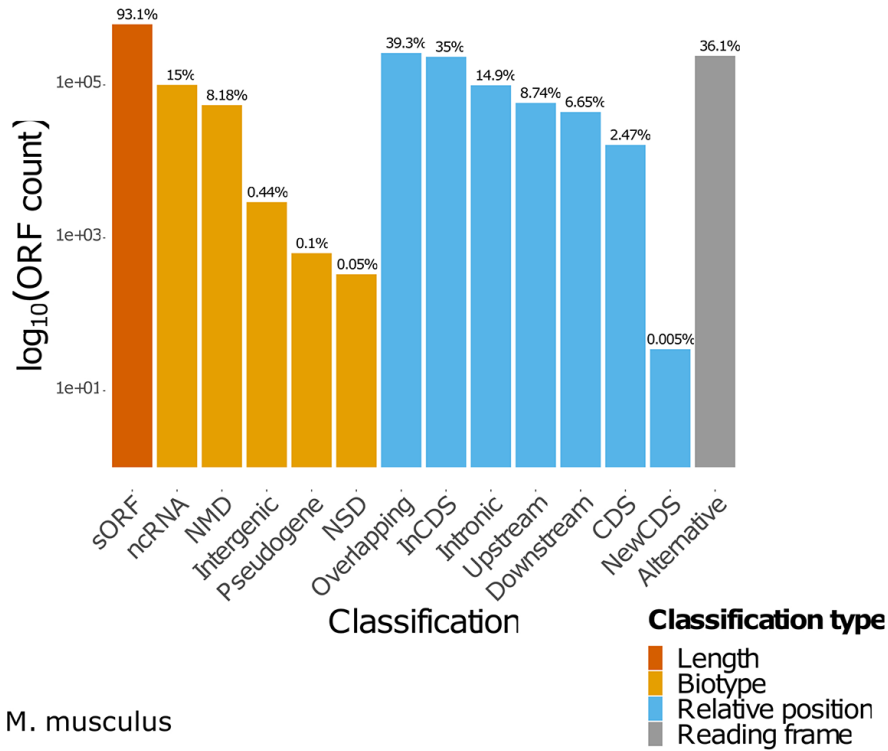
In addition to this user-friendly interface, it is possible to download from the website the content of the full MetamORF database in BED and FASTA formats.

### Using MetamORF to analyze the regulation of integrated stress response

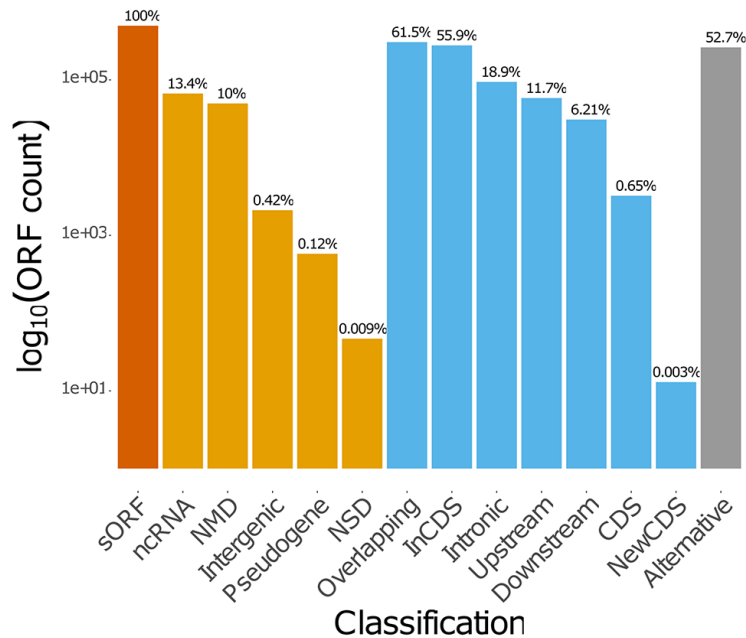
Several studies have reported the role of uORFs in the regulation of the translation during the integrated stress response (ISR) (13, 23, 26, 28). Notably, the mechanism by which the repression of the translation is alleviated under an ER stress has been elucidated for the mammalian transcription factor ATF4, the targets of which are responsible for cell adaptation to stress. Briefly, ATF4 CDS is preceded by two functional uORFs (58), both highly expressed under normal growth and stress conditions. Under the ISR, the small ribosomal subunit is expected to remain bound to the mRNA, scan through the uORF2 and acquire the eIF2•GTP•Met-tRNA<sup>Met</sup> and the large ribosomal subunit in time for initiation at the start codon of the CDS, a phenomenon known as ‘leaky scanning’. In addition, it has been also suggested that the translation of the CDS under stress may result from the ‘re-initiation’, a model in which the large ribosomal subunit and the initiation complex are recruited by the small subunit right after the termination of the translation of the uORF2, allowing thus the initiation at the CDS start codon. Both events are nevertheless technically difficult to distinguish and the exact process remains debated. Hence, assuming the presence of one uORF is sufficient to regulate the translation of the CDS (20), are targets of ATF4 and CHOP (another transcription factor activated upon stress) more likely to harbor uORFs than other genes? Are

2. sORFs identified in human and mouse genomes have been gathered as unique entries in a database: *MetamORF – 2.1*. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis

**A** *H. sapiens*



**B** *M. musculus*



**Figure 2.** Count of ORFs in each class. The bar plots represent the count of ORFs annotated for each class for (A) *H. sapiens* and (B) *M. musculus*. The percentages displayed over the bars indicate the proportion of ORFs annotated in the class over the total number of ORFs registered in the database for the species. NMD: non-sense-mediated decay; NSD: non-stop decay.



## 2. sORFs identified in human and mouse genomes have been gathered as unique entries in a database: MetamORF – 2.1. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis

classes and Kozak contexts, for instance. By homogenizing this information, MetamORF offers the possibility to compare datasets coming from different sources. We noticed that information regarding the Kozak context is missing most of the time, and start flanking sequences are usually not provided. Hence, MetamORF provides a new interesting set of information. It is noteworthy that we discarded 12 of 18 datasets because they lack crucial information regarding their integration into MetamORF. Although this is a rather drastic method, this is performed for the sake of data quality. In these conditions, the confidence in the data and the reliability in the existence of the sORF of interest can be assessed by the number of original experiments that identify the sORF (column 'EXP. COUNT' in the tables of the web interface). It is noteworthy that >97% of the unique ORF entries registered in MetamORF have been identified by at least one experimental method.

It should be noticed that a large amount (~80%) of the sORFs contained in our database have been described in the sORFs.org repository (37). Despite being the most prominent sORF database and offering the community data processed in a normalized way using their own workflow, sORFs.org does not provide metagene analyses (1). In addition, such analysis is made difficult by the absence of gene names and transcriptomic coordinates as well as the high redundancy of information contained in the sORF.org database (37), issues that we addressed with MetamORF. It is noteworthy that another sORF resource, namely OpenProt (59), does not contain ORFs shorter than 30 amino acids, whereas in MetamORF, sORFs of such size represent ~50% of the dataset. Of note, 54% of them have been detected in at least two data sources, therefore reinforcing their probability of existence. Hence, in comparison with existing resources (Supplementary Table S8), MetamORF is complementary and allows analyses at ORF, transcript, gene and locus levels. In addition, it opens the possibility of studying sORFs as a group, at a global scale.

The resource is accessible at <https://metamorf.hb.univ-amu.fr> and provides an intuitive querying interface to enable wet-laboratory researchers to easily question this large set of information. The web interface comes with advanced filters, notably on computed ORF classes, ORF start codons, identification methods, Kozak contexts and cell contexts. Such filters should help end-user biologists without computational skills to identify and collect information about the sORFs important for their topic of interest. Moreover, the implementation of MetamORF content in track hubs allows both quick and advanced visualization of data through the UCSC Genome Browser. Finally, the database content may be exported in various convenient formats widely used by the scientific community (e.g. FASTA and BED).

We believe that MetamORF is of interest not only to bioinformaticians working on short ORFs but also to a wider community, including any biologist who may benefit from knowledge regarding the sORFs located on their gene, transcript or region of interest. As ribosome profiling becomes more appreciated and proteomics starts allowing accurate identification of short peptides, new data describing sORFs in various conditions will be published in the next years, and our database is expected to grow accordingly. In particular, the next release of MetamORF is expected to include data describing the sORFs of other organisms such as *Drosophila melanogaster*. As a conclusion, we believe that MetamORF should help to address new questions in the future, in particular regarding the regulatory functions of the sORFs as well as the functions of the short peptides they may encode.

### Supplementary data

Supplementary data are available at *Database* online.

### Acknowledgement

We thank Andreas Zanzoni for helpful discussions.

### Funding

The project leading to this publication has received funding from the « Investissements d'Avenir » French Government program managed by the French National Research Agency (ANR-16-CONV-0001) and from Excellence Initiative of Aix-Marseille University - A\*MIDEX. It has also received funding from the Excellence Initiative of Aix-Marseille University - A\*Midex a French "Investissements d'Avenir programme" - Institute MarMaRa AMX-19-IET007.

*Conflict of interest.* The authors have no conflict of interest to declare.

### Data availability

Data sources are available on the editor's website or using the links provided in their original publications. The source code used to create the database and the full technical documentation (source code documentation, manual, database structure and dockerfiles) are available on GitHub (<https://github.com/TAGC-NetworkBiology/MetamORF>). Full content of the database can be downloaded in BED and FASTA formats from the MetamORF website, and up-to-date version of track hubs may be downloaded and/or used with your favorite genome browser from the link <https://metamorf.hb.univ-amu.fr/hubDirectory/hub.txt>. The dump of the database is available on request.

### References

1. Martinez, T.F., Chu, Q., Donaldson, C. *et al.* (2019) Accurate annotation of human protein-coding small open reading frames. *Nat. Chem. Biol.*, 16, 458–468

## 2. sORFs identified in human and mouse genomes have been gathered as unique entries in a database: *MetamORF – 2.1*. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis

2. Andrews,S.J. and Rothnagel,J.A. (2014) Emerging evidence for functional peptides encoded by short open reading frames. *Nat. Rev. Genet.*, **15**, 193–204.
3. Aspden,J.L., Eyre-Walker,Y.C., Phillips,R.J. *et al.* (2014) Extensive translation of small open reading frames revealed by Poly-Ribo-Seq. *eLife*, **3**, e03528.
4. Saghatelian,A. and Couso,J.P. (2015) Discovery and characterization of smORF-encoded bioactive polypeptides. *Nat. Chem. Biol.*, **11**, 909–916.
5. 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**, 11663.
6. Ingolia,N.T., Brar,G.A., Stern-Ginossar,N. *et al.* (2014) Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep.*, **8**, 1365–1379.
7. Hao,Y., Zhang,L., Niu,Y. *et al.* (2018) SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. *Brief. Bioinform.*, **19**, 636–643.
8. Pueyo,J.I., Magny,E.G. and Couso,J.P. (2016) New peptides under the s(ORF)ace of the genome. *Trends Biochem. Sci.*, **41**, 665–678.
9. Raj,A., Wang,S.H., Shim,H. *et al.* (2016) Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *eLife*, **5**, e13328.
10. Zanet,J., Chanut-Delalande,H., Plaza,S. *et al.* (2016) Small peptides as newcomers in the control of *Drosophila* development. *Curr. Top. Dev. Biol.*, **117**, 199–219.
11. Mackowiak,S.D., Zauber,H., Bielow,C. *et al.* (2015) Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol.*, **16**, 179.
12. Laumont,C.M., Daouda,T., Laverdure,J.-P. *et al.* (2016) Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nat. Commun.*, **7**, 10238.
13. Couso,J.-P. and Patraquim,P. (2017) Classification and function of small open reading frames. *Nat. Rev. Mol.*, **18**, 575–589.
14. Samandi,S., Roy,A.V., Delcourt,V. *et al.* (2017) Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. *eLife*, **6**, e27860.
15. McGillivray,P., Ault,R., Pawashe,M. *et al.* (2018) A comprehensive catalog of predicted functional upstream open reading frames in humans. *Nucleic Acids Res.*, **46**, 3326–3338.
16. Olexiuk,V., Crappé,J., Verbruggen,S. *et al.* (2016) sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res.*, **44**, D324–D329.
17. Wethmar,K., Barbosa-Silva,A., Andrade-Navarro,M.A. *et al.* (2014) uORFdb—a comprehensive literature database on eukaryotic uORF biology. *Nucleic Acids Res.*, **42**, D60–D67.
18. Johnstone,T.G., Bazzini,A.A. and Giraldez,A.J. (2016) Upstream ORFs are prevalent translational repressors in vertebrates. *EMBO J.*, **35**, 706–723.
19. Wang,H., Yang,L., Wang,Y. *et al.* (2019) RPFdb v2.0: an updated database for genome-wide information of translated mRNA generated from ribosome profiling. *Nucleic Acids Res.*, **47**, D230–D234.
20. Andreev,D.E., Arnold,M., Kiniry,S.J. *et al.* (2018) TASEP modelling provides a parsimonious explanation for the ability of a single uORF to derepress translation during the integrated stress response. *eLife*, **7**, e32563.
21. Fields,A.P., Rodriguez,E.H., Jovanovic,M. *et al.* (2015) A regression-based analysis of ribosome-profiling data reveals a conserved complexity to mammalian translation. *Mol. Cell.*, **60**, 816–827.
22. Erhard,F., Halenius,A., Zimmermann,C. *et al.* (2018) Improved Ribo-seq enables identification of cryptic translation events. *Nat. Methods*, **15**, 363–366.
23. Starck,S.R., Tsai,J.C., Chen,K. *et al.* (2016) Translation from the 5' untranslated region shapes the integrated stress response. *Science*, **351**, aad3867.
24. Crappé,J., Van Crielinge,W. and Menschaert,G. (2014) Little things make big things happen: a summary of micropeptide encoding genes. *EuPA Open Proteom.*, **3**, 128–137.
25. Orr,M.W., Mao,Y., Storz,G. *et al.* (2019) Alternative ORFs and small ORFs: shedding light on the dark proteome. *Nucleic Acids Res.*, **48**, 1029–1042.
26. Plaza,S., Menschaert,G. and Payre,F. (2017) In search of lost small peptides. *Annu. Rev. Cell Dev. Biol.*, **33**, 391–416.
27. Hazarika,R.R., Sostaric,N., Sun,Y. *et al.* (2018) Large-scale docking predicts that sORF-encoded peptides may function through protein-peptide interactions in *Arabidopsis thaliana*. *PLoS One*, **13**, e0205179.
28. Andreev,D.E., O'Connor,P.,B.F., Fahey,C. *et al.* (2015) Translation of 5' leaders is pervasive in genes resistant to eIF2 repression. *eLife*, **4**, e03971.
29. Zanet,J., Benrabah,E., Li,T. *et al.* (2015) Pri sORF peptides induce selective proteasome-mediated protein processing. *Science*, **349**, 1356–1358.
30. Cabrera-Quio,L.E., Herberg,S. and Pauli,A. (2016) Decoding sORF translation – from small proteins to gene regulation. *RNA Biol.*, **13**, 1051–1059.
31. Chen,J., Brunner,A.-D., Cogan,J.Z. *et al.* (2020) Pervasive functional translation of noncanonical human open reading frames. *Science*, **367**, 1140–1146.
32. Rodriguez,C.M., Chun,S.Y., Mills,R.E. *et al.* (2019) Translation of upstream open reading frames in a model of neuronal differentiation. *BMC Genomics*, **20**, 391.
33. Sharipov,R.N., Yevshin,I.S., Kondrakhin,Y.V. *et al.* (2014) RiboSeqDB – a repository of selected human and mouse ribosome footprint and RNA-seq data. *Virtual Biol.*, **1**, 37–46.
34. Evans,V.C., Barker,G., Heesom,K.J. *et al.* (2012) De novo derivation of proteomes from transcriptomes for transcript and protein identification. *Nat. Methods*, **9**, 1207–1211.
35. Liu,W., Xiang,L., Zheng,T. *et al.* (2018) TranslatomeDB: a comprehensive database and cloud-based analysis platform for translatome sequencing data. *Nucleic Acids Res.*, **46**, D206–D212.
36. Lee,S., Liu,B., Lee,S. *et al.* (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, E2424–E2432.
37. Olexiuk,V., Van Crielinge,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.

## 2. sORFs identified in human and mouse genomes have been gathered as unique entries in a database: MetamORF – 2.1. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis

38. Yates, A.D., Achuthan, P., Akanni, W. *et al.* (2020) Ensembl 2020. *Nucleic Acids Res.*, **48**, D682–D688.
39. Hansen, K.D., Sabuncuyan, S., Langmead, B. *et al.* (2014) AnnotationHub: large-scale hypomethylated blocks associated with Epstein-Barr virus-induced B-cell immortalization. *Genome Res.*, **24**, 177–184.
40. Diehl, A.D., Meehan, T.F., Bradford, Y.M. *et al.* (2016) The cell ontology 2016: enhanced content, modularization, and ontology interoperability. *J Biomed Semantics*, **7**, 44.
41. Sarntivijai, S., Lin, Y., Xiang, Z. *et al.* (2014) CLO: the cell line ontology. *J Biomed Semantics*, **5**, 37.
42. Chang, A., Schomburg, I., Placzek, S. *et al.* (2015) BRENDA in 2015: exciting developments in its 25th year of existence. *Nucleic Acids Res.*, **43**, D439–D446.
43. Welter, D., Osumi-Sutherland, D. and Jupp, S. (2018) Human cell atlas ontology. CEUR-WS.org, Vol. 2285, p. 1–2.
44. Golbreich, C., Grosjean, J. and Darmoni, S.J. (2013) The foundational model of anatomy in OWL 2 and its use. *Artif. Intell. Med.*, **57**, 119–132.
45. Bandrowski, A., Brinkman, R., Brochhausen, M. *et al.* (2016) The ontology for biomedical investigations. *PLoS One*, **11**, e0154556.
46. Sioutos, N., de Coronado, S., Haber, M.W. *et al.* (2007) NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J Biomed Inform.*, **40**, 30–43.
47. Malone, J., Holloway, E., Adamusiak, T. *et al.* (2010) Modeling sample variables with an experimental factor ontology. *Bioinformatics*, **26**, 1112–1118.
48. Abeyruwan, S., Vempati, U.D., Küçük-mcginty, H. *et al.* (2014) Evolving BioAssay Ontology (BAO): modularization, integration and applications. *J Biomed Semantics*, **5**, S5.
49. Huang, J., Gutierrez, F., Strachan, H.J. *et al.* (2016) OmniSearch: a semantic search system based on the Ontology for MicroRNA Target (OMIT) for microRNA-target gene interaction data. *J Biomed Semantics*, **7**, 25.
50. Jupp, S., Burdett, T., Malone, J. *et al.* (2015) A new ontology lookup service at EMBL-EBL. *SWAT4LS*, **2**, 118–119.
51. Hernández, G., Osnaya, V.G. and Pérez-Martínez, X. (2019) Conservation and variability of the AUG initiation codon context in eukaryotes. *Trends Biochem. Sci.*, **44**, 1009–1021.
52. Han, J., Back, S.H., Hur, J. *et al.* (2013) ER-stress-induced transcriptional regulation increases protein synthesis leading to cell death. *Nat. Cell Biol.*, **15**, 481–490.
53. Guan, B.-J., van Hoef, V., Jobava, R. *et al.* (2017) A unique ISR program determines cellular responses to chronic stress. *Mol. Cell*, **68**, 885–900.e6.
54. Raudvere, U., Kolberg, L., Kuzmin, I. *et al.* (2019) g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.*, **47**, W191–W198.
55. Ashburner, M., Ball, C.A., Blake, J.A. *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
56. The Gene Ontology Consortium. (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.
57. Kent, W.J., Sugnet, C.W., Furey, T.S. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
58. Vattam, K.M. and Wek, R.C. (2004) Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 11269–11274.
59. Brunet, M.A., Brunelle, M., Lucier, J.-F. *et al.* (2019) OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucleic Acids Res.*, **47**, D403–D410.

2. *sORFs identified in human and mouse genomes have been gathered as unique entries in a database: MetamORF – 2.1. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis*

## **SUPPLEMENTARY MATERIAL**

2. *sORFs identified in human and mouse genomes have been gathered as unique entries in a database: MetamORF – 2.1. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis*

**Supplementary methods**

**ORF classification**

- ORF annotations based on length
  - o sORF (short ORF): The ORF sequence is shorter than 100 amino acids (stop codon and intron excluded).
- ORF annotations based on the transcript biotype
  - o Intergenic: The ORF is located on a 'Long intergenic ncRNA' or 'lincRNA' biotype.
  - o ncRNA: The ORF is located on a 'Non coding', 'ncRNA', 'Processed transcript', 'processed\_transcript', 'Long non-coding RNA', 'lncRNA', '3' overlapping ncRNA', 'Macro lncRNA', 'Long intergenic ncRNA', 'lincRNA', 'miRNA', 'miscRNA', 'piRNA', 'rRNA', 'siRNA', 'snRNA', 'snoRNA', 'tRNA', or 'vaultRNA' biotype.
  - o Pseudogene: The ORF is located on a 'Pseudogene', 'IG pseudogene', 'Polymorphic pseudogene', 'Processed pseudogene', 'processed\_pseudogene', 'Transcribed pseudogene', 'transcribed\_processed\_pseudogene', 'transcribed\_unprocessed\_pseudogene', 'unprocessed\_pseudogene', 'transcribed\_unitary\_pseudogene', 'Translated pseudogene', 'Unitary pseudogene', 'unitary\_pseudogene' or 'Unprocessed pseudogene' biotype.
  - o NMD: The ORF is located on a 'Non sense mediated decay', 'NMD', 'nonsense\_mediated\_decay' or 'non\_stop\_decay' biotype.
  - o Readthrough: The ORF is located on a 'Readthrough' or 'Stop codon readthrough' biotype.
- ORF annotations based on the relative position
  - o Upstream: The start codon of the ORF is located upstream of the CDS start codon and the stop codon of the ORF is located upstream of the CDS stop codon. Note that if both the start and the stop codons are the same for the ORF than the CDS, then the ORF is annotated CDS instead.
  - o Downstream:
    - The start codon of the ORF is located downstream of the CDS start codon and the stop codon of the ORF is located downstream of the CDS stop codon. Note that if both the start and the stop codons are the same for the ORF than the CDS, then the ORF is annotated CDS instead, or
    - The ORF is located on a '3'overlapping ncRNA' biotype.
  - o Overlapping:
    - The start codon of the ORF is located upstream of the CDS start codon and the stop codon of the ORF is located downstream of the CDS start codon and upstream of the CDS stop codon, or
    - The start codon of the ORF is located downstream of the CDS start codon and upstream of the CDS stop codon and the stop codon of the ORF is located downstream of the CDS stop codon or
    - The ORF is located on an 'Antisense' biotype.
  - o Intronic: The ORF is located on a 'retained\_intron', 'sense\_intronic' or 'sense\_overlapping' biotype.
  - o InCDS: The start codon of the ORF is located downstream of the CDS start codon and the stop codon of the ORF is located upstream of the CDS stop codon. Note that if both the start and the stop codons are the same for the ORF than the CDS, then the ORF is annotated CDS instead.
  - o CDS: The start codon of the ORF is the same than the CDS start codon and the stop codon of the ORF is the same than the CDS stop codon.
  - o NewCDS: The start codon of the ORF is located upstream of the CDS start codon and the stop codon of the ORF is located downstream of the CDS stop codon. Note that if both the start and the stop codons are the same for the ORF than the CDS, then the ORF is annotated CDS instead.
- ORF annotations based on the reading frame
  - o Alternative: The ORF start is located on a different frame than the CDS start codon (i.e. the distances in bp between the first nucleotide of the ORF start and the first nucleotide of the CDS start is not a multiple of three).



2. *sORFs identified in human and mouse genomes have been gathered as unique entries in a database: MetamORF – 2.1. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis*

- ORF annotations based on the strand
  - Opposite: The ORF is located on the opposite strand of its transcript.

2. *sORFs identified in human and mouse genomes have been gathered as unique entries in a database: MetamORF – 2.1. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis*

Supp. Table S1 | Data sources description.

Publication	DOI	Database / Source description	Included	Criteria of exclusion if not included
Andreev et al., 2018, eLife (1)	10.7554/eLife.32563		No	ORF start and stop genomic absolute coordinates missing
Rodriguez et al., 2019, BMC Genomics (2)	10.1101/412106		No	ORF start and stop genomic absolute coordinates missing
Sharipov et al., 2014, Virtual Biology (3)	10.12704/vb/e18	RiboSeqDB	No	ORF start and stop genomic absolute coordinates missing
Evans et al., 2012, Nat. Methods (4)	10.1038/nmeth.2227	PITDB	No	ORF start and stop genomic absolute coordinates missing
Chew et al., 2016, Nat. Commun. (5)	10.1038/ncomms11663		No	ORF start and stop genomic absolute coordinates missing
Wethmar et al., 2014, Nucl. Ac. Res. (6)	10.1093/nar/gkt952	uORFdb	No	ORF start and stop genomic absolute coordinates missing
Fields et al., 2015, Mol. Cell (7)	10.1016/j.molcel.2015.11.013		No	Dataset included in sORFs.org
Liu et al., 2018, Nucl. Ac. Res. (8)	10.1093/nar/gkx1034	TranslatomeDB	No	Automated download of numerous files impossible
Wang et al., 2019, Nucl. Ac. Res. (9)	10.1093/nar/gky978	RPFdb	No	Automated download of numerous files impossible

2. *sORFs identified in human and mouse genomes have been gathered as unique entries in a database: MetamORF – 2.1. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis*

Hao et al., 2018, Brief Bioinform. (10)	10.1093/bib/bbx005	smPROT	No	ORF theoretical length different from the provided ORF length for more than 95% of the entries, suggesting unregistered splicing
Lee et al., 2012, Proc. Natl. Acad. Sci. USA (11)	10.1073/pnas.1207846109	TISdb. Files downloaded from “download” section.	No	No information provided about the splicing neither about the ORF length
McGillivray et al., 2018, Nucl. Ac. Res. (12)	10.1093/nar/gky188	Supplementary tables S3, S4, S6-S14	No	No information provided about the splicing neither about the ORF length
Mackowiak et al., 2015, Genome Biol. (13)	10.1186/s13059-015-0742-x	Additional file 2: Table S1. All sORF information for human	Yes	
Erhard et al., 2018, Nat. Meth. (14)	10.1038/nmeth.4631	Supplementary Table 3: Identified ORFs (Union of all ORFs detected either by PRICE, RP-BP or ORF-RATER, or contained in the annotation (Ensembl V75))	Yes	
Johnstone et al., 2016, EMBO (15)	10.15252/emboj.201592759	Dataset EV2: Location and translation data for all analyzed transcripts and ORFs in human	Yes	
Laumont et al., 2016, Nat. Commun. (16)	10.1038/ncomms10238	Supplementary Data 2: List of all cryptic MAPs detected in subject 1. Table presenting the genomic and proteomic features of all cryptic MAPs	Yes	

2. sORFs identified in human and mouse genomes have been gathered as unique entries in a database: MetamORF – 2.1. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis

Samandi et al., 2017, eLife (17)	10.7554/eLife.27860	<i>Homo sapiens</i> alternative protein predictions based on RefSeq GRCh38 (hg38) based on assembly GCF_000001405.26. Release date 01/01/2016 (tsv file). <i>Mus musculus</i> alternative protein predictions based on annotation version GRCm38. Release date 01/01/2016 (tsv file).	Yes
Olexiouk et al., 2018, Nucl. Ac. Res. (18)	10.1093/nar/gkx1130	sORFs.org. Download full database content from their website	Yes

2. sORFs identified in human and mouse genomes have been gathered as unique entries in a database: *MetamORF – 2.1*. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis

**Supp. Table S2 | Date of download of the data sources and cross-references**

File / data source	Date of download
HGNC cross-references for <i>H. sapiens</i>	08/06/2020
NCBI cross-references for <i>M. musculus</i>	08/06/2020
sORFs.org - <i>H. sapiens</i>	08/06/2020
Erhard <i>et al.</i> , 2018 - <i>H. sapiens</i> (14)	04/01/2019
Johnstone <i>et al.</i> , 2016 - <i>H. sapiens</i> (15)	04/01/2019
Laumont <i>et al.</i> , 2016 - <i>H. sapiens</i> (16)	04/01/2019
Mackowiak <i>et al.</i> , 2015 - <i>H. sapiens</i> (13)	20/03/2019
Samandi <i>et al.</i> , 2017 - <i>H. sapiens</i> (17)	04/01/2019
Johnstone <i>et al.</i> , 2016 - <i>M. musculus</i> (15)	04/01/2019
sORFs.org - <i>M. musculus</i>	08/06/2020
Mackowiak <i>et al.</i> , 2015 - <i>M. musculus</i> (13)	20/03/2019
Samandi <i>et al.</i> , 2017 - <i>M. musculus</i> (17)	04/01/2019

2. *sORFs identified in human and mouse genomes have been gathered as unique entries in a database: MetamORF – 2.1. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis*

**Supp. Table S3 | Homogenization of cell types**

<b>Species</b>	<b>Name of the original cell type (as provided by the data source)</b>	<b>Cell type name used in MetamORF</b>
<i>H. sapiens</i>	loayza_puch_2013	
	rooijers_2013	BJ
	ji_BJ_2015	
	B cells	B_cell
	mills_2016	Blood
	gonzalez_2014	Brain
	Human brain tumor	Brain_tumor
	ji_breast_2015	Breast
	loayza_puch_2016	Breast_tumor
	jakobsson_2017	HAP1
	crappe_2014	HCT116
	lee_2012	
	andreev_2015	
	sidrauski_2015	
	liu_2013_HEK	
	liu_HEK_2013	
	ingolia_2012	HEK293
	ingolia_2014	
	calviello_2016	
	iwasaki_2016	
	park_2017	
	zhang_2017	
	eichorn_2014	
	jan_2014	HEK293T
	Primary human foreskin fibroblasts (HFFs)	
	Primary human fibroblast (HFF)	HFF
	rutkowski_2015	
	wang_2015	
	niu_2014	
	yoon_2014	
liu_2013_HeLa	HeLa	
liu_Hela_2013		
stumpf_2013		

2. *sORFs identified in human and mouse genomes have been gathered as unique entries in a database: MetamORF – 2.1. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis*

	park_2016	
	zur_2016	
	shi_2017	
	werner_2015	
	xu_2016	hES
	gawron_2016	Jurkat
	cenik_2015	LCL
	Loayza_Puch_2016	MCF7
	rubio_2014	MDA-MB-231
	wiita_2013	MM1S
	su_2015	Monocyte
	grow_2015	NCCIT
	tanenbaum_2015	
	tirosh_2015	RPE-1
	wein_2014	Skeletal_muscle
	fritsch_2012	
	stern_ginossar_2012	THP-1
	elkon_2015	U2OS
	malecki_2017	Flp-In_T-REx-293
<i>M.</i>	eichorn_3t3_2014	3T3
<i>musculus</i>	jovanovic_2015	
	fields_2015	BMDC
	eichorn_bcell_2014	B_cell
	gonzalez_2014_mmu	
	cho_2015	Brain
	laguesse_2015	
	deklreck_2015	C2C12
	ingolia_2014_mmu	
	Ingolia_2011	E14
	ingolia_2011	
	Mouse gliomal cells	Glioma
	Mouse liver cell	
	eichorn_liver_2014	
	gao_liver_2014	Liver
	gerashchenko_2016	
	janich_2015	
	Mouse Embryonic Fibroblast (MEFs)	
	thoreen_2012	MEF
	lee_2012_mmu	

2. *sORFs identified in human and mouse genomes have been gathered as unique entries in a database: MetamORF – 2.1. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis*

---

gao_mef_2014	
reid_er_2016	
reid_cytosol_2016	
reid_2014	
Mouse Embryonic Stem Cells	MESC
katz_2014	NSC
guo_2010_mmu	Neutrophil
you_2015	R1E
blanco_2016	Skin_tumor
diaz_munoz_2015	Spleen_B_cell
castaneda_2014	Testis
hurt_2013	v6-5

---



2. sORFs identified in human and mouse genomes have been gathered as unique entries in a database: MetamORF – 2.1. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis

Supp. Table S4 | MetamORF cell types and ontologies

MetamORF cell type	Ontology terms*										
	CL	CLO	BTO	HCAO	FMA	OBI	NCIT	EFO	BAO	OMIT	
HCT116			BTO:0001109					EFO:0002824	CLO:0003665	OMIT:0023581	
THP-1			BTO:0001370					EFO:0001253	CLO:0009348		
HEK293		CLO:0001230	BTO:0000007					EFO:0001182		OMIT:0027010	
NCCIT		CLO:0007955	BTO:0004180								
HeLa			BTO:0000567				NCIT:C20226	EFO:0001185	CLO:0003684	OMIT:0007538	
HEK293T			BTO:0002181					EFO:0001184			
Brain	UBERON:0000955	UBERON:0000955	BTO:0000142	UBERON:0000955	FMA:50801	UBERON:0000955	NCIT:C12439	UBERON:0000955	UBERON:0000955	OMIT:0003277	
HFF	CL:1001608	CLO:0000556	BTO:0002245					EFO:0001209	CLO:0007634		
MDA-MB-231			BTO:0000815					EFO:0002779	BAO:0002670		
BJ		CLO:0001980	BTO:0003807					EFO:0005724			
MM1S		CLO:00037203						EFO:0002869	CLO:0009454		
U2OS			BTO:0001938					EFO:0002796	CLO:0007043	OMIT:0019249	
Jurkat		BTO:0000661	BTO:0000661			OBI:1110035					
RPE-1	CL:0002586	BTO:0002334					NCIT:C33470				
Skeletal_muscle	CL:0000188		BTO:0004392				NCIT:C48687				
hES		CLO:00037280	BTO:0001581							OMIT:0001087	
Neutrophil	CL:0000775		BTO:0000130				NCIT:C12533				
v6-5								EFO:0006308			
E14			BTO:0005136					EFO:0007075			
NSC	CL:0000047	CLO:0000051									
MEF	CL:2000042		BTO:0002572				NCIT:C24196	EFO:0004040			

2. sORFs identified in human and mouse genomes have been gathered as unique entries in a database: MetamORF – 2.1. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis

Spleen_B_cell	CL:0000236	BTO:0000776	NCIT:C12474	OMIT:0016 721
3T3	CLO:0001345			OMIT:0016 968
B_cell	CL:0000236	BTO:0000776 CL:0000236	NCIT:C12474	OMIT:0002 778
Liver	EFO:0000887 BTO:0000759 UBERON:0 FMA:63179	UBERON:00 02107	NCIT:C12392 UBERON:000 UBERON:000 UBERON:000	OMIT:0009 182
BMDC	BTC:0003857		NCIT:C15659 1	
Skin_tumor			DOID:3178	
Testis	UBERON:000 UBERON:0001363 UBERON:0 FMA:7210	UBERON:00 00473	NCIT:C12412 EFO:0000984 UBERON:000	OMIT:0014 592
C2C12	0473	BTC:0000165	EFO:0001098 BAO:000270 8	
R1E	CLO:0008700 BTO:0004500		EFO:0002076	OMIT:0037 111
HAP1			EFO:0007598	
Blood	CL:0000081 EFO:0000296 BTO:0000089 CL:0000081 FMA:62844	UBERON:0000178	CL:0000081 UBERON:000	OMIT:0003 133
Monocyte	CL:0000576 CL:0000576 BTO:0000876 CL:0000576 FMA:62864		NCIT:C12547 CL:0000576	
LCL	BTO:0003335		EFO:0005292	
MCF7	BTO:0000093		NCIT:C18096 EFO:0001203 CLO:0007606	OMIT:0028 025
MCF10A	CLO:0007599 BTO:0001939		EFO:0001200	
Flp-In_T-REx-293	CLO:0037238 BTO:0006149			
Brain_tumor	DOID:1319 BTO:0001573		NCIT:C2907 MONDO:000 DOID:1319	OMIT:0003 288
MESC	CL:0002322 CLO:0037317 BTO:0001581	FMA:82841	NCIT:C12935 EFO:0004038	OMIT:0001 088
Glioma	EFO:0000520 BTO:0000526		NCIT:C3059 EFO:0005543 DOID:006010	OMIT:0007 103
Breast	UBERON:000 UBERON:000 BTO:0000149 UBERON:0 FMA:19898	UBERON:00 00310	NCIT:C12971 UBERON:000 UBERON:000	OMIT:0003 296

\* Some ontology terms may refer themselves to external ontologies

2. *sORFs identified in human and mouse genomes have been gathered as unique entries in a database: MetamORF – 2.1. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis*

**Supp. Table S5 | Kozak contexts definitions.** The start codon contains the nucleotides +1 to +3. The same patterns with variation allowed on the nucleotides between +1 to +3 position were used to compute the Kozak contexts of sORFs with alternative start codons.

	-6	-5	-4	-3	-2	-1	+1	+2	+3	+4
Optimal	G	C	C	R	C	C	A	T	G	G
Strong	N	N	N	R	N	N	A	T	G	G
Moderate	N	N	N	R	N	N	A	T	G	A / T / C
or	N	N	N	Y	N	N	A	T	G	G
Weak	N	N	N	Y	N	N	A	T	G	A / T / C

R = A / G (purine), Y = C / T (pyrimidine)

**Supp. Table S6 | Regular expressions corresponding to Kozak contexts.** The Kozak contexts have been computed using the criteria described in the Supp. table S2. To perform this computation, regular expressions have been searched in the sequences flanking the ORF start codons.

Kozak context	Regular expression
Optimal	GCC[AG]CC.{3}G
Strong	.{3}[AG].{2}.{3}G
Moderate	(.{3}[AG].{2}.{3}[ATC] .{3}[CT].{2}.{3}G)
Weak	.{3}[CT].{2}.{3}[ACT]

**Supp. Table S7 | Source of the gene lists used to perform the enrichment analysis**

Gene list	Source	Description
ATF4 targets <sup>1</sup>	Han <i>et al.</i> , 2013, Nat. Cell. Biol.	Table S1: "Supplementary Table S2. List of ATF4 and CHOP target genes that have binding peaks within 3kb from TSS of annotated gene." restricted to ATF4 targets (i.e. genes with "Overlap=Common" or "ATF4_Only")
CHOP targets <sup>2</sup>	Han <i>et al.</i> , 2013, Nat. Cell. Biol.	Table S1: "Supplementary Table S2. List of ATF4 and CHOP target genes that have binding peaks within 3kb from TSS of annotated gene." restricted to ATF4 targets (i.e. genes with "Overlap=Common" or "CHOP_Only")
Genes congruently up-regulated <sup>3</sup>	Guan <i>et al.</i> , 2017, Mol. Cell.	Get upon request - Congruent (Transcriptional and translational) up-regulation at 16h (chronic ER stress)
Genes transitionally up-regulated <sup>4</sup>	Guan <i>et al.</i> , 2017, Mol. Cell.	Get on request - Translational up-regulation at 1h (acute ER stress)

2. *sORFs identified in human and mouse genomes have been gathered as unique entries in a database: MetamORF – 2.1. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis*

Universe	Gene ontology / gProfiler	All protein coding genes with at least one gene ontology annotation have been included in the universe. The lists of GO terms associated with their Ensembl gene IDs have been downloaded using the gProfiler web interface as a gmt file (data sources tab).
----------	---------------------------	---

**Supp. Table S8 | Comparison of MetamORF with existing sORF-related databases**

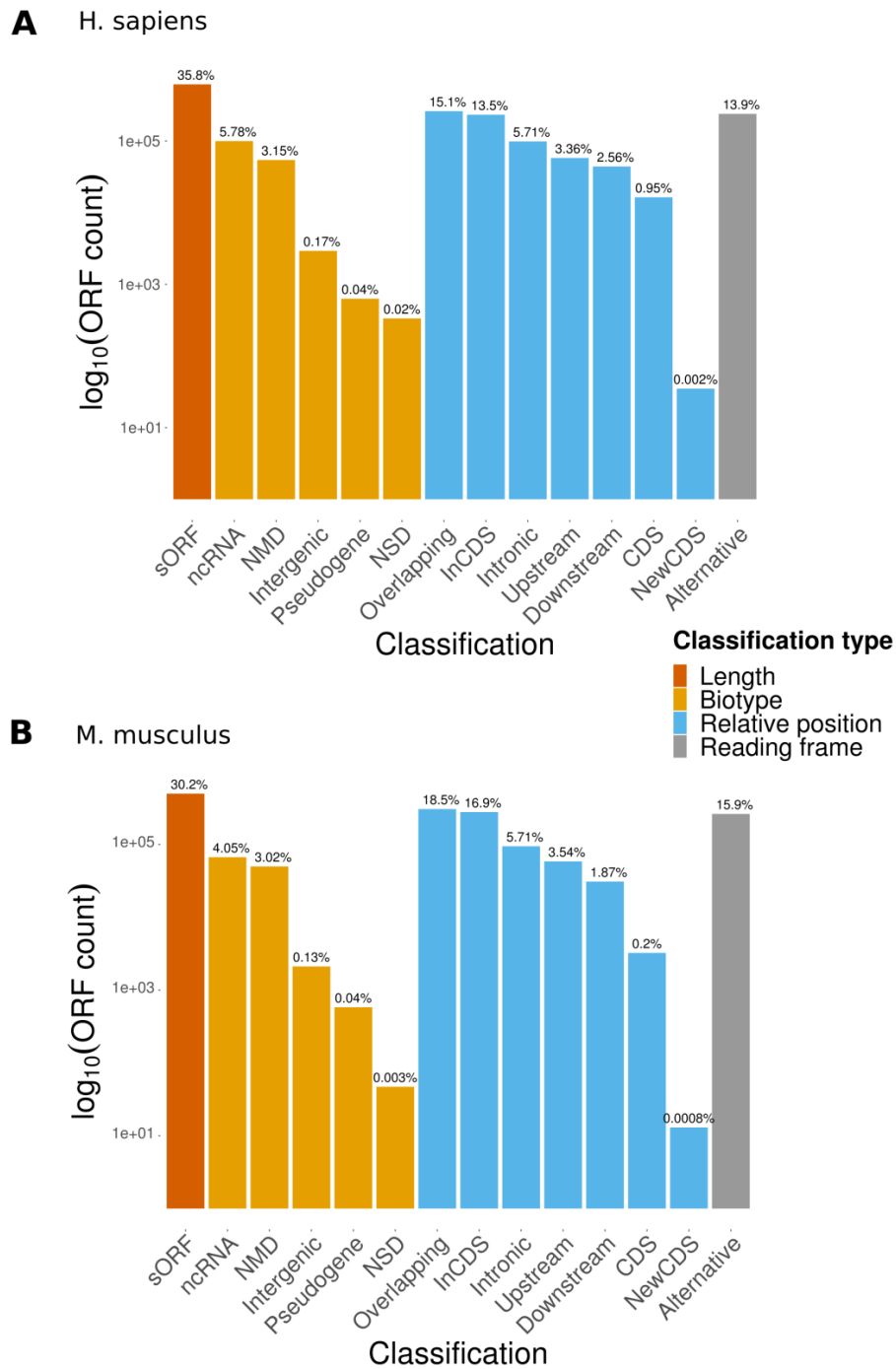
**REFERENCES**

1. Andreev,D.E., Arnold,M., Kiniry,S.J., Loughran,G., Michel,A.M., Rachinskii,D. and Baranov,P.V. (2018) TASEP modelling provides a parsimonious explanation for the ability of a single uORF to derepress translation during the integrated stress response. *eLife*, **7**.
2. Rodriguez,C.M., Chun,S.Y., Mills,R.E. and Todd,P.K. (2019) Translation of upstream open reading frames in a model of neuronal differentiation. *BMC Genomics*, **20**, 391.
3. Sharipov,R.N., Yevshin,I.S., Kondrakhin,Y.V. and Volkova,O.A. (2014) RiboSeqDB – a repository of selected human and mouse ribosome footprint and RNA-seq data. *Virtual Biology*, **1**, 37-46–46.
4. Evans,V.C., Barker,G., Heesom,K.J., Fan,J., Bessant,C. and Matthews,D.A. (2012) De novo derivation of proteomes from transcriptomes for transcript and protein identification. *Nat. Methods*, **9**, 1207–1211.
5. 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**, 11663.
6. Wethmar,K., Barbosa-Silva,A., Andrade-Navarro,M.A. and Leutz,A. (2014) uORFdb--a comprehensive literature database on eukaryotic uORF biology. *Nucleic Acids Res.*, **42**, D60-67.
7. Fields,A.P., Rodriguez,E.H., Jovanovic,M., Stern-Ginossar,N., Haas,B.J., Mertins,P., Raychowdhury,R., Hacohen,N., Carr,S.A., Ingolia,N.T., *et al.* (2015) A regression-based analysis of ribosome-profiling data reveals a conserved complexity to mammalian translation. *Mol. Cell.*, **60**, 816–827.
8. Liu,W., Xiang,L., Zheng,T., Jin,J. and Zhang,G. (2018) TranslatomeDB: a comprehensive database and cloud-based analysis platform for translatome sequencing data. *Nucleic Acids Res.*, **46**, D206–D212.
9. Wang,H., Yang,L., Wang,Y., Chen,L., Li,H. and Xie,Z. (2019) RPFdb v2.0: an updated database for genome-wide information of translated mRNA generated from ribosome profiling. *Nucleic Acids Res.*, **47**, D230–D234.
10. Hao,Y., Zhang,L., Niu,Y., Cai,T., Luo,J., He,S., Zhang,B., Zhang,D., Qin,Y., Yang,F., *et al.* (2018) SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. *Brief. Bioinform.*, **19**, 636–643.

2. *sORFs identified in human and mouse genomes have been gathered as unique entries in a database: MetamORF – 2.1. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis*

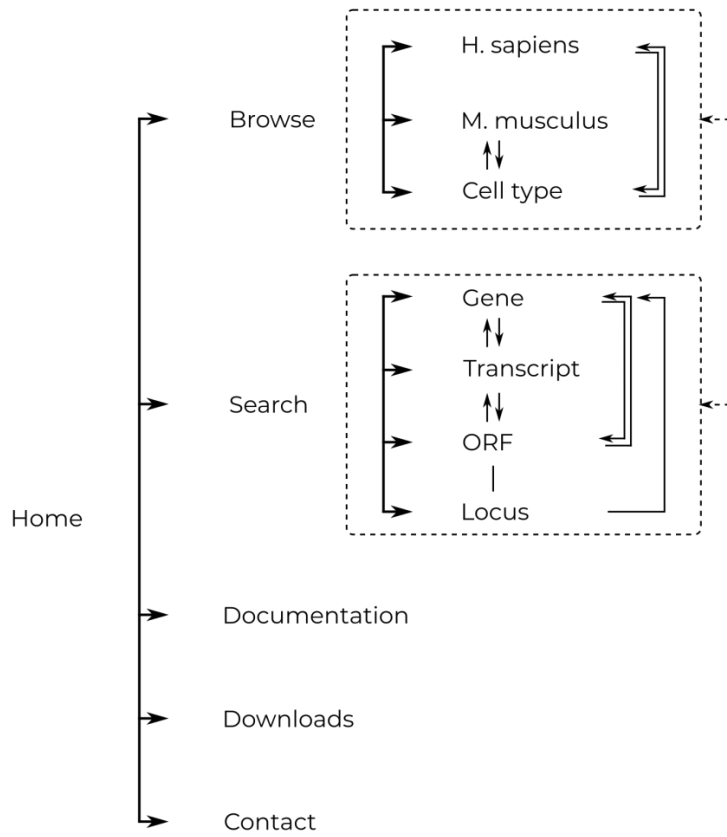
11. Lee, S., Liu, B., Lee, S., Huang, S.-X., Shen, B. and Qian, S.-B. (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, E2424-2432.
12. McGillivray, P., Ault, R., Pawashe, M., Kitchen, R., Balasubramanian, S. and Gerstein, M. (2018) A comprehensive catalog of predicted functional upstream open reading frames in humans. *Nucleic Acids Res.*, **46**, 3326–3338.
13. Mackowiak, S.D., Zauber, H., Bielow, C., Thiel, D., Kutz, K., Calviello, L., Mastrobuoni, G., Rajewsky, N., Kempa, S., Selbach, M., *et al.* (2015) Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol.*, **16**.
14. Erhard, F., Halenius, A., Zimmermann, C., L'Hernault, A., Kowalewski, D.J., Weekes, M.P., Stevanovic, S., Zimmer, R. and Dölken, L. (2018) Improved Ribo-seq enables identification of cryptic translation events. *Nat. Methods*, **15**, 363–366.
15. Johnstone, T.G., Bazzini, A.A. and Giraldez, A.J. (2016) Upstream ORFs are prevalent translational repressors in vertebrates. *EMBO J.*, **35**, 706–723.
16. Laumont, C.M., Daouda, T., Laverdure, J.-P., Bonneil, É., Caron-Lizotte, O., Hardy, M.-P., Granados, D.P., Durette, C., Lemieux, S., Thibault, P., *et al.* (2016) Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nat. Commun.*, **7**, 10238.
17. Samandi, S., Roy, A.V., Delcourt, V., Lucier, J.-F., Gagnon, J., Beaudoin, M.C., Vanderperre, B., Breton, M.-A., Motard, J., Jacques, J.-F., *et al.* (2017) Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. *eLife*, **6**.
18. 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.

2. sORFs identified in human and mouse genomes have been gathered as unique entries in a database: MetamORF – 2.1. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis



**Supp. Figure S1 | Count of ORFs in each class.** The barplot represent the count of ORFs annotated for each class for (A) *H. sapiens* and (B) *M. musculus*. The percentages displayed over the bars indicates the proportion of ORFs annotated in the class over the total number of annotations computed by the MetamORF workflow for the species.

2. sORFs identified in human and mouse genomes have been gathered as unique entries in a database: *MetamORF – 2.1*. A repository of unique, homogenized sORFs was required to get the data necessary to address the questions raised in this thesis



**Supp. Figure S2 | Relational map of MetamORF web interface.**

## **3. sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins**

### **3.1. Studying protein-protein interactions (PPIs) may help characterizing proteins of unknown functions**

#### **3.1.1. Study of sPEP interactions with canonical proteins should provide new insights about their functions**

Functions of **sPEPs** were mainly studied through low scale approaches so far. One possible strategy is to overexpress **sPEPs** in transfected cell lines or whole organism to monitor changes in phenotypes. Knock-down of **sORFs** (based on **clustered regularly interspaced palindromic repeat - Cas9 (CRISPR-Cas9)** technology<sup>1</sup> for instance) can also be performed but are more challenging as phenotypic changes may be due to loss of function or disruption of the transcription harboring the **sORF** [8, 36, 82, 106]. If such approaches have proven to be successful for the characterization of some peptides of unknown functions, they fail at performing large-scale functional characterization and annotation of full proteomes or peptidomes.

---

<sup>1</sup>Clustered regularly interspaced palindromic repeat - Cas9 (CRISPR-Cas9) is a technology allowing genome engineering in animals in plants. It originates from type II CRISPR-Cas systems, which provide bacteria with adaptive immunity to viruses and plasmids. The CRISPR-associated protein Cas9 is an endonuclease that uses a guide sequence within an RNA duplex, tracrRNA:crRNA, to form base pairs with DNA target sequences, enabling Cas9 to introduce a site-specific double-strand break in the DNA. This cost-effective and easy-to-use technology offers the possibility to target any DNA sequence of interest and has become widely used during the past few years. It allows to precisely and efficiently target, edit, modify, regulate, and mark genomic loci of a wide array of cells and organisms [42]. In the particular case of short open reading frames (sORFs), this technology may allow to edit or replace a particular sORF with an unrelated sequence or with a sequence that has specific internal changes. This allows to examine the consequences of a change in nucleotide and/or in amino acids of the product of translation. It may also be used to create fusion peptides by adding a tag to the encoded peptide [8, 82].



### 3. *sPEP* functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.1. Studying protein-protein interactions (PPIs) may help characterizing proteins of unknown functions

However, Plaza *et al.* [102] pointed out the fact that the functions of most *sPEPs* characterized so far mainly rely on specific interactions with larger proteins. In addition, it has been demonstrated that for canonical proteins, cellular processes arise from the dynamic organization of proteins in networks of physical interactions, notably through the dynamic association of individual proteins into complexes and signaling pathways [101, 103, 128]. Saghatelian and Couso [113] as well as Makarewich and Olson [82] estimated also that the importance of *protein-protein interactions* (PPIs) in known *sPEP* functions suggests that the identification of *sPEP-RefProt interactions* (*sPEPRIs*) will be an expedient route for characterizing the molecular functions of uncharacterized *sPEPs*. Hazarika *et al.* [52] were the first (to the best of my knowledge) to propose a computational method for large-scale prediction of *sPEP-RefProt interactions* (*sPEPRIs*). This last consists in screening of *sPEP* binding pockets on the peptide surfaces and predicting models of interactions based on biophysical properties of *sPEPs* and their interactors, and has been successfully applied to investigate the roles of *sPEPs* in *A. thaliana*. Finally, Adai *et al.* [2] stressed out that studying novel proteins as component of networks is an important method of function discovery, providing additional information than individual studies of these proteins in isolation or linear pathways.

In addition, it has been demonstrated that the human interactome is composed of functional network modules, defined as groups of proteins densely connected through their interactions and involved in the same biological processes [21, 22]. This feature has been exploited in the past to annotate proteins of unknown functions [18, 22, 162]. Hence, it is likely for the human *sPEP-RefProt interaction network* (*sPEPRIN*) to be also composed of functional network modules containing *sPEPs* and *RefProts* that are involved in the same biological processes; and we may expect to perform large-scale annotation of *sPEPs* based on this assumption.

#### 3.1.2. Proteins functions can be predicted by studying their interactions with annotated proteins

In the early 2000s, functions of many canonical proteins were still unknown. In 2021, You *et al.* noticed that there were more than 200 million proteins in UniProtKB, while less than 0.1 % of them had experimental GO annotation because of the the high cost of biochemical experiments and the challenge they are rising [52, 158]. Furthermore, these methods usually require to obtain large quantities of proteins and most of them perform better at detecting high-affinity interactions than transient, low-affinity ones [101]. To overcome this issue and characterize the functions of these proteins, many computational methods have been and continue to be developed, including large-scale studies of PPIs [21, 158]. Whilst the first computational methods relied mainly on sequence similarities, more recent ones were based on identification of functional modules after network clustering and the assignment of functions to proteins of unknown function on the basis of the functional annotation of their neighbors [21].

3. *sPEP* functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.1. Studying protein-protein interactions (PPIs) may help characterizing proteins of unknown functions

These methods are based on the observation that proteins of similar cellular functions tend to be close in the interaction graphs [22]. The first methods in this line of idea predicted functions by assigning to proteins the three most frequent cellular functions represented among their direct interaction partners. However this local approach did not consider the graph as a whole but only the immediate protein neighborhood [22]. In addition, it was highlighted that more complex relationships exist, such as between proteins that are part of the same complexes or pathways, and thus involved in the same protein biological processes [22].

Brun et al. [21] hypothesized that the more two proteins share common interactors, the more likely they are to be functionally related. They demonstrated that proteins involved in the same molecular complexes, pathways or cellular processes are clustered, making possible the prediction of cellular functions for uncharacterized proteins. They also emphasize that dense PPIs are the sign of common involvement of proteins in certain biological processes [22]. Based on these observations, they implemented methods performing graph clustering and assigning to the resulting classes biological functions according to the functional annotations of their members following a classical majority rule (*i.e.* the most frequent functions or those shared by more than half of annotated proteins in the class are assigned to proteins of unknown functions in the class). The cellular functions of uncharacterized proteins were then predicted by taking into account the functions assigned to the class and the direct interaction partners of uncharacterized proteins present within the class [22]. The modularity of the network allowed identifying clusters of proteins acting together in particular biological processes using appropriate graph partitioning [18]. As a consequence of these approaches, Becker et al. [18] defined clusters as 'functional modules', *i.e.* groups of proteins involved in the same pathway or the same cellular process.

For instance, such system approach based on PPI clustering have notably been used in the past for the identification of **extreme multifunctional protein (EMF)**, a class of proteins whose multiple functions are very dissimilar to one another and thus involved in unrelated cellular functions [18, 30, 162].

### 3.1.3. Short linear motifs and domains mediate PPIs

Many PPIs are mediated by **short linear motifs (SLiMs)** and domains through **domain-domain interactions (DDIs)** [89] and **domain-SLiM interactions (DMIs)** [39, 70, 102, 162]. Because the presence of particular sequence signatures, including **short linear motifs (SLiMs)** and domains, are a good indicator of a protein's function [117]. Hence, the study of these features in **sPEPs** should provide an insight to their biochemical and molecular functions whilst the study of their interactions with canonical proteins should help identifying the cellular functions and processes in which **sPEPs** are involved.

**SLiMs**, a.k.a. **eukaryotic linear motifs (ELMs)** or motifs are compact disordered mono-partite motifs that constitute an important class of interfaces of interactions

3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.2. A workflow was required to infer sPEP-protein interactions (sPEPRIs) from protein sequences*

[39]. **SLiMs** are short stretches of 3-10 contiguous amino acids residues (6 in average, 1-23 at most) that have been observed in many species and viruses. They are often found in disordered regions and mediate transient **PPIs** with low affinity [39, 146, 149, 162]. **SLiMs** have a lower affinity for their binding partners than globular domains, allowing them to engage in reverse and transient interactions [39]. They can be important regulators of protein function and **PPIs** [81] and have a role in many processes, including cell signaling and the cell cycle for instance [70, 149]. Six class types of **SLiMs** have been identified according to the functional site and are registered in the ELM database: proteolytic cleavage sites (CLVs), post-translational modification sites (MODs), subcellular targeting sites (TRGs), ligand-binding sites (LIGs), docking sites (DOCs) and degradation sites (DEGs) [39, 146]. Mackowiak et al. [81] demonstrated that **sPEPs** tend to be disordered and rich in protein interaction motifs. A disorder analysis with IUPred [41], a software for the prediction of intrinsically unstructured regions of proteins, demonstrated that **sPEPs** are much more disordered than canonical peptides and that conserved **sPEPs** adopt a more stable structure only upon binding to other proteins or nucleic acids.

Andrews and Rothnagel [8] hypothesized that **sPEPs** could also mimic binding domains of interacting proteins. However, **sPEPs** usually do not support typical multidomain structures of canonical proteins because of their short size [36, 102], and they are thus less likely to harbor domains than **SLiMs**. It should also be noticed that the surface area available for stable interaction with other molecules diminishes with size [48], making interaction through **SLiM** more likely than through domains for **sPEPs**.

### **3.2. A workflow was required to infer sPEP-protein interactions (sPEPRIs) from protein sequences**

Because most peptides and proteins can only fulfill their functions through interactions with other proteins, network biology has emerged as a major field of biology. It notably aims at understanding relationships among proteins. The study of **PPIs** has been proved to be of major interest for the characterization of unknown proteins which interactions with annotated ones are known, either from experimental approach or computational predictions. In particular, it is now known that proteins being part of the same complexes and/or involved in similar cellular processes or pathways tends to cluster in networks representing **PPIs**. As previously reported, this knowledge has notably been exploited in the early 2000s to annotate and propose functional characterization of canonical proteins of unknown functions.

To date, only a handful of **sPEPs** have been successfully characterized. Although these ones are involved in many distinct biological functions and lots of putative functions have been proposed for **sPEPs** because of their short size, the actual functions of the wide majority remain to be determined. One objective of my thesis was to

3. *sPEP* functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.2. A workflow was required to infer *sPEP*-protein interactions (*sPEPRIs*) from protein sequences

explore the functions exerted by *sPEPs* and to propose a characterization of these novel peptides. Hence, I decided to take advantage of *PPIs* networks for addressing such objective. However, no large-scale study of *sPEP*-canonical protein interactions (*sPEPRIs*) was available in 2018 for human cells (the only one published was available for *A. thaliana*, and, to the best of my knowledge, there is still no *sPEPRI* network in human available at the time of writing). In addition, acquisition of experimental evidences of *sPEPRIs* can be hard, expensive and would require specific equipment and training, in particular because *sPEPs* seem to have a short half life, be present in low amount, difficult to recover and identify by *MS* and last but not least most experimental methods for the detection or validation of *PPI* are low-scale. Hence, it is unfortunately highly unlikely that large-scale experimental interactomes of *sPEPs* will be released in a near future. To tackle this issue, I decided instead to take advantage of computational approaches to predict the first human *sPEPRI* network.

The pitfall of experimental biology in the identification of large-scale interactomes evoked above also applies to canonical *PPIs*. Although some methods (such as *Y2H*<sup>2</sup>) perform better at identifying interactions for canonical proteins than peptides, they globally remain expensive when compared with *in silico* prediction, slower (considering that prediction tools are available) and could be harder to perform, in particular when it comes to proteins of pathogens whose manipulation is under strict control (such as SARS-CoV-2, Lassa or Marburg viruses for instance). Those reasons, in line with other topics of interest for our laboratories regarding the study of host-pathogen interactions, brought us to develop *mimicINT*, a tool aiming at predicting host-pathogen protein interactions.

*mimicINT* is based on the same paradigm and assumptions as the method used by our team in 2017 for predicting interactions between *F. nucleatum* and human proteins. A. Zanzoni and his colleagues [146, 162] reported that virulence factors often display structures resembling host components in form and function to interact with host proteins, and that pathogen proteins often carry a range of mimics, which resemble structures of the host at the molecular level, what is usually referred as molecular mimicry. Briefly, the method described in 2017 is based on the observation that molecular mimicry occurs in pathogenic viruses and bacteria, as a result of evolution, and allows the emergence of domains and motifs in pathogenic proteins similar to the ones harbored by host proteins. In addition, as the genesis of a rudimentary functional motif necessitates only a handful of mutations, *SLiMs* have a greater propensity to

---

<sup>2</sup>Yeast two-hybrid system (Y2H) is a method that aims at detecting protein-protein interactions developed by Fields and Song in 1989. Briefly, the *cDNA* encoding a protein of interest is cloned into a vector allowing the expression of the protein (the "bait") fused to a transcription factor binding-domain (BD). The *cDNA* encoding another protein of interest is cloned into a vector allowing also the expression of the protein (the "prey") fused to a transcription factor activation domain (AD). If the two expressed proteins interact in the nucleus of the yeast, the proximity of BD and AD allows for activation of the transcription factor of reporter genes under the control of promoters containing sequences bound by BD. By running multiple Y2H experiments in parallel, this technology is now commonly used for large-scale screening of protein-protein interactions [98].

3. *sPEP* functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.2. A workflow was required to infer *sPEP*-protein interactions (*sPEPRIs*) from protein sequences

evolve convergently, and many [SLiMs](#) are indeed conserved across large evolutionary distances [39]. Hence, by looking at host domain and motifs (as well as experimentally validated pathogenic motifs) in pathogens, and under the assumption of this mimicry and that templates of interactions are also valid for host-pathogen interactions, it is possible to infer the full spectrum of theoretically possible [PPIs](#) between a pathogen and its host.

At the beginning of 2020, A. Zanzoni and C. Brun were involved in a project entitled "The role of diet-dependent human microbiome encoded T3SS-dependent effectors in modulating health" (DIME) with the group of P. Falter-Braun (Germany). This project notably planned to predict interactions between human proteins and proteins of (non-pathogenic) multiple bacteria. The sequences of these proteins were collected from metagenomic analyses. Under the observation that commensal bacteria share cell-to-cell but also proteins interactions with their host, and that they evolved together across time, it makes sense to assume that mimicry principle can also apply to non-pathogenic bacteria. This led our team to plan the development of a computational approach that aimed to improve and facilitate the application of the method initially described by A. Zanzoni in 2017.

When COVID-19 pandemic happened, such tool was unfortunately not yet developed. In response to this global threat, our lab started RiPCoN (Rapid interaction profiling of SARS-CoV-2 for network-based deep drug-repurpose learning), an European project led in collaboration with two other teams (P. Falter-Braun, Germany and P. Aloy, Spain). This project notably integrated a work package related to the prediction and analysis of interactions between human and SARS-CoV-2 proteins. Because of the urgency to develop this computational method, and the requirement of similar approach for the DIME project and my own PhD project, I had the opportunity to join this initiative, with the objective to develop *mimicINT* and infer the host-pathogen interactome as fast as possible. The interactomes I predicted in the frame of the RiPCoN and DIME projects are not yet published but should lead to publications in a near future. Because the study of these [protein-protein interaction networks \(PPINs\)](#) is out of the scope of this manuscript, these results will not be discussed here.

For the sake of clarity, I refer to the host proteins (for multi-species interactions) or canonical proteins ([RefProt](#)) as *target* proteins. On the other hand, I refer to bacterial / viral proteins or [sPEPs](#) as *query* proteins when describing the *mimicINT* approach (Fig. 3.1). *mimicINT* relies on the identification of interfaces of interactions (domains and [SLiMs](#)) on both target and query proteins to infer the interactions between them. It uses as input the sequences of the proteins (as a FASTA file) (Fig. 3.1A). Domains are identified on both target and query sequences by using a third-party software (InterProScan [62]) and the domain signatures from the InterPro database [20] (Fig. 3.1B). The detection of [SLiMs](#) on query proteins exploits the motif definitions provided in the ELM database [71] and is performed by another third-party program (SLiMProb, from the SLiMSuite package [43]) (Fig. 3.1C). As [SLiMs](#) are usually located in disorder regions, the disorder propensity of each amino-acid is assessed by the IUPred algorithm



3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.2. A workflow was required to infer sPEP-protein interactions (sPEPRIs) from protein sequences*

and the average disorder propensity score is computed for each sequence. We defined this score as the ratio of the number of residues considered as disordered (according to a threshold previously defined by the user) over the length of the protein. Several versions of mimicINT have been developed and one of them allows to collect target domains directly from the InterPro database instead of performing their detection from the amino acid sequences, in order to consider only occurrences of domains whose functionality has already been demonstrated.

Then, mimicINT infers the interactions between target and query proteins. This inference is performed based on experimentally validated templates of interactions among globular domains (identified based on three-dimensional protein structures, from the 3DID database [89]) as well as between domains and SLiMs (templates of interactions validated in Eukaryotes, from the ELM database [70]) (Fig. 3.1D). mimicINT checks whether any of the query protein contains at least one domain or SLiM for which an interaction template is available. In such case, it infers the interaction between the given protein and all the target proteins containing the cognate domain.

Because they belong to the same species, we may reasonably expect that human sPEPs display interfaces of interactions which resemble structures of the canonical proteins at the molecular level. Thus mimicINT can be used to infer interaction between human sPEPs and RefProts.

Because mimicINT is prone to over-estimation, we also developed two complementary strategies to identify the interactions the most likely to be true positives. As motif-binding domains of the same group (*e.g.* SH3, PDZ) show different interaction specificities, we implemented a strategy (based on published method using HMM [153]) to take into consideration these different specificities. This approach assigns a domain score that can then be used to rank or filter inferred domain-SLiM interactions (DMIs). Because SLiMs are short and degenerate in sequence (*i.e.* there are few fixed amino acid positions for most of them), their detection is also prone to over-prediction. As the level of degeneracy of a SLiM correlates with the stochastically occurring motif count in a proteome, therefore making expectation of random occurrence not equal for all ELM classes, I developed a second approach based on Monte-Carlo simulations. This method is based on published work [49] and aims at assessing the probability of a given SLiM to occur by chance in query sequences. By comparing natural occurrences with occurrences in randomly generated sequences, mimicINT is able to compute p-values that can then be used to filter SLiM occurrences in the query sequences.

Then, SLiM and domain predictions are based on the strong assumption that the mimicry principle applies, *i.e.* that sPEPs are using the same motifs and domain than canonical proteins. This is an important assumption that can be argued, as most of them have never been experimentally detected on sPEPs so far. However, we may easily oppose that functional SLiMs have already been reported experimentally on canonical short peptides [70].

Interaction inferences are also likely to be incomplete due to the limited number of experimentally validated templates of interactions, a recurrent issue in computational

3. *sPEP* functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.2. A workflow was required to infer *sPEP*-protein interactions (*sPEPRIs*) from protein sequences

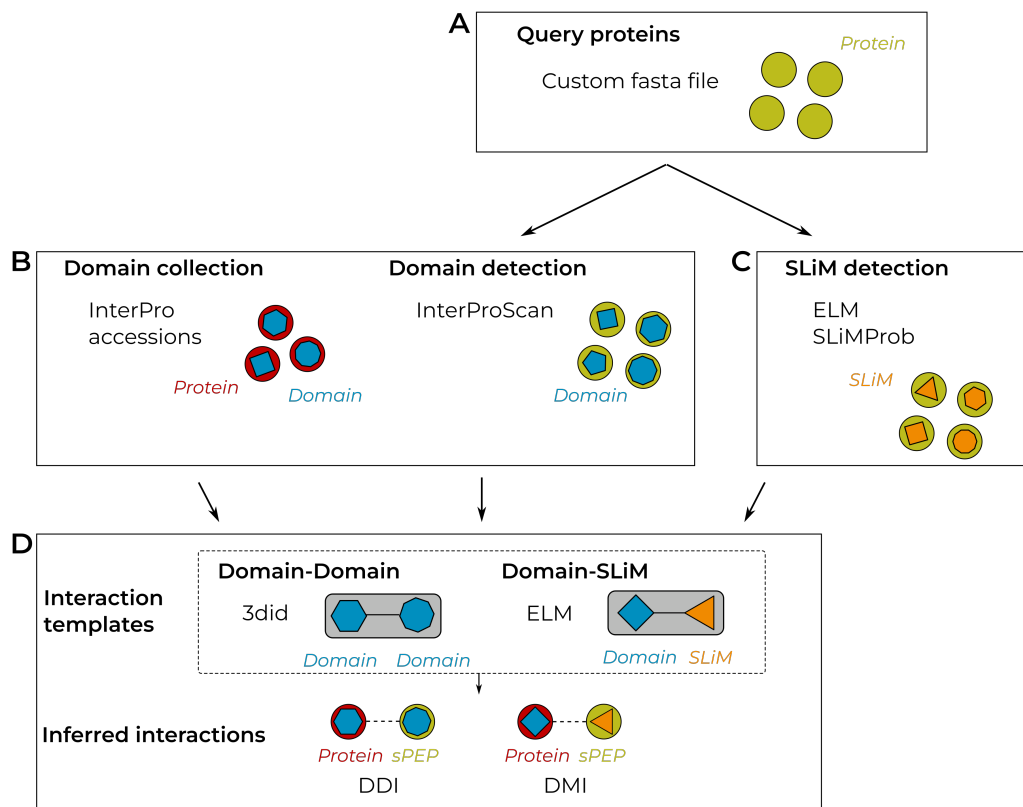


Figure 3.1.: **Overview of the mimicINT workflow.** (A) By providing a fasta file of query protein (bacterial / viral protein or *sPEP*) sequences, mimicINT allows identifying both (B) the domain and (C) SLiM mediated interfaces of interactions. (D) Using publicly available templates of interactions, mimicINT infers the interactions between the query and target proteins (host proteins or canonical proteins (RefProts)).

3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.2. A workflow was required to infer sPEP-protein interactions (sPEPRIs) from protein sequences*

biology. However, due to the limited number of residues that make direct contact with the binding partner in SLiM interfaces [39], mimicINT is more prone to over-prediction than under-prediction. Additionally, many false positive instances will probably not co-occur with their corresponding binding partner due to restricted protein expression as a result of cell compartmentalisation, cell state or tissue specificity [39].

Finally, mimicINT does not take into account the secondary structure of the proteins. sPEP hydrophobicity is not considered whilst it is known that unstructured interfaces have a strong preference for hydrophobic residues [39], a feature that could be integrated in future releases of mimicINT to improve the quality of inferences. Additionally, conformation changes can be induced by peptide binding and potentiate specific interaction [8], a property of proteins that is difficult to integrate in PPIs prediction tools based exclusively on the peptide sequences.

Despite these drawbacks, mimicINT provides the possibility to predict interaction interfaces along with the PPIs (or sPEPRIs) they are able to mediate from the sole amino acid sequence of peptides, a major interest as we are usually missing many information about newly discovered proteins (or sPEPs). In addition, when most of network-based methods totally ignore the protein sequence information [158], mimicINT provides the opportunity to take both into consideration the protein sequences (through the identification of interaction interfaces) and the topology of the PPIN to characterize unknown proteins.

It is important to note that mimicINT includes many steps as well as the use of several third-party softwares. Because of the rapid evolution of operating systems' kernels and software versions, inconsistency or even incompatibility can quickly appear when trying to deploy such complex workflows on new computers. Hence, to ease the deployment and ensure reproducibility and scalability on HPC<sup>3</sup> clusters, I took advantage of Snakemake [87] and containerized environments based on Docker [85] (<https://docs.docker.com/>) and Singularity (<https://github.com/sylabs/singularity>) technologies. Snakemake is a workflow management system based on Python language. It allows easing the development of workflows, and ensuring the automatic execution of a set of predefined rules. In addition, it allows easy execution on HPC clusters through the easy communication with workload managers, such as SLURM (<https://slurm.schedmd.com>). Containerized environments enable to separate applications from the infrastructure on which they run, so it makes software delivery easier.

---

<sup>3</sup>High-performance computing refers to the practice of using massive computational resources in a way that delivers much higher performance than one could get out of a typical workstation. HPC aggregates a great number of processors (CPUs) and allows to manipulate big volumes of data. Such computational resources are now used in many fields (meteorology, ecology, fluid mechanics, astronomy, molecular biology, genetics, finance etc.). A HPC cluster is constituted of thousands of compute servers (called nodes) that are connected together. The nodes work in parallel with each other, and each one can work independently of the others. Such architecture allows for massive parallelization of computations, which may decrease by decades the computational time. As a matter of comparison, I used more than 3 millions of CPU hours in the frame of my thesis, which would roughly correspond to 20+ years of computation on a common bioinformatician's workstation (assuming a workstation is provided 16 CPUs fully available).



3. *sPEP* functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.2. A workflow was required to infer *sPEP*-protein interactions (*sPEPRIs*) from protein sequences

Such technologies were initially developed by and for pure computer scientists, but they are more and more used in the field of bioinformatics. As many other computational biologist, I strongly support and encourage the use of such environments as they help ensuring deployment, portability, reproducibility and scalability on **HPC** clusters. It is to note that all the (bio)informatics tools I developed and analysis I run during my thesis were performed using Docker and/or Singularity. Docker offers the opportunity to provide micro-services, such as web server, but usually requires to have administrator access to the infrastructure it is executed on. On the contrary, Singularity can be used without administrator access, but is not appropriate for the development of servers. It should be highlighted that all the tools previously described are open-source, a practice that tends to be encourage over years in our application field.

Finally, *mimicINT* may also benefit to experimental biologists, as for scientist willing to predict interactions between newly discovered proteins and human proteins or to predict yet unknown interactions between canonical proteins for instance. Unfortunately, the use of *mimicINT* requires advanced computational skills which makes it hard to use for most biologists. For this reason, we decided to build a web server, allowing to run *mimicINT* online, through an user-friendly interface. This task brought some difficulties notably because it requires to be able to manage putative numerous job submissions on the server at the same time. We started to develop this web server in collaboration with three students (M. Cristianini, a Master 2 student, L. Drets and K. Maldonado, two technology degrees graduate students (*DUT*)) I had the opportunity to co-supervise. This task is still under progress and we expect the deployment of this web server to happen in the upcoming months.

**Choteau SA**, Cristianini M, Maldonado K, Drets L, Boujeant M, Brun C, Spinelli L, Zan-zoni A (2022). *mimicINT*: a workflow for microbe-host protein interaction inference. *bioRxiv*, 10.1101/2022.11.04.515250.

3. *sPEP* functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.2. A workflow was required to infer *sPEP*-protein interactions (*sPEPRIs*) from protein sequences

## *mimicINT*: a workflow for microbe-host protein interaction

### inference

Sébastien A. Choteau<sup>1</sup>, Marceau Cristianini<sup>1</sup>, Kevin Maldonado<sup>1</sup>, Lilian Drets<sup>1</sup>,

Mégane Boujeant<sup>1</sup>, Christine Brun<sup>1,2</sup>, Lionel Spinelli<sup>1,\*</sup>, Andreas Zanzoni<sup>1,\*</sup>, #

<sup>1</sup>Aix-Marseille Univ, INSERM, TAGC, UMR\_S1090, Turing Centre for Living Systems, Marseille, France, <sup>2</sup>CNRS, Marseille, France

\* Equal contribution.

# To whom correspondence should be addressed: andreas.zanzoni@univ-amu.fr.

### Abstract

The increasing incidence of emerging infectious diseases is posing serious global threats. Therefore, there is a clear need for developing computational methods that can assist and speed-up experimental research to better characterize the molecular mechanisms of microbial infections. In this context, we developed *mimicINT*, a freely available computational workflow for large-scale protein-protein interaction inference between microbe and human by detecting putative molecular mimicry elements that can mediate the interaction with host proteins: short linear motifs (SLiMs) and host-like globular domains. *mimicINT* exploits these putative elements to infer the interaction with human proteins by using known templates of domain-domain and SLiM-domain interaction templates. *mimicINT* provides (i) robust Monte-Carlo simulations to assess the statistical significance of SLiM detection which suffers from

3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.2. A workflow was required to infer sPEP-protein interactions (sPEPRIs) from protein sequences*

false positive, and (ii) interaction specificity filter to account for differences between motif-binding domains of the same family.

*mimicINT* is implemented in Python and R, and it is available at: <https://github.com/TAGC-NetworkBiology/mimicINT>.

## **Introduction**

Most pathogens interact with their hosts to reach an advantageous niche and ensure their successful dissemination. For instance, viruses interfere with important host-cell processes through protein-protein interactions to coordinate their life cycle (Yamauchi and Helenius, 2013). It has been shown that host cell networks subversion by pathogen proteins can be achieved through interface mimicry of endogenous interactions (i.e., interaction between host proteins) (Franzosa and Xia, 2011; Garamszegi *et al.*, 2013). This strategy relies on the presence in pathogen protein sequences of host-like elements, such as globular domains and short linear motifs (SLiMs), that can mediate the interaction with host proteins (Davey *et al.*, 2011; Hagai *et al.*, 2014; Via *et al.*, 2015).

Over the last years, many computational methods have been developed to predict pathogen-host protein interactions, some of which are based on the detection of sequence or structural mimicry elements (Arnold *et al.*, 2012; Nourani *et al.*, 2015). Such approaches allowed, for instance, to suggest potential molecular mechanisms underlying the implication of gastrointestinal bacteria in human cancer (Zanzoni *et al.*, 2017; Guven-Maiorov *et al.*, 2017) or to discriminate between viral strains with different oncogenic potential (Lasso *et al.*, 2019), thus showing that protein-protein interaction predictions can be instrumental in untangling microbe-host disease

3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.2. A workflow was required to infer sPEP-protein interactions (sPEPRIs) from protein sequences*

associations. Nevertheless, the source code of many of these tools are not freely available to the community (e.g., (Becerra *et al.*, 2017; Guven-Maiorov *et al.*, 2017; Lasso *et al.*, 2019)) providing the predictions through a database (e.g., (Lasso *et al.*, 2019)), or can be only used through a web interface (e.g. (Guyen-Maiorov *et al.*, 2020)), thus limiting the prediction reproducibility and tool usability.

In this context, and inspired by our previous work (Zanzoni *et al.*, 2017), we present *mimicINT*, a computational workflow for large-scale interaction inference between microbe and human proteins by detecting host-like elements and using experimentally identified interaction templates (Mosca *et al.*, 2014; Kumar *et al.*, 2020).

## **Implementation**

*mimicINT* detects putative molecular mimicry elements in microbe sequences of interest that can mediate the interaction with host proteins (Figure 1). *mimicINT* is written in Python and R languages and exploits the Snakemake workflow manager for automated execution (Köster and Rahmann, 2018). It consists of four main steps: (i) the detection of host-like elements in microbe sequences; (ii) the collection of domains on the host protein (iii); the interaction inferences between microbe and host proteins; and (iv) the functional enrichment analysis on the list of inferred host interactors.

In the first step, *mimicINT* takes as input the FASTA-formatted sequences of microbe proteins (e.g., viral or other pathogen proteins susceptible to be found at the pathogen-host interface) to detect host-like elements: domains and SLiMs. The domain identification is performed by the stand-alone version of InterProScan (Jones

3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.2. A workflow was required to infer sPEP-protein interactions (sPEPRIs) from protein sequences*

*et al.*, 2014) using the domain signatures from the InterPro database (Blum *et al.*, 2021). By default, *mimicINT* retains InterProScan matches with an *E*-value below  $10^{-5}$ , a threshold value commonly used for detecting profile-based domain signatures in protein sequences in the context of interaction inference (Schleker *et al.*, 2012). The host-like SLiMs detection exploits the motif definitions available in the ELM database (Kumar *et al.*, 2020) and is carried out by the SLiMProb tool from the SLiMSuite software package (Edwards *et al.*, 2020). As SLiMs are usually located in disordered regions (Davey *et al.*, 2012), SLiMProb uses the IUPred algorithm (Dosztányi, 2018) to compute the disorder propensity of each amino acid in the query sequences, and generates an average disorder propensity score for every detected SLiM occurrence. For SLiM detection, the default IUPred disorder propensity threshold is set to 0.2, a value commonly used to limit false negatives (Edwards and Palopoli, 2015; Edwards *et al.*, 2020), and the minimum size of the predicted disorder region is set to 5, the optimal size to detect true positive SLiM occurrences (Paulsen, 2019). Nevertheless, the user can choose all running parameters for the host-like element detection in the *mimicINT* configuration file.

In the second step, *mimicINT* gathered the domain annotations of the host proteins from the InterPro database through a REST API query.

In the third step, *mimicINT* infers the interactions between host and microbe proteins. This analysis takes as input the list of known interactions templates gathered from two resources: (i) the 3did database (Mosca *et al.*, 2014), a collection of domain-domain interactions extracted from three-dimensional protein structures (Rose *et al.*, 2013), and (ii) the ELM database (Kumar *et al.*, 2020) that provides a list of experimentally identified SLiM-domain interactions in Eukaryotes. The inference

3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.2. A workflow was required to infer sPEP-protein interactions (sPEPRIs) from protein sequences*

checks whether any of the microbe proteins contains at least one domain or SLiM for which an interaction template is available. In this case, it infers the interaction between the given protein and all the host proteins containing the cognate domain (*i.e.*, the interacting domain in the template). As motif-binding domains of the same group, like SH3 or PDZ, show different interaction specificities (Gfeller *et al.*, 2011) for the SLiM-domain interaction inference, we have implemented a previously proposed strategy (Weatheritt *et al.*, 2012) to take these differences into account (see Supplementary Methods). This approach assigns a "domain score" that can be used to rank or filter inferred SLiM-domain interactions. Once this step is completed, the inferred interactions are stored in both tab-delimited and JSON files to facilitate the import in other applications, such as Cytoscape (Shannon *et al.*, 2003).

In the final step, in order to identify the host cellular functions potentially targeted by the pathogen proteins, *mimicINT* executes a functional enrichment analysis of host inferred interactors. This analysis statistically assesses the over-representation of functional categories, such as Gene Ontology terms and biological pathways (e.g., KEGG and Reactome), using the g:Profiler R client (Raudvere *et al.*, 2019).

Given the degenerate nature of SLiMs (Davey *et al.*, 2012), their detection is prone to generate false positive occurrences. For this reason, we implemented an optional sub-workflow that, using Monte-Carlo simulations, assesses the probability of a given SLiM to occur by chance in query sequences and, thus, can be used to filter out potential false positives (Hagai *et al.*, 2014) (see Supplementary Methods).

To ease deployment and ensure reproducibility and scalability on high-performance computing infrastructures, *mimicINT* is provided as a containerized application based

3. *sPEP* functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.2. A workflow was required to infer *sPEP*-protein interactions (*sPEPRIs*) from protein sequences

on Docker and Singularity (Merkel, 2014; Kurtzer *et al.*, 2017). *mimicINT* is available at <https://github.com/TAGC-NetworkBiology/mimicINT>.

## Results

We sought to evaluate the ability of *mimicINT* to correctly infer SLiM-domain interactions, as this inference can generate many false positives (Weatheritt *et al.*, 2012), using the default parameters for SLiM detection (see Implementation). To do so, we used as controls two datasets of established motif-mediated interactions (MDI) from the ELM database (Kumar *et al.*, 2020) (see Supplementary Methods): (i) 103 interactions between 87 viral and 44 human proteins (vMDI); (ii) 31 interactions between 16 bacterial and 23 human proteins (bMDI). We were able to correctly infer the majority of these interactions (91 vMDI, true positive rate = 88.3%; 21 bMDI, true positive rate = 67.7%). As the availability of negative SLiM-mediated interaction datasets is very limited (Weatheritt *et al.*, 2012; Idrees *et al.*, 2018; Kumar *et al.*, 2020), we estimated the false positive rate (FPR) by applying *mimicINT* to two sets of randomly generated interactions sets (degree-controlled, vMDI<sub>md</sub> and bMDI<sub>md</sub>, respectively). Thirty-four vMDI<sub>md</sub> and 7 bMDI<sub>md</sub> were inferred as motif-mediated (FPR = 33% and FPR = 23%, respectively). We next annotated the human proteins in the two random sets with domain similarity scores. We kept only interactions for which the domain score was above 0.4 (Weatheritt *et al.*, 2012), thereby reducing the number of random interactions predicted as motif-mediated to 9 (FPR = 8.7%) for vMDI<sub>md</sub> and 2 (FPR = 6.4%) for bMDI<sub>md</sub>. Finally, we tested *mimicINT* on two sets of experimentally verified negative 37 viral-human and 4 bacterial-human protein

3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.2. A workflow was required to infer sPEP-protein interactions (sPEPRIs) from protein sequences*

interactions from the Negatome 2.0 database (Blohm *et al.*, 2014). Only two virus-human interactions (5.4%) were inferred as motif-mediated by *mimicINT*.

In the light of these results, we used *mimicINT* to infer the interactions between human proteins and the Marburg virus (MARV), an emerging infectious agent for which experimental protein interaction data is scarce (23 interactions for VP24 protein in IMEx interaction databases (Orchard *et al.*, 2012)).

We downloaded MARV protein sequences (7 proteins, Proteome ID: UP000180448) from UniprotKB in FASTA format. For domain detection, we considered only InterProScan matches in MARV sequences and ran *mimicINT* with default parameters.

In total, we inferred 11,431 interactions between 7 MARV and 2757 human proteins (see Supplementary Data). The vast majority of the inferred interactions, namely 10,101, are motif-domain interactions (MDI, 7 MARV and 2324 human proteins), and the remaining 1,339 are domain-domain interactions (DDI, 5 MARV and 479 human proteins). Interestingly, we observed an significant enrichment of known targets of other viruses among inferred interactors (1096 human proteins, 39.7% of the total, odds ratio = 1.3, P-value =  $1.8 \times 10^{-8}$ , one-sided Fisher's Exact test) (Orchard *et al.*, 2012): 62 (13% of DDI interactors, odds ratio = 0.2, P-value = 1, one-sided Fisher's Exact test) are involved in 133 inferred DDIs, and 1059 (45% of MDI interactors, odds ratio = 1.3, P-value =  $6.7 \times 10^{-6}$ , one-sided Fisher's Exact test) participated in 4591 inferred MDIs. By setting a stringent cutoff of 0.4 on the domain similarity scores, the number of inferred MDI decreases to 2082 (7 MARV and 597 human proteins), while the proportion of known viral targets among human interactors slightly increases (i.e.,



3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.2. A workflow was required to infer sPEP-protein interactions (sPEPRIs) from protein sequences*

50%, 299 proteins, odds ratio = 1.4, P-value =  $4.7 \times 10^{-5}$ , one-sided Fisher's Exact test).

None of the 23 experimentally identified interactions of the MARV VP24 proteins were identified by *mimicINT*, probably due to the fact that they were detected by an affinity-based purification method (Pichlmair *et al.*, 2012), which is more suited to identify indirect protein associations rather than direct interactions (Snider *et al.*, 2015). However, 17 MARV inferred interactions (17 MDI and 4 DDI) are supported by experimental evidence in the closely related Zaire Ebola Virus (Orchard *et al.*, 2012; Batra *et al.*, 2018).

The functional enrichment analysis performed by *mimicINT* on the full list of inferred host interactors returned a list of 975 enriched annotations at  $FDR < 0.01$  (see Supplementary Data). We next filtered out the functional categories annotating less than 5 or more 500 proteins obtaining a list of 763 enriched annotations (241 GO biological processes, 63 GO Cellular components, 6 CORUM complexes, 130 KEGG and 237 Reactome pathways), which points towards cellular processes and pathways related to viral infection and immune system (see Supplementary Data), thus further reinforcing the biological relevance of the inferred interactions.

## Conclusions

We present *mimicINT*, a computational workflow enabling large-scale interaction inference between microbe and host sequences. Given the increasing frequency of (re-)emerging infectious diseases, *mimicINT* can be instrumental to better understand

3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.2. A workflow was required to infer sPEP-protein interactions (sPEPRIs) from protein sequences*

the molecular details underlying microbial infections and to identify proteins and interactions as candidate points for therapeutic intervention. Although we developed *mimicINT* as a tool to infer protein interactions at the microbe-human interface, the workflow can be used to infer interaction among human proteins as well, or applied to organisms whose proteins bear either domains or SLiMs participating in known interaction templates.

## **Acknowledgments**

The authors thank Paul de Boissier for helping in the early development of the workflow. The authors are also grateful to the members of the DIME project for fruitful scientific discussions and advices. Centre de Calcul Intensif d'Aix-Marseille is acknowledged for granting access to its high performance computing resources.

*Author contributions:* Conceptualization: S.A.C, C.B., L.S. and A.Z. Methodology: S.A.C., L.S. and A.Z. Software: S.A.C., M.C., K.M., L.D., L.S. and A.Z. Formal Analysis: S.A.C. and A.Z. Investigation: S.A.C., M.B. and A.Z. Writing – original draft: S.A.C. and A.Z. Writing – review & editing: C.B., L.S. and A.Z. Visualization: S.A.C. and A.Z. Supervision: C.B., L.S. and A.Z. Project Administration: C.B., L.S. and A.Z. Funding Acquisition: C.B and A.Z.

## **Funding**

This work was supported by: the JPI HDHL-INTIMIC action co-funded by the Agence Nationale de la Recherche [ANR-17-HDIM-0001, DIME]; France 2030, the French Government program managed by the French National Research Agency [ANR-16-CONV-

3. *sPEP* functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.2. A workflow was required to infer *sPEP*-protein interactions (*sPEPRIs*) from protein sequences

0001] and from Excellence Initiative of Aix-Marseille University - A\*MIDEX; and the European Union's Horizon 2020 Research and Innovation Programme [Project ID 101003633, RiPCoN]. SAC received funding from the "Espoirs de la recherche" program managed by the French Fondation pour la Recherche Médicale (FDT202106013072).

## References

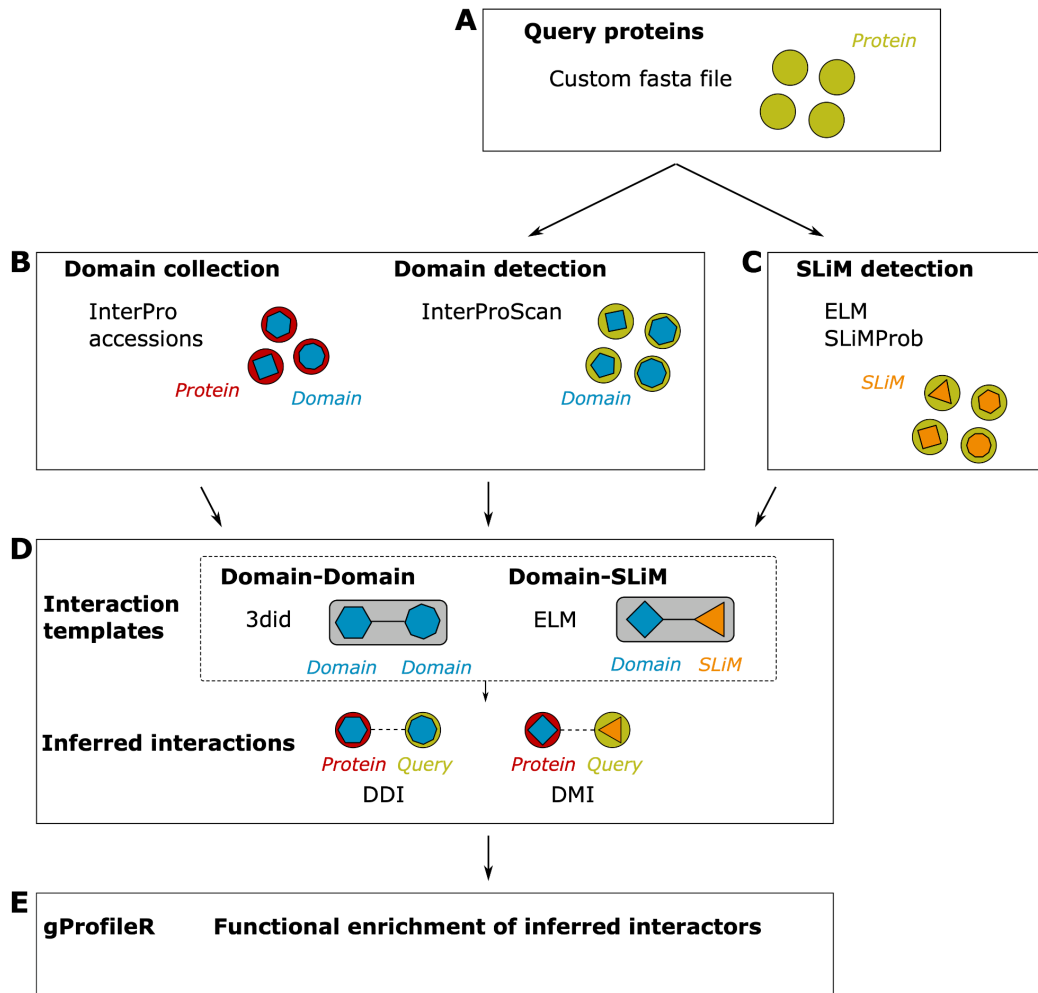
- Arnold,R. *et al.* (2012) Computational analysis of interactomes: current and future perspectives for bioinformatics approaches to model the host-pathogen interaction space. *Methods*, **57**, 508–518.
- Batra,J. *et al.* (2018) Protein Interaction Mapping Identifies RBBP6 as a Negative Regulator of Ebola Virus Replication. *Cell*, **175**, 1917–1930.e13.
- Becerra,A. *et al.* (2017) Prediction of virus-host protein-protein interactions mediated by short linear motifs. *BMC Bioinformatics*, **18**, 163.
- Blohm,P. *et al.* (2014) Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic Acids Res*, **42**, D396–400.
- Blum,M. *et al.* (2021) The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res*, **49**, D344–D354.
- Davey,N.E. *et al.* (2012) Attributes of short linear motifs. *Molecular bioSystems*, **8**, 268–81.
- Davey,N.E. *et al.* (2011) How viruses hijack cell regulation. *Trends in biochemical sciences*, **36**, 159–69.
- Dosztányi,Z. (2018) Prediction of protein disorder based on IUPred. *Protein Sci*, **27**, 331–340.
- Edwards,R.J. *et al.* (2020) Computational Prediction of Disordered Protein Motifs Using SLiMSuite. *Methods Mol Biol*, **2141**, 37–72.
- Edwards,R.J. and Palopoli,N. (2015) Computational prediction of short linear motifs from protein sequences. *Methods Mol. Biol.*, **1268**, 89–141.
- Franzosa,E.A. and Xia,Y. (2011) Structural principles within the human-virus protein-protein interaction network. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 10538–10543.
- Garamszegi,S. *et al.* (2013) Signatures of pleiotropy, economy and convergent evolution in a domain-resolved map of human-virus protein-protein interaction networks. *PLoS Pathog.*, **9**, e1003778.
- Gfeller,D. *et al.* (2011) The multiple-specificity landscape of modular peptide recognition domains. *Mol Syst Biol*, **7**, 484.
- Guyen-Maiorov,E. *et al.* (2020) HMI-PRED: A Web Server for Structural Prediction of Host-Microbe Interactions Based on Interface Mimicry. *J Mol Biol*, **432**, 3395–3403.

3. *sPEP* functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.2. A workflow was required to infer *sPEP*-protein interactions (*sPEPRIs*) from protein sequences

- Guven-Maiorov, E. *et al.* (2017) Prediction of Host-Pathogen Interactions for *Helicobacter pylori* by Interface Mimicry and Implications to Gastric Cancer. *J Mol Biol*, **429**, 3925–3941.
- Hagai, T. *et al.* (2014) Use of host-like peptide motifs in viral proteins is a prevalent strategy in host-virus interactions. *Cell reports*, **7**, 1729–39.
- Idrees, S. *et al.* (2018) SLiM-Enrich: computational assessment of protein-protein interaction data as a source of domain-motif interactions. *PeerJ*, **6**, e5858.
- Jones, P. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.
- Köster, J. and Rahmann, S. (2018) Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, **34**, 3600.
- Kumar, M. *et al.* (2020) ELM—the eukaryotic linear motif resource in 2020. *Nucleic Acids Res*, **48**, D296–D306.
- Kurtzer, G.M. *et al.* (2017) Singularity: Scientific containers for mobility of compute. *PLoS One*, **12**, e0177459.
- Lasso, G. *et al.* (2019) A Structure-Informed Atlas of Human-Virus Interactions. *Cell*, **178**, 1526–1541.e16.
- Merkel, D. (2014) Docker: lightweight linux containers for consistent development and deployment. *Linux journal*, **2014**, 2.
- Mosca, R. *et al.* (2014) 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res*, **42**, D374–379.
- Nourani, E. *et al.* (2015) Computational approaches for prediction of pathogen-host protein-protein interactions. *Front Microbiol*, **6**, 94.
- Orchard, S. *et al.* (2012) Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat. Methods*, **9**, 345–350.
- Paulsen, K. (2019) Optimising intrinsic disorder prediction for short linear motif discovery.
- Pichlmair, A. *et al.* (2012) Viral immune modulators perturb the human molecular network by common and unique strategies. *Nature*, **487**, 486–490.
- Raudvere, U. *et al.* (2019) g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res*, **47**, W191–W198.
- Rose, P.W. *et al.* (2013) The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Res*, **41**, D475–482.
- Schleker, S. *et al.* (2012) Prediction and comparison of *Salmonella*-human and *Salmonella*-*Arabidopsis* interactomes. *Chemistry & biodiversity*, **9**, 991–1018.
- Shannon, P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
- Snider, J. *et al.* (2015) Fundamentals of protein interaction network mapping. *Mol Syst Biol*, **11**, 848.
- Via, A. *et al.* (2015) How pathogens use linear motifs to perturb host cell networks. *Trends Biochem. Sci.*, **40**, 36–48.
- Weatheritt, R.J. *et al.* (2012) The identification of short linear motif-mediated interfaces within the human interactome. *Bioinformatics*, **28**, 976–982.
- Yamauchi, Y. and Helenius, A. (2013) Virus entry at a glance. *J Cell Sci*, **126**, 1289–1295.
- Zanzoni, A. *et al.* (2017) Perturbed human sub-networks by *Fusobacterium nucleatum* candidate virulence proteins. *Microbiome*, **5**, 89.

3. *sPEP* functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.2. A workflow was required to infer *sPEP*-protein interactions (*sPEPRIs*) from protein sequences

## Figures



**Figure 1: Overview of the *mimiciNT* workflow.** By providing a fasta file of protein sequences of the query species (e.g., microbe sequences) (A), *mimiciNT* allows identifying both the domain (B) and SLiM (C) mediated interfaces of interactions. Using publicly available templates of interactions, *mimiciNT* infers the interactions between the proteins of the query and target (i.e., host) species (D). Finally, it provides a list of functional annotations that are significantly enriched in inferred protein targets (E).

3. *sPEP* functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.2. A workflow was required to infer *sPEP*-protein interactions (*sPEPRIs*) from protein sequences

## **SUPPLEMENTARY MATERIAL**

3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.2. A workflow was required to infer sPEP-protein interactions (sPEPRIs) from protein sequences*

## Supplementary Information

*mimic*INT: a workflow for microbe-host protein interaction

inference

Sébastien A. Choteau<sup>1</sup>, Marceau Cristianini<sup>1</sup>, Kevin Maldonado<sup>1</sup>, Lilian Drets<sup>1</sup>,

Mégane Boujeant<sup>1</sup>, Christine Brun<sup>1,2</sup>, Lionel Spinelli<sup>1,\*</sup>, Andreas Zanzoni<sup>1,\*,#</sup>

<sup>1</sup>Aix-Marseille Univ, INSERM, TAGC, UMR\_S1090, Turing Centre for Living Systems, Marseille, France, <sup>2</sup>CNRS, Marseille, France

### Supplementary Methods

**Computation of the motif-binding domain similarity scores.** To identify motif-binding domains that can be specifically associated to a given ELM motif class, we use the same strategy proposed by Weatheritt et al. in 2012 (Weatheritt *et al.*, 2012), which assumes that a domain significantly similar to a known motif-binding domain should also bind the same motif. We first compiled a list of experimentally identified motif binding domains by gathering the original list from Weatheritt et al. complemented by more recent annotations from the ELM database (Kumar *et al.*, 2020) (August 2020). Obsolete ELM class identifiers from Weatheritt et al. were mapped to current ELM identifiers using the "Renamed ELM classes file ([http://elm.eu.org/infos/browse\\_renamed.tsv](http://elm.eu.org/infos/browse_renamed.tsv)) and duplicated domain annotations were removed. In total, we collected 538 domains in 415 human proteins known to bind 212 ELM motif classes (73% of the 290 motif classes present in ELM, August 2020). The sequences of these 415 annotated proteins were fetched from UniprotKB (UniProt Consortium, 2019). We next gathered the sequences of 1452 reference

3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.2. A workflow was required to infer sPEP-protein interactions (sPEPRIs) from protein sequences*

Eukaryota proteomes (22,262,113 protein sequences in total) from UniprotKB (August 2020). We removed redundancy using the CD-HIT algorithm (Fu *et al.*, 2012) to generate a database of 21,414,544 non-identical sequences. We used the GOPHER tool (Davey *et al.*, 2007) from the SLiMSuite package (Edwards *et al.*, 2020) to identify orthologous sequences of the annotated proteins in the database of non-identical eukaryotic sequences by reciprocal BLAST best hits. Selected orthologous proteins were aligned using the multiple sequence alignment algorithm Clustal Omega (v. 1.2.4) (Sievers *et al.*, 2011). Once the position of the motif-binding domain was identified within the alignment, we removed aligned domains with indels covering >10% of the annotated domain sequence. We iteratively realigned the sequences until a set of proteins was identified with <10% indels coverage. In total, we selected 701 multiple sequence alignments that were used as input for generating domain-specific HMM profiles with the `hmmbuild` program from the HMMER package v.3.1.1 (Eddy, 1998). Subsequently, we scanned a representative set of the human proteome (20,350 “reviewed” sequences from UniprotKB) with the domain-specific HMMs using the `hmmsearch` program. We used a *E*-value cutoff of 0.01 to select the best hits and we rejected those hits with a length of <90% of the annotated motif-binding domain sequence length. Finally, the *E*-value of the best-scoring domain was converted into a domain similarity score using the iELM script downloaded from [http://elmint.embl.de/program\\_file/](http://elmint.embl.de/program_file/) (Weatheritt *et al.*, 2012). Doing so, we computed at least one motif-binding domain similarity score for 1,461 human proteins.

**Statistical significance of the SLiMs detected on the microbe sequences.** To assess the probability of a given motif to occur by chance in microbe sequences, we implemented a previously proposed approach (Hagai *et al.*, 2014) to randomly shuffle



3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.2. A workflow was required to infer sPEP-protein interactions (sPEPRIs) from protein sequences*

the disordered regions of each sequence of a microbe of interest to generate a large set of randomized microbe proteins. The number of shuffled sequences to be generated by *mimicINT* can be chosen by the user in the corresponding configuration file (see the *mimicINT* online documentation for more details). By default, *mimicINT* creates a set of 100,000 randomly shuffled proteins, with the assumption that the input sequences belong to the same microbe species or strain. Once the shuffled sequences are generated, the occurrences of each detected motif are compared in each microbe input sequence to the occurrences observed in the corresponding set of shuffled sequences. In order to compute the probability ( $P$ ) of each detected motif to occur by chance, *mimicINT* counts the number of times ( $m$ ) out of the shuffled sequences ( $N$ ) where there are at least the same number of instances of the given motif in the input sequence:

$$P = \frac{m + 1}{N + 1}$$

For example, if a given motif occurs twice in the input sequence, the methods count how many times the same motif is detected at least twice in the corresponding set of randomly shuffled sequences.

**Virus-human and bacteria-human motif-domain interaction datasets.** We gathered two interaction datasets that are known to be mediated by motif-domain interfaces from the ELM database (Kumar *et al.*, 2020) (January 2022). The first consists of 103 interactions between 87 viral and 44 human proteins, and the second consists of 31 interactions between 16 bacterial and 23 human proteins. As the availability of negative motif-mediated interaction datasets is very limited (Weatheritt *et al.*, 2012; Idrees *et al.*, 2018; Kumar *et al.*, 2020), we generated for each dataset a

3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.2. A workflow was required to infer sPEP-protein interactions (sPEPRIs) from protein sequences*

corresponding random set as follows: we shuffled human interactors by randomly sampling a set of proteins (44 for the virus-human dataset and 23 for the bacteria-human dataset) proteins from the list of human proteins annotated with at least one motif-binding domain according to InterPro (i.e, 3940 human proteins in total). We conserved the degree of both viral/bacterial and human proteins. In addition, we fetched a manually curated negative human protein interaction dataset from the Negatome Database 2.0 (manually\_stringent set) (Blohm *et al.*, 2014) extracting 57 virus-human and 4 bacteria-human interactions. We further filtered these sets against the interaction data stored in IMEx consortium databases (4 virus-human interactions removed), and kept only interactions described in research articles only (i.e., we excluded 2 review and 1 conference abstract papers, 16 virus-human interactions removed) in order to have negative data supported by direct experimental evidence in peer-reviewed papers. Doing so, we obtained a list of 37 negative virus-human protein interactions and 4 negative bacteria-human protein interactions.

## Supplementary References

- Blohm,P. *et al.* (2014) Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic Acids Res*, **42**, D396-400.
- Davey,N.E. *et al.* (2007) The SLiMDisc server: short, linear motif discovery in proteins. *Nucleic Acids Res.*, **35**, W455-459.
- Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
- Edwards,R.J. *et al.* (2020) Computational Prediction of Disordered Protein Motifs Using SLiMSuite. *Methods Mol Biol*, **2141**, 37–72.
- Fu,L. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
- Hagai,T. *et al.* (2014) Use of host-like peptide motifs in viral proteins is a prevalent strategy in host-virus interactions. *Cell reports*, **7**, 1729–39.
- Idrees,S. *et al.* (2018) SLiM-Enrich: computational assessment of protein-protein interaction data as a source of domain-motif interactions. *PeerJ*, **6**, e5858.
- Kumar,M. *et al.* (2020) ELM-the eukaryotic linear motif resource in 2020. *Nucleic Acids Res*, **48**, D296–D306.

3. *sPEP* functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.2. A workflow was required to infer *sPEP*-protein interactions (*sPEPRIs*) from protein sequences

Sievers, F. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*, **7**, 539.

UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*, **47**, D506–D515.

Weatheritt, R.J. *et al.* (2012) The identification of short linear motif-mediated interfaces within the human interactome. *Bioinformatics*, **28**, 976–982.

3. *sPEP* functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. *mimicINT* is of major interest to explore the human *sPEP*-ome

### 3.3. *mimicINT* is of major interest to explore the human *sPEP*-ome

I thus used the *mimicINT* workflow to perform a large-scale prediction of *sPEPRIs* in monocytes. This prediction was based on the sequences of 11,404 canonical proteins (*RefProts*, from UniProtKB) expressed in monocytes (according to the Human Protein Atlas) and of 10,475 putative *sPEPs* identified in monocytes by *Ribo-seq* (from Meta-mORF). By running this computation, I present there the first large-scale *sPEP-RefProt interaction network* in human. After filtering the interactions the most likely to be functional (based on domain scores and Monte-Carlo simulations), I finally inferred a total of 250,959 unique interactions (65,508 *domain-domain interactions* (DDIs) and 185,451 *DMIs*). It is to note that when looking at interactions independently of their interfaces (*i.e.* counting only once several interactions involving the same canonical protein and *sPEP*), 154,407 interactions between *RefProts* and *sPEPs* were actually inferred (42,097 *DDIs* and 112,414 *DMIs*). One may highlight that even for a computational biologist, the manipulation of such big interactomes is quite unusual and raised many computational issues. As a matter of comparison, the Human Reference Interactome (HuRI), that registers all experimentally validated interactions among *RefProts* in *H. sapiens*, contains 64,006 interactions (among 6,047 proteins) at the time of writing.

I first performed a descriptive analysis of the interfaces of interactions harbored by *sPEPs* in monocytes. In line with published results, this analysis suggested that most *sPEPs* are involved in key biological functions, including notably regulatory functions and metabolism, immunology responses and cytoskeleton organization. These results were confirmed by looking for *GO*<sup>4</sup> enrichment in interacting *RefProts*.

I then took advantage of the fact that canonical proteins clustering in canonical *PPIs* have been demonstrated to be involved in similar complexes or pathways. As detailed above, we may reasonably expect that *sPEPs* and *RefProts* involved in the same complexes or pathways are clustering together in *sPEPRIs*. I thus performed a systemic annotation of *sPEPs* with *GO* terms, based on graph clustering and assignment of class biological functions according to the functional annotations of their members following a classical majority rule. This large-scale annotation of *sPEPs* with *GO* terms is also the first to be proposed at the time, and it would be valuable to integrate it to the web interface of Meta-mORF in order to make this information more easily accessible to end-user biologists. Interestingly, the wide majority of *sPEPs* (90 % of them) were annotated with metabolic process-related *GO:BP* terms, a result that

---

<sup>4</sup>The Gene Ontology (*GO*) is a resource initially published in 2000 by a consortium (the Gene Ontology consortium) willing to "produce a dynamic, controlled vocabulary that can be applied to all eukaryotes, even as knowledge of gene and protein roles in cells is accumulating and changing". This consortium defined three independent ontologies: biological process (BP), molecular function (MF) and cellular component (CC). Each of these ontologies provides annotations for the (partially characterized) genes and proteins. *GO* is one of the resources the most commonly used by biologist that keeps to be updated [9, 134].

3. *sPEP* functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. *mimicINT* is of major interest to explore the human *sPEP*-ome

highlights the importance of these peptides in the regulation of the cell metabolism. This is particularly interesting as *sORFs* (especially *uORFs*) have been reported to impair the translation under stress, and we may legitimately wonder the exact nature of the relationship between *sORF cis* and *trans* roles.

Because they are located on the same transcripts, and thus transcribed at the same time, I wondered if the *sPEP* located on genes annotated with a certain *GO:BP* term were preferentially interacting with the *RefProts* encoded by these exact same genes. I discovered that for 72 % of *GO:BP* terms (in the *GO* generic subset), the *sPEPs* encoded by the genes of these terms were indeed preferentially interacting with *RefProts* of the same term. These terms were notably related to metabolism (protein folding, metabolic processes etc.), cell cycle (mitotic cell cycle), cytoskeleton (cytokinesis, cytoskeleton organization etc.) and immune responses (inflammatory response etc.). Despite further evidence are required, this result support the hypothesis of coordinated transcriptional regulation and the idea that being located on the same *RNAs* for a peptide and a protein functionally related may represent an advantage for the coordination of their expression.

Further evidence, in particular experimental validation of interactions and experimental characterization of *sPEP* functions are clearly required. However, these results stress out the wide array of cellular functions, biological processes and the number of pathways in which *sPEPs* seem to be involved. In addition, some of these processes are clearly crucial to the cell, such as metabolism pathways. Interestingly, *RNA* metabolism is among the processes in which *sPEPs* seem to be involved. This is of particular interest as *sORFs* are already known to act as *cis* regulators of the translation, and these results suggest they may also play an additional role in the regulation of the translation as *trans* regulators. Finally, I am personally confident that *sPEPs* will become a topic of primary interest in the future, in particular in regard to applications in multiple domain, included (but not restricted to) human and veterinary medicine, ecology and agronomy.

It is to note that I also performed large-scale prediction of *sPEPRIs* from the full set of 20,368 *RefProt* and 659,735 putative *sPEPs* respectively registered in UniProtKB and MetaMORF. Whilst *mimicINT* was designed from the beginning to be adaptable for really large fasta files, it is quite unusual and unexpected to perform bioinformatics analyses with such large datasets. This led me to release several versions of *mimicINT* over time in order to make it scalable for huge amount of data. *mimicINT* demonstrated to be efficient for handling unusually large datasets, despite such computation can only be considered on *HPC* clusters and required to use the resources from the intensive computational centre of Aix-Marseille (> 2 millions CPUh consumed). This interactome and most of the tools necessary for its analysis are now available, and we may expect that analysing this network could be of great interest for the discovery of novel *sPEPs* functions or the experimental validation of interactions, despite being run outside of a particular cell context. It is thus likely that one of the future step of C. Brun's lab will be to analyse in details this larger, non cell-specific, interactome.

3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. mimicINT is of major interest to explore the human sPEP-ome*

**Choteau SA, Pierre P, Spinelli L, Zanzoni A, Brun C (2022).** Short open reading frames-encoded peptides in human monocytes are involved in ubiquitous regulatory functions, metabolism and immunology responses. *In preparation.*

3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. mimicINT is of major interest to explore the human sPEP-ome*

Short open reading frames-encoded peptides in human monocytes are involved in ubiquitous regulatory functions, metabolism and immunology responses

Sébastien A. Choteau, Philippe Pierre, Lionel Spinelli, Andreas Zanzoni, Christine Brun

## Abstract

Short Open Reading Frames (sORFs) are ubiquitous genomic elements that have been overlooked for years. Some may encode functional peptides, so-called sORF-encoded peptides (sPEPs). Most of these last have failed to be annotated notably due to their short length (< 100 residues) and the use of alternative start codons (other than AUG). So far, the roles of only few sPEPs have been characterized and sPEPs whose functions have been determined are involved in a wide range of key biological processes (apoptosis, DNA repair, transcriptional regulation, mTOR signaling, antigen presentation, cardiac activity regulation etc.). However, the functions of most sPEPs remains unknown.

In this study, we propose a system approach to determine the functions of sPEPs in monocytes. We first predicted the interactions of sPEPs with canonical proteins (RefProts) and we analyzed the interfaces of interactions as well as the set of RefProts interacting with sPEPs. Based on the topology of the sPEP-canonical protein interaction network, we then predicted the function of the sPEPs. Our results suggest that the majority of sPEPs are involved in key biological functions, including regulatory functions and metabolism, immunology responses and cytoskeleton organization. Finally, we showed that sPEPs preferentially target RefProts involved in the same processes as their cognate RefProt.

Our results suggest that sPEPs may be key regulator of both ubiquitous and specialized functions. They therefore should be of growing interest for the future proteome-wide analyses.

## 1 Introduction

Open reading frames shorter than 100 codons were initially thought to be non functional and discarded in most gene annotation programs with the notion they had no coding potential [1, 12, 30, 40, 44]. More recent studies demonstrated that these sequences, called short open reading frames (sORFs), may actually encode functional peptides [19, 31, 32, 34, 44]. sORF-encoded peptides (sEPs or sPEPs, a.k.a. micropeptides) have notably been described in eukaryotic cells and are encoded by sORFs located on all classes of RNAs (including presumptive ncRNAs) [7, 31, 44]. Because (i) messenger RNAs (mRNAs) are usually considered as monocistronic, (ii) the use of alternative start codons and (iii) their short sizes, sPEPs have been missed for long [6].

However, due to the growing body of evidences that sPEPs are stable within cells and have regulatory functions, the study of this novel class of peptides has intensified [17]. Recent studies have demonstrated sPEPs to be involved in various cellular processes and diseases, notably cell proliferation, signaling, cell growth, death, metabolism or development [44]. It has even

### 3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. mimicINT is of major interest to explore the human sPEP-ome*

been suggested that sPEPs may constitute a new pool of cancer-related peptides that could be targeted by immunotherapy [34]. As an example, Laumont *et al.* identified 168 novel major histocompatibility complex class I (MHC-I)-associated peptides that were derived from sORFs [22], demonstrating that sPEPs can also be involved in specialized functions such as antigen presentation.

Human monocytes are an heterogeneous population of innate immune cells that may differentiate into macrophages and play a major role in the initiation of immune responses. They are able to express molecules of the MHC-I and MHC-II, which make them of particular interest as numerous sPEPs have been determined to be able to fixate the MHC-I. Indeed, they may be presented as self-antigens with high predicted binding affinities [7, 16, 22]. Additionally, because the presentation of peptides by MHC molecules is largely independent of the amino acid sequence, and many sPEPs may not need proteosomal degradation before entering the MHC-I presentation pathway, a certain fraction of sPEPs is likely to be involved in immunological functions [7, 22].

We recently gathered 664,771 unique sORFs in the full human genome among which 10,475 have been identified to be transcribed in monocytes according to ribosome profiling experiments [10]. Although for most of them there is no strong insight about their actual translation into functional sPEPs, it has been suggested that a sizable fraction of sORFs are translated [19]. Hence, sPEPs could constitute a major pool of functional peptides overlooked so far.

Whilst some methods (such as proteogenomics) succeed at identifying large pool of peptides, there is currently a lack of experimental method leading to the systematic determination of the functions of novel peptides. Consequently, the functions of most sPEPs are currently unknown and to our knowledge, no systematic annotation of sPEPs has been performed so far. To overcome this obstacle, we propose here to study the interactions of the sPEPs with the canonical proteins (designated as RefProts hereafter), for which the functions are known and functional annotations are available. Indeed, protein-protein interactions (PPIs) drive biological functions [23] and it has been demonstrated that protein functions can be assigned on the basis of the annotation of their neighbors in the PPI network [5]. Hence, we hypothesize that analyzing the interactions between sPEPs and RefProts will allow performing a systematic functional annotation of the sPEPs. Nonetheless, as highlighted earlier by Gray *et al.* [17], sPEPs are typically missing from large-scale protein localization and interaction studies [17]. As we recently developed mimicINT [11], a computational method that allows inferring protein-protein interactions (PPIs) based on the presence of short linear motifs and globular domains in amino acid sequences, we herein predicted interactions between sPEPs and RefProts using this method, integrated those predicted interactions in the human interactome and investigate network modules and topology to predict sPEP functions. We then asked whether sPEPs do participate to specific functions in monocytes? Are there processes in which sPEPs are preferentially involved? Are sPEPs preferentially interacting with RefProts involved in the same processes as the RefProts located on the same transcript as their coding sequence, suggesting a regulatory function?

To that extent, we decided to (i) identify the SLiMs and domains with the highest occurrences as sPEP interaction interfaces, to assess the biological processes to which sPEPs are participating; (ii) check the most common sPEP annotations, by annotating sPEP from network clustering followed by function assignment based on a majority rule; and (iii) determine for some GO terms whether there is an enrichment of RefProts involved in the GO term among the RefProts interacting with sPEPs encoded by genes annotated with the GO term.



3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. mimicINT is of major interest to explore the human sPEP-ome*

## 2 Results and discussion

### 2.1 250,959 sPEP-RefProt interactions have been inferred in monocytes

This study presents here the first large-scale network of sPEP-RefProt interactions in human, to the best of our knowledge. In this paper, we aim at exploring sPEP functions by studying these interactions. In particular, interfaces of interactions (domains and SLiMs) provide information about the molecular and biochemical functions of the proteins that harbor them. Indeed, interfaces may mediate interactions with other proteins that notably allow them to take part in particular complexes or pathways, to be addressed to certain subcellular compartments or to be submitted to post-translational modifications. Hence, our first goal (i) was to study which interfaces are the most commonly used by sPEPs to predict their putative functions. Then, proteins involved in the same complexes or metabolic pathways are known to cluster in the canonical PPI network, and the topology of the PPI has been successfully exploited in the past to perform assignment of cellular functions to uncharacterized proteins. Consequently, our second objective (ii) was to take advantage of the sPEP-RefProt interaction (sPEPRI) network topology to predict sPEP functions based on clustering and using a classical majority annotation rule. Finally, a growing community of scientists hypothesized that the co-expression of sPEPs and RefProt from the same transcript could facilitate the integration of sPEPs in cellular pathways related to the main protein product [26, 36]. Hence, our last objective (iii) was to determine for all Gene Ontology (GO) terms of the generic GO subset whether RefProts interacting with the sPEPs encoded by the genes these terms annotate tend to be themselves annotated with this particular GO term.

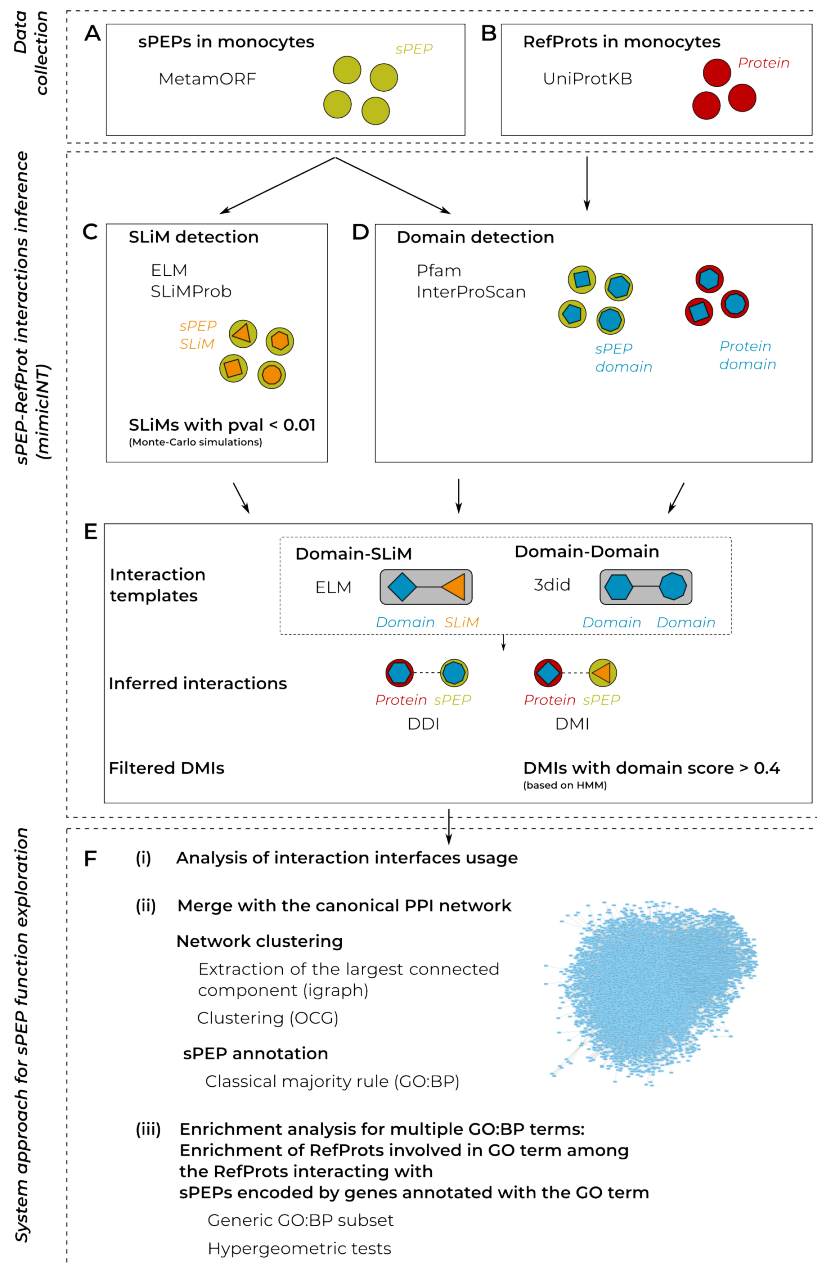
The amino acid sequences of 10,475 putative sORF-encoded peptides (sPEPs) identified by ribosome profiling in monocytes were collected from MetamORF, a repository of unique sORFs identified by computational and experimental methods we previously developed (Fig. 1A). 11,404 canonical proteins (RefProts) have been identified in monocytes at the protein level according to the Human Protein Atlas and their amino acid sequences were downloaded from the UniProtKB database (Fig. 1B). Finally, 250,959 interactions and 154,407 binary interactions (*i.e.* interactions involving distinct partners) between 4,393 sPEPs and 3,981 RefProts in monocytes were predicted with mimicINT, a computational method we previously implemented to infer protein-protein interactions from their sequences (Fig. 1C-E, Table 2). Overall, 41% of the sPEPs (4,393 / 10,475) are predicted to interact with 35% of the RefProt (3,981 / 11,404) within monocytes. Additionally, mimicINT prediction is based upon the detection of interfaces of interactions on sPEPs and RefProts and it identified a total of 17,101 (398 domains and 16,703 SLiMs) and 21,258 interfaces (21,258 domains) respectively on sPEPs and RefProts (Table 1).

Table 1: **Counts of inferred sPEP-RefProt interactions**

Type of interaction	Total interactions	Binary interactions
Domain-domain interactions (DDIs)	65,508	42,097
Domain-SLiM interactions (DMIs)	185,451	112,414
All interactions	250,959	154,407

**NB:** Two interactions involving the same couple of sPEP and RefProt interactors but mediated through two different set of interfaces are counted as two in the count of total interactions whilst counted as one in the number of binary interactions.

3. *sPEP* functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. *mimicINT* is of major interest to explore the human *sPEP*-ome



**Figure 1: Method of inference of the sPEP-RefProt interactions (sPEPRIs).** (A) sPEPs sequences identified in monocytes have been collected from MetamORF and (B) RefProts sequences expressed in monocytes have been collected from UniProtKB. (C-E) sPEP-RefProt interactions have been inferred using the *mimicINT* workflow. Briefly, (C) Short linear motifs (SLiMs) occurrences have been detected on sPEPs (using SLiMProb and data from ELM) and filtered based on pvalues computed by Monte-Carlo simulations; (D) Pfam signatures of globular domains have been detected on RefProts and sPEPs (using InterProScan) and; (E) templates of domain-domain interactions (DDIs, from 3DID) as well as of domain-SLiM interactions (DMIs, from ELM) were used to infer DDIs and DMIs between sPEPs and RefProts. DMIs were then filtered based on domain scores computed by looking for Hidden Markov Models. (F) A system approach has finally been used to explore the functions to which sPEPs participate.

3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. mimicINT is of major interest to explore the human sPEP-ome*

**2.2 The SLiMs and domains mediating interactions in sPEPs are related to ubiquitous regulatory functions, metabolism, immunology processes and cytoskeleton**

**2.2.1 97% of interaction interfaces on sPEPs are SLiMs**

First, we checked what interfaces of interaction are more often harbored by sPEPs, as they can provide an insight about the putative functions of the peptides. When inferring the interactions between sPEPs and RefProts in monocytes, mimicINT detects two types of protein interfaces for PPIs: short linear motifs (SLiMs), that are short stretches of amino acids mainly located in disordered regions [21], as well as globular domains. Globular domains are constituted in median of 88 amino acids (ranging from 4 to 6,907 residues in *H. sapiens*, according to data from the InterPro database [4]) whilst the SLiMs are short stretches of 1 to 23 residues (6.3 in average, according to [13]) preferentially located in highly disordered regions [13]. As by definition sPEPs are short (< 100 residues), we hypothesized sPEPs are more likely to harbor SLiMs than domains. Therefore, mimicINT appears to be suited to predict sPEP interactions since SLiMs are expected to be their favorite interface of interactions.

Table 2: **Counts of domain and SLiM occurrences in sPEPs and RefProts**

		sPEPs	RefProts
Counts in monocytes		10,475	11,404
Interacting		4,393	3,981
Domains	#Occ.	398	21,258
	#Occ. interacting	336	7,267
	#Pfam domains	120	5,475
	#Pfam domains interacting	104	676
SLiMs	#Occ.	16,703	-
	#Occ. interacting	12,949	-
	#ELM classes	60	-
	#ELM classes interacting	43	-

We found that sPEPs harbor respectively 16,703 and 398 distinct occurrences of SLiMs and domains, which confirms that sPEPs harbor more SLiMs than domains (Table 1).

In order to check if the presence of SLiMs and domain was indeed related to the size and disorder level of the sPEPs, we also compared the lengths of the sPEPs based upon their presence. As expected, the sPEPs harboring domains are the longest, whilst the length of the sPEPs seems not to be related with the presence of SLiMs (Fig. S2), which could be easily explained by the fact that long sequences can also present disordered regions harboring SLiMs. We noticed that the general disorder propensity scores are the highest for sPEPs harboring SLiMs (median—average score: 1—0.86; all sPEPs: 0.53—0.57; sPEPs with domains: 0.25—0.33), a result that was expected as the SLiMs were identified in disordered regions.

In accordance with the presence of multiple SLiMs on sPEPs when compared to domains, we were expecting the sPEP-RefProt interactions to be mainly mediated by SLiMs. Indeed, 74% (185,451/250,959) of predicted interactions are domain-SLiM interactions (Table 1) and the SLiM occurrences constituted 97% (12,949/13,285) of the interfaces of interactions harbored by sPEPs (Table 2).

Interestingly, the 12,949 SLiM occurrences mediating at least one interaction on sPEPs belonged to 43 classes of SLiMs (ELM classes) over the 60 ones with a p-value lower than 0.01. In addition, the 336 occurrences of domains mediating at least one interaction on sPEPs belonged

3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. mimicINT is of major interest to explore the human sPEP-ome*

to 104 classes of domains (Pfam signatures) over the 17,929 families registered in Pfam. On the other hand, 3,754 SLiM occurrences, belonging to 17 ELM classes, on sPEPs do not mediate any interaction. In addition, the 62 occurrences of domains of sPEP not mediating any interaction belonged to 16 classes of domains (Pfam signatures). Because of the diversity of interfaces harbored by sPEPs (147 distinct classes of interfaces), we suggest that looking at the most common ones could provide insights about the molecular functions to which sPEPs participate.

**2.2.2 Most commonly used domains are related to metabolism, immunology responses and cytoskeleton**

We thus next investigated the classes of domains used by the sPEPs to interact with RefProts. As previously stated, 26% (65,508/250,959) of interactions are DDIs and 3% (336/13,285) of the interfaces of interactions on sPEPs are domains. In accordance with this relatively low number of interactions mediated by domains, 92% (4,058/4,393) sPEPs are not harboring any domain able to mediate an interaction with RefProts.

We first only considered the most represented domains responsible of interaction(s) with RefProt and selected the 10 most commonly used domains and noticed these domains are mainly related to the cytoskeleton, immunology responses and metabolism (amino acid degradation notably) (Table 3 and Table S1).

Table 3: **Top 10 domains (Pfam) based on occurrence counts in sPEPs**

Pfam accession	Domain name	Function family	#Occ. <sup>a</sup>	#Interactions <sup>b</sup>
PF00038	Intermediate filament protein	Cytoskeleton	15	285
PF00048	Small cytokines (intercrine/chemokine), IL-8 like	Immunology	14	2,100
PF00112	Papain family cysteine protease	AA degradation	13	4,797
PF01231	Indoleamine 2,3-dioxygenase	AA degradation	10	20
PF04699	ARP2/3 complex, 16 kDa subunit (p16-Arc)	Actin	9	1,458
PF09286	Pro-kumamolisin, activation domain	Peptide cleavage	9	45
PF00113	Enolase, C-terminal TIM barrel domain	Glycolysis	8	56
PF12146	Serine aminopeptidase, S33	AA degradation	8	208
PF00129	Class I Histocompatibility antigen, domains alpha 1 and 2	Immunology	7	1,253
PF00262	Calreticulin family	Ca <sup>2+</sup> regulation	7	91
PF00340	Interleukin-1 / 18	Immunology	7	2,303
PF02841	Guanylate-binding protein, C-terminal domain	Immunology	7	91
PF04045	ARP2/3 complex, 34 kDa subunit (p34-Arc)	Actin	7	133
PF07654	Immunoglobulin C1-set domain	Immunology	7	5,859

<sup>a</sup>: Number of occurrences of the domain in sPEPs

<sup>b</sup>: Number of sPEPRIs mediated by the domain

We then considered all the domains mediating at least one interaction with a RefProt and mapped their Pfam accessions to GO terms (Table S1). The figure 2 presents the visualization of the GO biological process (GO:BP) terms using the REVIGO software. This result

3. *sPEP* functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. *mimicINT* is of major interest to explore the human *sPEP*-ome

shows that the domains used as interaction interfaces by the *sPEPs* are mainly related with immunology responses (inflammatory response, antigen processing and presentation, regulation of TOR signaling etc.), cytoskeleton (regulation of actin filament polymerization), as well as ubiquitous regulatory functions (regulation of apoptotic process, protein folding, translation etc.) and metabolic processes (proteolysis, ubiquitin-dependent catabolic process, pentose-phosphate shunt etc.), underlining the possible role of the *sPEPs* in those processes.



Figure 2: **Summarized visualization of the GO:BP terms mapped to the Pfam accessions mediating at least one interaction with a RefProt.** The size of the boxes are related with the number of GO:BP terms aggregated with the GO term shown.

We finally considered the domains that do not mediate any interaction and noticed these domains are mainly related to immunology responses (cytokine production, antigen processing and presentation, scavenger and T-cell receptors, inflammatory responses etc.), cytoskeleton and transport (Table 4 and S2).

All together, these results suggest *sPEPs* may be involved in many biological processes, some of them which are crucial for the cell (metabolism) or specifically related to the functions of monocytes (notably to immunology responses, the regulation of apoptosis and the cytoskeleton). These findings are in line with our current knowledge of *sPEPs* that are known to be notably involved in cell proliferation, signaling, growth, cell death, metabolism and development, cytoskeleton organization and antigen presentation in eukaryotes [12, 16, 22, 34, 44].

### 2.2.3 Most commonly used SLiMs are related to housekeeping regulatory functions

As SLiMs are preferred interfaces of interactions when compared with domains, we also investigated the ELM classes preferentially used by the *sPEPs* to interact with RefProts. As stated

3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. mimicINT is of major interest to explore the human sPEP-ome*

Table 4: **Top 10 domains (Pfam) not mediating interactions, based on occurrence counts in sPEPs**

Pfam accession	Domain name	Function family	#Occ. <sup>a</sup>	#Interactions <sup>b</sup>
PF00335	Tetraspanin family	Membrane proteins	15	0
PF05038	Cytochrome b558 alpha-subunit	Oxidative phosphorylation (phagocytes)	8	0
PF03821	Golgi 4-transmembrane spanning transporter	Transport	7	0
PF09307	CLIP, MHC2 interacting	Immunology (MHC-II)	5	0
PF02394	Interleukin-1 propeptide	Immunology (Cytokines)	4	0
PF03227	Gamma interferon inducible lysosomal thiol reductase (GILT)	Immunology (scavenger receptors)	4	0
PF05283	Multi-glycosylated core protein 24 (MGC-24), sialomucin	Immunology (Cytokines)	4	0
PF03836	RasGAP C-terminus	Cytoskeleton	3	0
PF02535	ZIP Zinc transporter	Transport	2	0
PF07946	Protein of unknown function (DUF1682)	ER biology	2	0
PF10601	LITAF-like zinc ribbon domain	Immunology (pathogen sensing)	2	0
PF11029	DAZ associated protein 2 (DAZAP2)	-	2	0

<sup>a</sup>: Number of occurrences of the domain in sPEPs

<sup>b</sup>: Number of sPEPRIs mediated by the domain

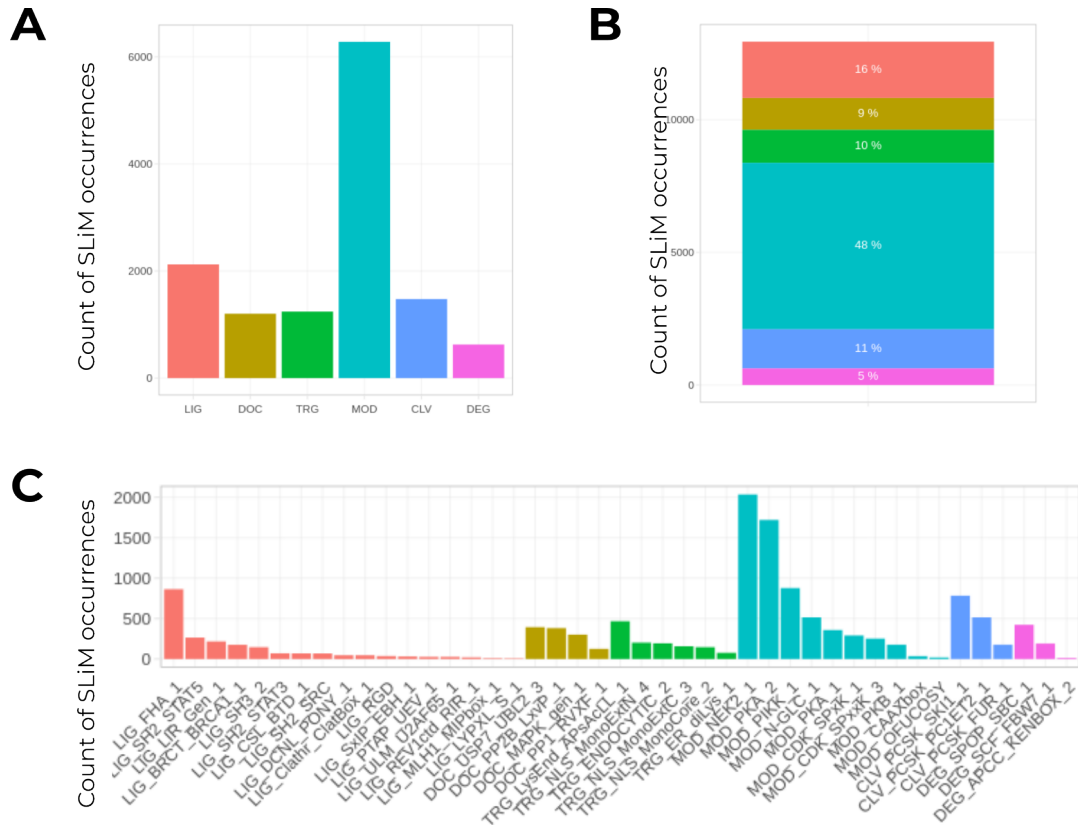
above, 74% (185,451/250,959) of interactions are DMIs and 97% (12,949/13,285) of the interfaces of interactions on sPEPs are SLiMs. As expected regarding this high number of interactions mediated by SLiMs, 95% (4,162/4,393) of the sPEPs are harboring at least one SLiM able to mediate interactions with RefProts. More precisely, 68% of the sPEPs harbor between 1 and 3 SLiMs mediating interactions.

The ELM database classifies the SLiM classes into six distinct types: ligand-binding sites (LIGs), docking sites (DOCs), subcellular targeting sites (TRGs), post-translational modification sites (MODs), proteolytic cleavage sites (CLVs) and degradation sites (DEGs).

Classes of interacting SLiMs belonging to the MOD class are the most commonly harbored by sPEPs (48%) whilst the DEG, known to favor the degradation of the protein that harbors it, are the less commonly used (5%) (Fig. 3). The LIG (16%), CLV (11%), TRG (10%) and DOC (9%) class types of SLiMs have similar numbers of occurrences. Many distinct SLiM classes are constituting the occurrences of LIG and MOD motifs, a result that was expected regarding the fact that the LIG and MOD class types respectively gather 47% (26/60) and 22% (13/60) of the SLiM classes.

Like for the domains, we hypothesized that the type of SLiMs the most commonly used to mediate interactions with RefProts were likely to provide insight about the biological processes in which the sPEPs are involved. Thus, we first only considered the most represented SLiMs

3. *sPEP* functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. *mimicINT* is of major interest to explore the human *sPEP*-ome



3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. mimicINT is of major interest to explore the human sPEP-ome*

responsible of interaction(s) with RefProt and selected the 10 most commonly used SLiM classes. These classes are involved in many biological processes, in particular in cell cycle regulation, DNA repair, metabolism and protein metabolism (Table 5 and Table S3).

Table 5: **Top 10 SLiM classes based on occurrence counts in sPEPs**

SLiM class	Site name (from ELM)	Function family	Pattern prob.	#Occ. <sup>a</sup>	#Interactions <sup>b</sup>
MOD_NEK2.1	NEK2 phosphorylation site	Cell cycle	0.0097983	2035	20350
MOD_PKA.2	PKA Phosphorylation site	Metabolism	0.0094575	1720	51600
MOD_PIKK.1	PIKK phosphorylation site	DNA repair	0.0092301	877	3508
LIG_FHA.1	FHA phosphopeptide ligands	Cell cycle, DNA repair	0.0086622	863	1726
CLV_PCSK_SKI1.1	PCSK cleavage site	Proteolytic processing of peptides	0.0068205	783	783
MOD_N-GLC.1	N-glycosylation site	Translation	0.0050178	515	515
CLV_PCSK_PC1ET2.1	PCSK cleavage site	Metabolism	0.0039028	515	1545
TRG-NLS_MonoExtN.4	NLS classical Nuclear Localization Signals	Nuclear localization signal	0.0012764	467	17746
DEG_SPOP_SBC.1	SPOP SBC docking motif	Cell cycle, protein degradation	0.000938	423	846
DOC_USP7_UBL2.3	USP7 binding motif	Cell survival, response to viral infections	0.0037418	394	394

<sup>a</sup>: Number of occurrences of the SLiMs in sPEPs

<sup>b</sup>: Number of sPEPRIs mediated by the SLiMs

On the other hand, classes of SLiMs not interacting belonging to the TRG class are the most commonly harbored by sPEPs (53%) whilst the DEG and DOC are the less commonly used (2%) (Fig. 4). The MOD (19%) and LIG (17%) class types of SLiMs have similar numbers of occurrences whilst CLV (6%) are less represented. As expected, more distinct SLiM classes are constituting the occurrences of LIG and MOD motifs than the other classes.

Considering the most represented SLiMs not used as interface of interaction with RefProt and selecting the 10 most commonly used SLiM classes, we noticed that these classes are involved in various biological processes, in particular in cell cycle regulation, cell survival and protein degradation (Table 6 and Table S4). It should be noticed that apart from LIG\_KEPE.2, all ELM classes are involved in exactly one template of interaction. This implies that the lack of interaction computed is not biased by the lack of experimental evidence of interaction for those ELMs, but instead by the absence of interacting interfaces on canonical proteins.

The analysis of the SLiM usage comforts the hypothesis that sPEPs are involved in diverse biological processes, some of them being ubiquitous processes, such as cell cycle and metabolism regulations.

#### 2.2.4 RefProts interacting with sPEPs are involved in metabolism, immunology responses and cytoskeleton

The results previously presented are based upon the assumption that the processes in which are involved the sPEPs are dictated by the domains and SLiMs mediating interaction with RefProts. In order to strengthen this hypothesis, we next investigated the biological process in which the RefProts interacting with sPEPs are involved. As the functions of RefProts are relatively quite well known (when compared to functions of sPEPs), we wondered in which biological processes are involved the RefProts interacting with sPEPs in monocytes. To address this question, we



3. *sPEP* functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. *mimicINT* is of major interest to explore the human *sPEP*-ome

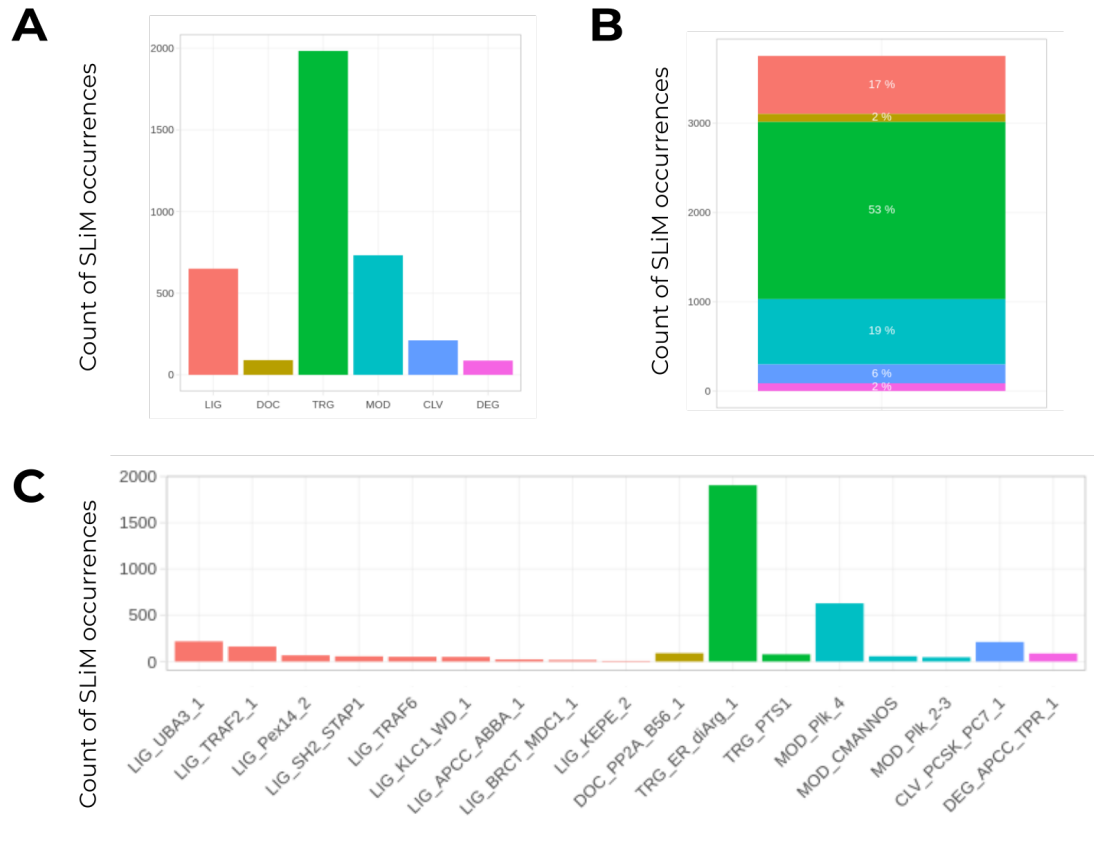


Figure 4: **Non-interacting SLiM classes harbored by sPEPs.** For each SLiM class type, (A) the count of occurrences and (B) proportion of occurrences belonging to the class have been computed. (C) The count of occurrences has also been computed for each individual SLiM class.

3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. mimicINT is of major interest to explore the human sPEP-ome*

Table 6: **Top 10 non interacting SLiM classes based on occurrence counts in sPEPs**

SLiM class	Site name (from ELM)	Function family	Pattern prob.	#Occ. <sup>a</sup>	#Interactions <sup>b</sup>
TRG_ER_diArg.1	di Arginine retention/retrieving signal	ER localization signal	0.0053693	1906	0
MOD_Plk.4	Polo-like kinase phospho-sites	Cell cycle, cytokinesis	0.0060193	629	0
LIG_UBA3.1	Binding motif for UBA3 adenylation domain	Ubiquitination, protein degradation	0.0011962	218	0
CLV_PCSK_PC7.1	PCSK cleavage site	Proteolytic processing of peptides	0.0005087	211	0
LIG_TRAF2.1	TRAF2 binding site	Cell survival	0.0042998	162	0
DOC_PP2A_B56.1	PP2A holoenzyme B56-docking site	Cell cycle, cytoskeleton, growth factor signaling	0.0014581	90	0
DEG_APCC_TPR.1	APCC_TPR-docking motifs	Cell cycle, protein degradation	0.0000136	87	0
TRG_PTS1	PTS1	Peroxisomal localization signal	0.0000152	78	0
LIG_Pex14.2	Pex14 ligand motif	Peroxisomal import	0.0004628	68	0
MOD_CMANNOS	C-Mannosylation site	Protein glycosylation	0.0000469	57	0

<sup>a</sup>: Number of occurrences of the SLiMs in sPEPs

<sup>b</sup>: Number of sPEPRIs mediated by the SLiMs

looked for GO term enrichments among the RefProts interacting with at least one sPEPs (all RefProts expressed in monocytes were used as background).

The figure 5 presents the visualization of the GO biological process (GO:BP) terms using the REVIGO software (data from Table S5). Significant enrichments have been found for 757 GO:BP terms, most of which are related with metabolism (protein metabolic process, macromolecule metabolic process, regulation of metabolic process, regulation of mRNA metabolic process, protein folding, proteolysis, ubiquitin-dependent catabolic process etc.), immunology responses (immune response, immune system process, inflammatory response, antigen processing and presentation, cytokine production etc.), cytoskeleton (cytoskeleton organization, regulation of actin cytoskeleton organization, regulation of actin filament polymerization, endocytosis etc.), signaling (protein phosphorylation etc.) as well as some other regulatory processes (apoptotic process, response to stress, response to stimulus etc.).

### 2.3 90% of sPEPs are annotated with metabolic processes-related GO terms

We then decided to exploit the topology of the network of sPEP-RefProt interactions to perform *in silico* systematic annotation of the sPEPs. To that extent, we first merged the sPEP-RefProt interactions (sPEPRI) network with the canonical protein-protein interactions (PPI) network build as described in the methods section. The resulting network, designated hereafter as the "merged network", is a binary network containing thus sPEPs and RefProts. Two types of interactions were considered then: experimentally identified interactions between RefProts as well as predicted interactions between sPEPs and RefProts. We then identified overlapping clusters and for each cluster, we got the Gene Ontology biological process (GO:BP) terms designating at least half of the RefProts of the cluster. These latter were finally transferred to all the proteins (RefProts and sPEPs) constituting the cluster. This method allowed us to annotate the sPEPs

3. *sPEP* functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. *mimicINT* is of major interest to explore the human *sPEP*-ome

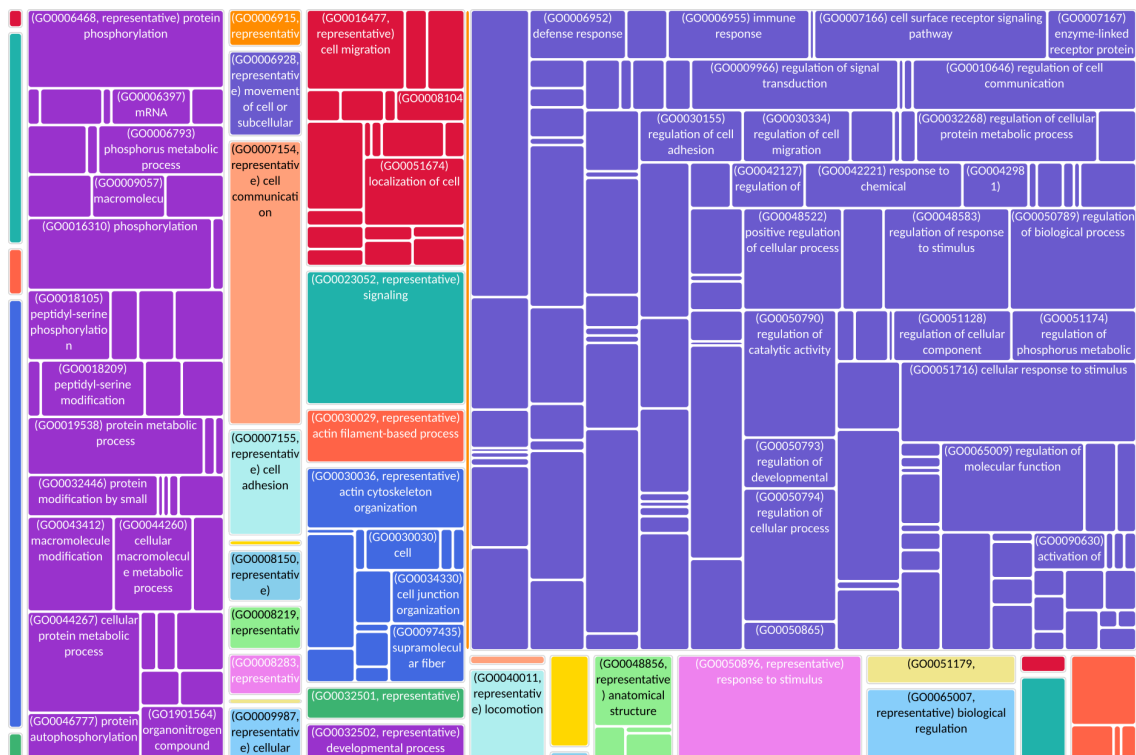


Figure 5: Summarized visualization of the GO:BP term enrichment analysis on interacting RefProts.

3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. mimicINT is of major interest to explore the human sPEP-ome*

with GO:BP terms (Tables S6 and S7).

101 distinct GO:BP terms were used to annotate 3,251 sPEPs and a total of 13,907 annotations (GO:BP) have been assigned to these sPEPs. The 1,142 remaining sPEPs did not received any annotation, as the criterion to annotate their cluster(s) has not been met. The 50 most common GO:BP terms are shown in Table 7.

Table 7: **Top 50 GO:BP terms annotating sPEPs**

Process family	GO:BP term	GO description	#sPEPs
Metabolic process	GO:0080090	regulation of primary metabolic process	1875
	GO:0051171	regulation of nitrogen compound metabolic process	1701
	GO:0031323	regulation of cellular metabolic process	1686
	GO:0031326	regulation of cellular biosynthetic process	27
	GO:0044271	cellular nitrogen compound biosynthetic process	24
	GO:0031325	positive regulation of cellular metabolic process	22
	GO:1901362	organic cyclic compound biosynthetic process	19
	GO:0009889	regulation of biosynthetic process	15
	GO:0044283	small molecule biosynthetic process	12
	GO:0062012	regulation of small molecule metabolic process	12
Macromolecule metabolic process	GO:0031324	negative regulation of cellular metabolic process	7
	GO:0060255	regulation of macromolecule metabolic process	2389
	GO:0010604	positive regulation of macromolecule metabolic process	38
	GO:0009059	macromolecule biosynthetic process	33
	GO:0044260	cellular macromolecule metabolic process	29
	GO:0010556	regulation of macromolecule biosynthetic process	22
	GO:0010605	negative regulation of macromolecule metabolic process	11
Nitrogen compound metabolic process	GO:0043412	macromolecule modification	9
	GO:0006139	nucleobase-containing compound metabolic process	1194
	GO:0090304	nucleic acid metabolic process	420
	GO:0019219	regulation of nucleobase-containing compound metabolic process	50
	GO:0016070	RNA metabolic process	41
	GO:0051173	positive regulation of nitrogen compound metabolic process	35

3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. mimicINT is of major interest to explore the human sPEP-ome*

<i>Process family</i>	<i>GO:BP term</i>	<i>GO description</i>	<i>#sPEPs</i>
	GO:0034654	nucleobase-containing compound biosynthetic process	28
	GO:0051252	regulation of RNA metabolic process	22
Protein metabolic process	GO:0036211	protein modification process	1728
	GO:0019538	protein metabolic process	228
	GO:0016485	protein processing	61
	GO:0031293	membrane protein intracellular domain proteolysis	61
	GO:0006508	proteolysis	12
	GO:0051246	regulation of protein metabolic process	11
Lipid metabolic process	GO:0045540	regulation of cholesterol biosynthetic process	61
Gene expression	GO:0010468	regulation of gene expression	413
	GO:0006357	regulation of transcription by RNA polymerase II	26
	GO:0045893	positive regulation of transcription, DNA-templated	7
Cell cycle	GO:0007049	cell cycle	24
Response to stimulus	GO:0036500	ATF6-mediated unfolded protein response	61
	GO:0033554	cellular response to stress	15
	GO:0071310	cellular response to organic substance	12
Signaling	GO:0007165	signal transduction	981
	GO:0035556	intracellular signal transduction	25
	GO:0009966	regulation of signal transduction	22
	GO:0023051	regulation of signaling	20
	GO:0007166	cell surface receptor signaling pathway	11
Cell communication	GO:0010646	regulation of cell communication	20
Cellular component organization	GO:0007040	lysosome organization	61
	GO:0006996	organelle organization	53
Transport	GO:0060627	regulation of vesicle-mediated transport	70
	GO:0051049	regulation of transport	7
Biological regulation	GO:0065009	regulation of molecular function	14
Cellular localization	GO:0008104	protein localization	7

The figure 6 presents the visualization of the GO:BP terms using the REVIGO web interface. These results show that sPEPs have been mainly annotated with terms related with metabolic processes (nucleic acid metabolic process, protein metabolic process, regulation of cellular metabolic process, regulation of macromolecule metabolic process etc.), stress response (ATF6-mediated stress response), signal transduction and regulation of gene expression. It should

3. *sPEP* functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. *mimicINT* is of major interest to explore the human *sPEP*-ome

be highlighted that these are major biological processes that are ubiquitous in eukaryotes, suggesting the *sPEPs* may be major regulatory peptides.

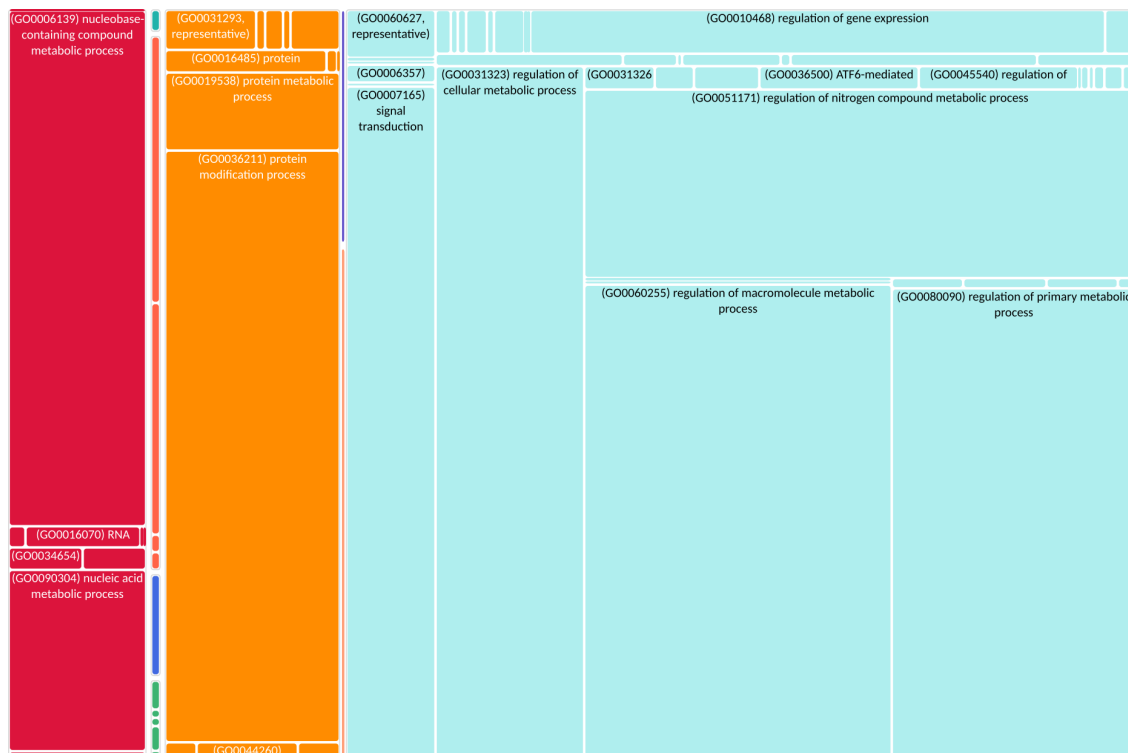


Figure 6: Summarized visualization of the GO:BP terms used to annotate the *sPEPs*. The size of the boxes are related to the number of *sPEPs* that have been annotated with the term shown or a similar term.

56 distinct GO:BP terms used to annotate the *sPEPs* were related to metabolism (*i.e.* have either "inferred related to", "inferred is a" or "inferred part" relation to the GO:BP term "metabolic process" (GO:0008152) according to AMIGO [8], Table S8). More precisely, 82% of the *sPEP* annotations are terms related to metabolism (12,401 / 15,090) and 90% of the *sPEPs* have at least one GO:BP annotation term related to metabolism (2,924 / 3,251). This result strongly suggests the importance of *sPEPs* in the metabolic processes.

## 2.4 *sPEPs* preferentially target the RefProts involved in the same biological processes as the RefProts encoded by their transcripts

Because the existence of sORFs that interact with the canonical protein encoded by their transcript, a growing number of scientists suggested that the co-expression from the same mRNA could facilitate the functionalisation of *sPEPs* and their integration in cellular pathways related to the main protein product of their transcript [26, 36]. Hence, we wondered whether *sPEPs* were preferentially interacting with RefProts involved in the same biological processes as their

3. *sPEP* functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. *mimicINT* is of major interest to explore the human *sPEP*-ome

own genes. To that extent we checked, for a list of GO terms, if there were more RefProts annotated with the GO term among the RefProts interacting with the *sPEPs* encoded by the genes annotated with the GO term than expected by chance (Fig. 7). GO terms used here were those of the generic GO subset (a.k.a. GO slim, a cut-down versions of the GO containing a subset of the terms that give a broad overview of the ontology content without the detail of the specific fine grained terms), restricted to the BP branch.

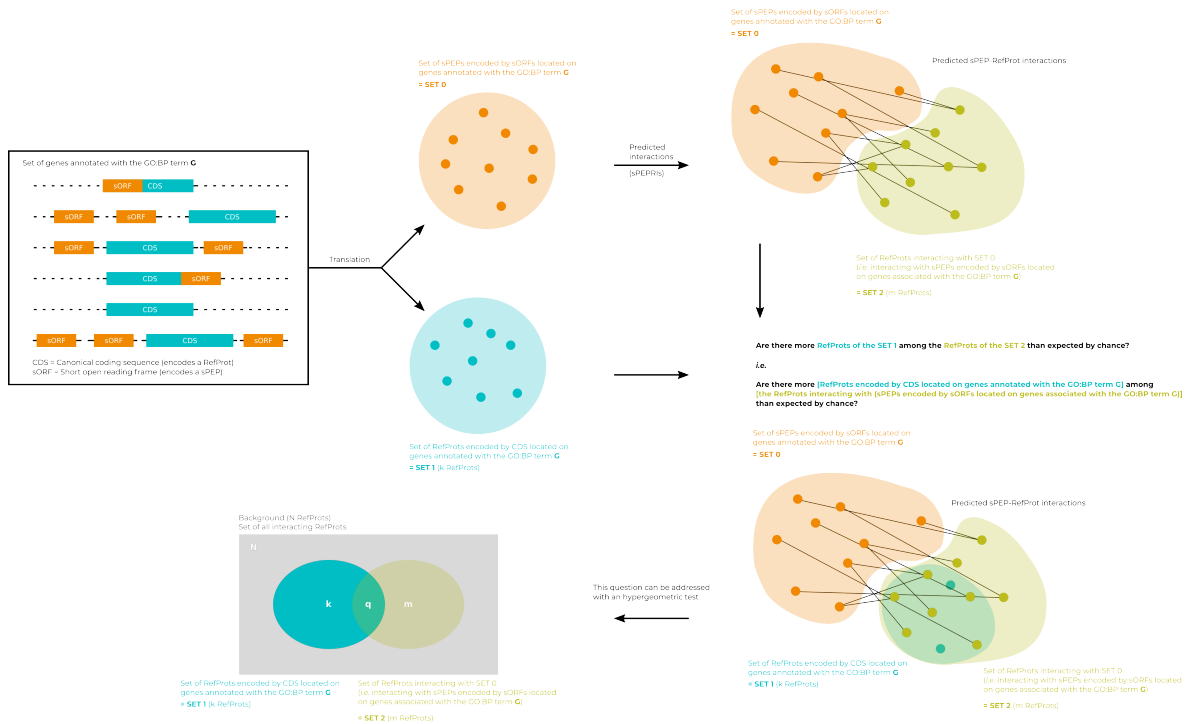


Figure 7: For a certain GO:BP term, are there more RefProts encoded by CDS located on genes annotated with the GO term among the RefProts interacting with *sPEPs* encoded by sORFs located on genes associated with the GO term than expected by chance?

A significant enrichment has been found for 71% GO:BP terms (52 with a FDR lower than 0.05 among 73 terms tested), notably related to metabolism (protein folding, metabolic processes etc.), cell cycle (mitotic cell cycle), cytoskeleton (cytokinesis, cytoskeleton organization etc.) and immune responses (inflammatory response etc.) (Fig. 8). It should be highlighted that the highest odd ratio computed equals 22.76 for the cellular amino acid metabolic process (GO:0006520, FDR =  $10^{-22}$ ) and that odd ratio over 5 are computed mostly for terms related to genetic expression and protein metabolism (mRNA metabolic process, protein catabolic process, cellular amino acid metabolic process, ribosome biogenesis, cytoplasmic translation and gene silencing by RNA) and terms related to the cytoskeleton organization (cytoskeleton organization, establishment or maintenance of cell polarity) (Table 8 and S9).

For each of these GO:BP terms, we then wondered which are the annotations of the RefProts interacting with the *sPEPs* from genes associated with the term. To that extent, defining the

3. *sPEP* functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. *mimicINT* is of major interest to explore the human *sPEP*-ome



Figure 8: Summarized visualization of the GO:BP terms for which there are more RefProts annotated with the term among the RefProts interacting with the *sPEPs* encoded by genes annotated with the term. The size of the boxes are proportional to the number of GO terms aggregated.



3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. mimicINT is of major interest to explore the human sPEP-ome*

Table 8: **Top GO:BP terms ranked by decreasing odd ratio**

GO ID	GO term	Size of the universe <sup>a</sup>	Size of the GO set <sup>b</sup>	Number of sPEP interactors <sup>c</sup>	Intersection <sup>d</sup>	pval	odd ratio	FDR
GO:0006520	cellular amino acid metabolic process	3981	49	307	31	$2.43 \cdot 10^{-23}$	22.8	$1.61 \cdot 10^{-22}$
GO:0042254	ribosome biogenesis	3981	105	448	74	$3.77 \cdot 10^{-48}$	22.3	$6.87 \cdot 10^{-47}$
GO:0002181	cytoplasmic translation	3981	67	510	48	$4.65 \cdot 10^{-29}$	18.8	$3.40 \cdot 10^{-28}$
GO:0016071	mRNA metabolic process	3981	418	1328	343	$5.22 \cdot 10^{-105}$	12.0	$3.81 \cdot 10^{-103}$
GO:0007163	establishment or maintenance of cell polarity	3981	99	1637	86	$1.54 \cdot 10^{-21}$	9.9	$8.64 \cdot 10^{-21}$
GO:0022600	digestive system process	3981	16	190	4	$5.82 \cdot 10^{-3}$	6.8	$9.65 \cdot 10^{-3}$
GO:0030163	protein catabolic process	3981	400	1298	288	$2.34 \cdot 10^{-65}$	6.5	$8.55 \cdot 10^{-64}$
GO:0031047	gene silencing by RNA	3981	44	336	15	$1.38 \cdot 10^{-06}$	5.8	$3.46 \cdot 10^{-6}$
GO:0007010	cytoskeleton organization	3981	541	1951	434	$4.58 \cdot 10^{-58}$	5.1	$1.11 \cdot 10^{-56}$
GO:0016192	vesicle-mediated transport	3981	787	2765	701	$1.05 \cdot 10^{-46}$	4.5	$1.53 \cdot 10^{-45}$

<sup>a</sup>: Size of the universe (*i.e.* all RefProts expressed in monocytes)

<sup>b</sup>: size of the GO set (*i.e.* all RefProts annotated with the GO term) ("Set1")

<sup>c</sup>: Number of RefProts interacting with sPEPs encoded by a gene annotated with the GO term ("Set2")

<sup>d</sup>: Number of RefProts annotated with the GO term and interacting with sPEPs encoded by a gene annotated with

"Set1" as the set of RefProts encoded by CDS located on gene annotated with the GO:BP term, we looked over the same set of GO:BP terms for three types of enrichments:

- Enrichment in RefProts interacting with sPEPs [encoded by genes annotated with the GO terms]. This set of RefProts has been defined as "Set2".
- Enrichment in RefProts interacting with sPEPs [encoded by genes annotated with the GO terms], and being themselves annotated with the GO term. This set of RefProts has been defined as "Set1 inter Set2".
- Enrichment in RefProts interacting with sPEPs [encoded by genes annotated with the GO terms], but not being themselves annotated with the GO term. This set of RefProts has been defined as "Set1 - Set2".

We noticed that significant enrichments were found for GO:BP terms related to similar biological processes as the RefProts interacting with the sPEPs from genes associated with the term (Table S10). These results suggest that even RefProts that are not annotated with a certain GO:BP term but involved in near-cognate biological processes are preferentially interacting with the same sPEPs as the RefProt actually annotated with the GO:BP term.

Finally, it has been shown that genes are able to produce polycistronic transcripts in Eukaryotes [34], notably in drosophila (*e.g.* the tarsal-less gene produces a polycistronic transcript translated into sPEPs) and vertebrates [28, 29], thus more and more studies have been questioning the monocistronic organization of Eukaryotic genomes for the past few years. Hence, we checked the proportion of sPEPs able to interact with the RefProt encoded by their transcript and vice-versa. We observed that 6% (267/4,393) of the sPEPs with at least one interaction with a RefProt are targeting their own RefProt and that 2% (93/3,981) of the RefProts with at least one interaction with a sPEPs are interacting with their own sPEPs.

### 3 Limitations of the study

This study have been performed exclusively on human monocytes, and our findings have been discussed in the scope of this particular species and cell type. In addition, it should be noticed that the list of sPEPs in monocytes has been inferred from the list of sORFs identified by ribosome profiling methods. Hence, as it has been previously highlighted, some of them may not be translated as stable and functional peptides under normal conditions because the ribosome

3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. mimicINT is of major interest to explore the human sPEP-ome*

occupancy is not necessarily associated with an effective translation of a functional protein [24] [17].

Finally, the interactions of sPEPs with RefProts have been inferred by a computational method that is based on the detection of interface interaction. This method, based on mimicINT has the great advantage to provide a comprehensive inference of putative sPEP-RefProt interactions based solely on amino acid sequences. This is of particular interest as experimental data are missing about sPEP biophysical properties (*e.g.* profiles of hydrophobicity) and structures. However, it should be noticed that this method does not take into account the subcellular location of RefProts and sPEPs nor the accessibility of the interaction interfaces for the interactors, making it prone to over-estimation of interactions.

## 4 Conclusion

We first looked at the domain and SLiM usage by the sPEPs and noticed that most of the short linear motifs and domains mediating interactions are involved in several fundamental regulatory functions, such as metabolism, cytoskeleton organization or immunology processes.

Then, we investigated the topology of the sPEP-RefProt interaction network in order to propose sPEP annotations based on cluster identification. This allowed us to annotate most of the sPEPs with GO:BP terms related to metabolic processes, stress response, signal transduction and regulation of the gene expression.

To our knowledge, this study is the first to present a network of sPEP-RefProt interactions in *H. sapiens* as well as GO term annotations for human sPEPs. In addition, our results suggest that most of the sPEPs are likely to be involved in many biological processes, both central to the cell (such as protein, DNA and RNA metabolism, gene expression, or cytoskeleton organization) and related to specialized biological functions (such as immunological responses).

Finally, we performed a functional analysis that suggests that 72% of the time sPEPs encoded by genes annotated with a particular biological process are preferentially interacting with RefProts of the same process.

We think that these findings may be of major importance for the exploration of the regulation of biological processes by sPEPs, which should be consider as a part of the cell proteome in the future.

## 5 Methods

### 5.1 Collection of sPEPs identified in monocytes

The sequences of sPEPs have been collected from MetamORF [10] (<https://metamorf.hb.univ-amu.fr>), a repository of unique short open reading frames identified by both experimental and computational approaches we recently developed. Using the web interface, amino acid sequences of all 10,475 sORFs identified in human monocytes by ribosome profiling have been downloaded as fasta format (Fig. 1A). MetamORF provides classes for the registered ORFs, using an homogeneous nomenclature we previously described [10]. This nomenclature is based upon the ORF length (sORF), transcript biotype (*e.g.* intergenic, ncRNA), relative positions (*e.g.* upstream, downstream) and reading frames (alternative) information.

3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. mimicINT is of major interest to explore the human sPEP-ome*

## 5.2 Collection of RefProts expressed in monocytes

All reviewed sequences of proteins experimentally identified in monocytes according to the Human Proteome Atlas [43] (<https://www.proteinatlas.org>) have been downloaded from UniProtKB [42] (<https://www.uniprot.org/uniprotkb>) as fasta format (Fig. 1B).

## 5.3 Prediction of sPEP-RefProt interactions (sPEPRIs)

The interactions between the sPEPs and the RefProts have been predicted using mimicINT, a workflow for microbe-host protein interactions inference we recently developed [11]. Briefly, mimicINT performs large-scale interaction inference between microbe and human proteins by detecting putative molecular mimicry elements that can mediate the interactions with host proteins: host-like short linear motifs (SLiMs, Fig. 1C) and globular domains (Fig. 1D). It exploits these putative elements to infer the interactions with human proteins by using known templates of domain-domain and domain-SLiM interactions (Fig. 1E). Because sPEPs and RefProts belong to the same species, we may reasonably expect that human sPEPs display interfaces of interactions that resemble structures of the RefProts at the molecular level. Based on this assumption, interactions between sPEPs and RefProts have been inferred using mimicINT.

**Identification of short linear motifs (SLiMs) on sPEPs.** Occurrences of short linear motifs (SLiMs) have been identified on sPEPs using the SLiMProb software [15] (SLiMSuite v1.4.0) (parameters of SLiMProb: minregion = 10, iumethod = short, iucut = 0.4). To that extent, classes of SLiMs registered in the Eukaryotic Linear Motif (ELM) database [21] with at least one true positive instance in *H. sapiens* and a pattern probability lower than 0.01, along with their regular expression, have been provided to SLiMProb. The disorder propensity of each amino acid of the sPEP sequences has been computed by the IUPred [14] software (v1.0) *via* the use of SLiMProb. We defined as general disordered propensity, the ratio of the number of residues with a propensity greater than the selected threshold (0.4) over the length of the sPEP. The functionality of the SLiMs has finally been assessed in a similar fashion as previously proposed by Hagai *et al.* [18]. For each of the 10,475 sPEPs, we created a set of 10,000 shuffled sequences, by randomly shuffling the content of the disordered regions between all the 10,475 sPEPs (Fig. S1). This shuffled set was then used to compare the number of occurrences of SLiMs in the original sequences to their number in the 10,000 shuffled sequences, thereby assessing the likelihood of each SLiM observed in the original sequences of sPEPs to occur by chance. As highlighted by Hagai *et al.* [18], it may be hypothesized that the SLiMs that occur in the original sequences but occur very rarely in the shuffled set are likely to be functional, whereas the functionality of SLiMs that occur frequently in shuffled sequences cannot be inferred. All SLiMs with an empirical probability computed as greater than 0.01 have been discarded.

**Identification of domains on RefProts and sPEPs.** Occurrences of domains have been identified on the RefProts and on the sPEPs using the InterProScan software [20] (v5.41-78.0) looking for Pfam signatures [25]. All occurrences with an e-value greater than  $10^{-5}$  were discarded.

**Prediction of domain - SLiM interactions (DMIs) and domain - domain interactions (DDIs).** Templates for interactions between globular domains and SLiMs have been collected from the Eukaryotic Linear Motif (ELM) database [21]. As the occurrences of domains have been previously detected on RefProts and those of SLiMs on sPEPs, using these templates of interactions allowed inferring the domain-SLiM interactions (DMIs). Templates for interactions between globular domains have been collected from the three-dimensional interacting domains

### 3. *sPEP* functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. *mimicINT* is of major interest to explore the human *sPEP*-ome

(3DID) database [27] as flat format. As the occurrences of domains have been previously detected on both *sPEPs* and RefProts, using these templates of interactions allowed inferring the domain-domain interactions (DDIs).

**Scoring the domain - SLiM interactions (DMIs).** In order to select the DMIs the more likely to be functional, "domain scores" have been computed in a similar fashion as described by Weatheritt *et al.* [45] and all DMIs inferred using *mimicINT* and with a domain score lower than 0.4 have been discarded.

#### 5.4 Annotation of interactors based on network clustering

**Merging the *sPEPPI* network with the canonical protein-protein interaction (PPI) network.** The *sPEP*-RefProt interactions (*sPEPRI*) network has been merged with the canonical protein-protein interaction (PPI) network downloaded from MoonDB [35] (2021 update, unpublished release) and restricted to the RefProt expressed in monocytes according to the Human Proteome Atlas [43]. For the sake of clarity, the resulting network is referred hereafter as the "merged interactome" and contains both *sPEP*-RefProt interactions and RefProt-RefProt interactions.

**Clustering of the "merged interactome".** The largest connected component has been extracted from the "merged interactome" using python-igraph [39] (v0.9.1). This component has been clustered with OCG [3] (default parameters) that aim to maximize the modularity of the classes. Each cluster generated has then been annotated using Gene Ontology (GO) biological process (BP) terms using a classical majority rule as previously described [9]. Briefly, the clusters were annotated according to the BP GO annotations of its constituent proteins. A cluster was annotated with a GO term if at least 50% of annotated RefProts in that cluster shared that GO term. In such cases, all member RefProts and *sPEPs* inherited the annotation(s) of the cluster. Both direct GO annotations and all parent terms were taken into account. Clusters that could be annotated only to the root of the ontology were annotated 'unknown'.

#### 5.5 Enrichment analysis

Enrichment analyses have been performed either using the gProfiler [33] R [38] (v3.6.0) package (parameters: `correction_method = 'gSCS'`) or one-sided Fisher's exact tests followed by Benjamini-Hochberg procedure for multiple correction. False discovery rates (FDR) lower than 0.05 have been considered as significant.

#### 5.6 Collection and visualization of Gene Ontology (GO) terms

**Lists of GO terms and GO annotations of RefProts.** Enrichment analyses using Gene Ontology (GO) [2] [41] terms were performed either on the full set of terms or on the generic GO subset as provided by the GO consortium (downloaded as obo format). The lists of genes related to the GO terms have been collected from gProfiler [33] as gmt format.

**Collection of GO terms based on Pfam accessions.** Pfam [25] accessions have been mapped to GO terms using mapping of GO terms to Pfam entries provided by Gene Ontology (GO) [2] [41].

### 3. *sPEP* functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. *mimicINT* is of major interest to explore the human *sPEP*-ome

**Visualisation of GO terms.** GO terms identified by mapping or enrichment analysis have been visualized using the REVIGO software [37], which allows summarizing long lists of GO terms by finding a representative subsets of the terms based on semantic similarity measures.

#### 5.7 Data availability

Third party softwares and data are available on the editor’s website or using the links provided by the authors in the original publications. The scripts used in this study are available on GitHub (<https://github.com/TAGC-NetworkBiology/InteractORF>). The containerized environments and data are available on Zenodo.

## Acknowledgments

Centre de Calcul Intensif d’Aix-Marseille is acknowledged for granting access to its high performance computing resources.

## Funding

This work has been supported by the “Investissements d’Avenir” French Government program managed by the French National Research Agency (ANR-16-CONV-0001) and by the Excellence Initiative of Aix-Marseille University - A\*MIDEX. SC received a fellowship from the “Espoirs de la recherche” program managed by the French Fondation pour la Recherche Médicale (FDT202106013072).

## References

1. Andrews, S. J. & Rothnagel, J. A. Emerging evidence for functional peptides encoded by short open reading frames. *Nature Reviews Genetics* **15**, 193–204. ISSN: 1471-0056, 1471-0064. <http://www.nature.com/articles/nrg3520> (Mar. 2014).
2. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. en. *Nature Genetics* **25**, 25–29. ISSN: 1061-4036, 1546-1718. [http://www.nature.com/articles/ng0500\\_25](http://www.nature.com/articles/ng0500_25) (2022) (May 2000).
3. Becker, E., Robisson, B., Chapple, C. E., Guénoche, A. & Brun, C. Multifunctional proteins revealed by overlapping clustering in protein interaction network. en. *Bioinformatics* **28**, 84–90. ISSN: 1460-2059, 1367-4803. <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr621> (2022) (Jan. 2012).
4. Blum, M. *et al.* The InterPro protein families and domains database: 20 years on. eng. *Nucleic Acids Research* **49**, D344–D354. ISSN: 1362-4962 (Jan. 2021).
5. Brun, C. *et al.* Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biology* **5**, R6. ISSN: 14656906. <http://genomebiology.biomedcentral.com/articles/10.1186/gb-2003-5-1-r6> (2022) (2003).
6. Brunet, M. A., Leblanc, S. & Roucou, X. Reconsidering proteomic diversity with functional investigation of small ORFs and alternative ORFs. en. *Experimental Cell Research* **393**, 112057. ISSN: 00144827. <https://linkinghub.elsevier.com/retrieve/pii/S0014482720302895> (2022) (Aug. 2020).

3. *sPEP* functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. *mimicINT* is of major interest to explore the human *sPEP*-ome

7. Cabrera-Quio, L. E., Herberg, S. & Pauli, A. Decoding sORF translation – from small proteins to gene regulation. *RNA Biology* **13**, 1051–1059. ISSN: 1547-6286, 1555-8584. <https://www.tandfonline.com/doi/full/10.1080/15476286.2016.1218589> (Nov. 2016).
8. Carbon, S. *et al.* AmiGO: online access to ontology and annotation data. en. *Bioinformatics* **25**, 288–289. ISSN: 1460-2059, 1367-4803. <https://academic.oup.com/bioinformatics/article/25/2/288/220714> (2022) (Jan. 2009).
9. Chapple, C. E. *et al.* Extreme multifunctional proteins identified from a human protein interaction network. en. *Nature Communications* **6**, 7412. ISSN: 2041-1723. <http://www.nature.com/articles/ncomms8412> (2022) (Nov. 2015).
10. Choteau, S. A., Wagner, A., Pierre, P., Spinelli, L. & Brun, C. MetamORF: a repository of unique short open reading frames identified by both experimental and computational approaches for gene and metagene analyses. en. *Database* **2021**, baab032. ISSN: 1758-0463. <https://academic.oup.com/database/article/doi/10.1093/database/baab032/6307706> (2022) (June 2021).
11. Choteau, S. A. *et al.* *mimicINT: a workflow for microbe-host protein interaction inference* en. preprint (Bioinformatics, Nov. 2022). <http://biorxiv.org/lookup/doi/10.1101/2022.11.04.515250> (2022).
12. Couso, J.-P. & Patraquim, P. Classification and function of small open reading frames. *Nature Reviews Molecular Cell Biology* **18**, 575–589. ISSN: 1471-0072, 1471-0080. <http://www.nature.com/doi/10.1038/nrm.2017.58> (July 2017).
13. Davey, N. E. *et al.* Attributes of short linear motifs. en. *Mol. BioSyst.* **8**, 268–281. ISSN: 1742-206X, 1742-2051. <http://xlink.rsc.org/?DOI=C1MB05231D> (2022) (2012).
14. Dosztányi, Z., Csizmók, V., Tompa, P. & Simon, I. The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates between Folded and Intrinsically Unstructured Proteins. en. *Journal of Molecular Biology* **347**, 827–839. ISSN: 00222836. <https://linkinghub.elsevier.com/retrieve/pii/S0022283605001294> (2022) (Apr. 2005).
15. Edwards, R. J., Paulsen, K., Aguilar Gomez, C. M. & Pérez-Bercoff, Á. en. in *Intrinsically Disordered Proteins* (eds Kragelund, B. B. & Skriver, K.) 37–72 (Springer US, New York, NY, 2020). ISBN: 978-1-07-160523-3. [https://link.springer.com/10.1007/978-1-0716-0524-0\\_3](https://link.springer.com/10.1007/978-1-0716-0524-0_3) (2022).
16. Erhard, F. *et al.* Improved Ribo-seq enables identification of cryptic translation events. *Nature Methods* **15**, 363–366. ISSN: 1548-7091, 1548-7105. <http://www.nature.com/doi/10.1038/nmeth.4631> (Mar. 2018).
17. Gray, T., Storz, G. & Papenfort, K. Small Proteins; Big Questions. en. *Journal of Bacteriology*. ISSN: 0021-9193, 1098-5530. <https://journals.asm.org/doi/10.1128/JB.00341-21> (2021) (July 2021).
18. Hagai, T., Azia, A., Babu, M. M. & Andino, R. Use of Host-like Peptide Motifs in Viral Proteins Is a Prevalent Strategy in Host-Virus Interactions. en. *Cell Reports* **7**, 1729–1739. ISSN: 22111247. <https://linkinghub.elsevier.com/retrieve/pii/S2211124714003702> (2022) (June 2014).
19. Hinnebusch, A. G., Ivanov, I. P. & Sonenberg, N. Translational control by 5'-untranslated regions of eukaryotic mRNAs. en. *Science* **352**, 1413–1416. ISSN: 0036-8075, 1095-9203. <https://www.science.org/doi/10.1126/science.aad9868> (2021) (June 2016).

3. *sPEP* functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. *mimicINT* is of major interest to explore the human *sPEP*-ome

20. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. en. *Bioinformatics* **30**, 1236–1240. ISSN: 1367-4803, 1460-2059. <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu031> (2022) (May 2014).
21. Kumar, M. *et al.* The Eukaryotic Linear Motif resource: 2022 release. en. *Nucleic Acids Research* **50**, D497–D508. ISSN: 0305-1048, 1362-4962. <https://academic.oup.com/nar/article/50/D1/D497/6414054> (2022) (Jan. 2022).
22. Laumont, C. M. *et al.* Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nature Communications* **7**, 10238. ISSN: 2041-1723. <http://www.nature.com/doifinder/10.1038/ncomms10238> (Jan. 2016).
23. Luck, K. *et al.* A reference map of the human binary protein interactome. en. *Nature* **580**, 402–408. ISSN: 0028-0836, 1476-4687. <http://www.nature.com/articles/s41586-020-2188-x> (2022) (Apr. 2020).
24. Makarewich, C. A. & Olson, E. N. Mining for Micropeptides. en. *Trends in Cell Biology* **27**, 685–696. ISSN: 09628924. <https://linkinghub.elsevier.com/retrieve/pii/S0962892417300648> (2022) (Sept. 2017).
25. Mistry, J. *et al.* Pfam: The protein families database in 2021. en. *Nucleic Acids Research* **49**, D412–D419. ISSN: 0305-1048, 1362-4962. <https://academic.oup.com/nar/article/49/D1/D412/5943818> (2022) (Jan. 2021).
26. Moro, S. G., Hermans, C., Ruiz-Orera, J. & Albà, M. M. Impact of uORFs in mediating regulation of translation in stress conditions. en. *BMC Molecular and Cell Biology* **22**, 29. ISSN: 2661-8850. <https://bmcmolcellbiol.biomedcentral.com/articles/10.1186/s12860-021-00363-9> (2021) (Dec. 2021).
27. Mosca, R., Céol, A., Stein, A., Olivella, R. & Aloy, P. 3did: a catalog of domain-based interactions of known three-dimensional structure. en. *Nucleic Acids Research* **42**, D374–D379. ISSN: 0305-1048, 1362-4962. <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkt887> (2022) (Jan. 2014).
28. Mouilleron, H., Delcourt, V. & Roucou, X. Death of a dogma: eukaryotic mRNAs can code for more than one protein. eng. *Nucleic Acids Research* **44**, 14–23. ISSN: 1362-4962 (Jan. 2016).
29. Mudge, J. M. *et al.* Standardized annotation of translated open reading frames. eng. *Nature Biotechnology* **40**, 994–999. ISSN: 1546-1696 (July 2022).
30. Olexiuk, V. *et al.* sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Research* **44**, D324–D329. ISSN: 0305-1048, 1362-4962. <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1175> (Jan. 2016).
31. Plaza, S., Menschaert, G. & Payre, F. In Search of Lost Small Peptides. *Annual Review of Cell and Developmental Biology* **33**, 391–416. ISSN: 1081-0706, 1530-8995. <http://www.annualreviews.org/doi/10.1146/annurev-cellbio-100616-060516> (Oct. 2017).
32. Pueyo, J. I., Magny, E. G. & Couso, J. P. New peptides under the s(ORF)ace of the genome. *Trends in Biochemical Sciences* **41**, 665–678. ISSN: 09680004. <https://linkinghub.elsevier.com/retrieve/pii/S0968000416300317> (Aug. 2016).
33. Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). en. *Nucleic Acids Research* **47**, W191–W198. ISSN: 0305-1048, 1362-4962. <https://academic.oup.com/nar/article/47/W1/W191/5486750> (2022) (July 2019).

3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. mimicINT is of major interest to explore the human sPEP-ome*

34. Renz, P. F., Valdivia-Francia, F. & Sandoel, A. Some like it translated: small ORFs in the 5'UTR. en. *Experimental Cell Research* **396**, 112229. ISSN: 00144827. <https://linkinghub.elsevier.com/retrieve/pii/S001448272030478X> (2021) (Nov. 2020).
35. Ribeiro, D. M., Briere, G., Bely, B., Spinelli, L. & Brun, C. MoonDB 2.0: an updated database of extreme multifunctional and moonlighting proteins. en. *Nucleic Acids Research* **47**, D398–D402. ISSN: 0305-1048, 1362-4962. <https://academic.oup.com/nar/article/47/D1/D398/5146199> (2022) (Jan. 2019).
36. Samandi, S. *et al.* Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. *eLife* **6**. ISSN: 2050-084X. <https://elifesciences.org/articles/27860> (Oct. 2017).
37. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. en. *PLoS ONE* **6** (ed Gibas, C.) e21800. ISSN: 1932-6203. <https://dx.plos.org/10.1371/journal.pone.0021800> (2022) (July 2011).
38. Team, R. C. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. <https://www.R-project.org/> (2018).
39. Team, T. I. C. *igraph* June 2022. <https://zenodo.org/record/3630268> (2022).
40. Tharakan, R. & Sawa, A. Minireview: Novel Micropeptide Discovery by Proteomics and Deep Sequencing Methods. *Frontiers in Genetics* **12**, 651485. ISSN: 1664-8021. <https://www.frontiersin.org/articles/10.3389/fgene.2021.651485/full> (2021) (May 2021).
41. The Gene Ontology Consortium *et al.* The Gene Ontology resource: enriching a GOLD mine. en. *Nucleic Acids Research* **49**, D325–D334. ISSN: 0305-1048, 1362-4962. <https://academic.oup.com/nar/article/49/D1/D325/6027811> (2022) (Jan. 2021).
42. The UniProt Consortium *et al.* UniProt: the universal protein knowledgebase in 2021. en. *Nucleic Acids Research* **49**, D480–D489. ISSN: 0305-1048, 1362-4962. <https://academic.oup.com/nar/article/49/D1/D480/6006196> (2022) (Jan. 2021).
43. Uhlen, M. *et al.* Towards a knowledge-based Human Protein Atlas. en. *Nature Biotechnology* **28**, 1248–1250. ISSN: 1087-0156, 1546-1696. <http://www.nature.com/articles/nbt1210-1248> (2022) (Dec. 2010).
44. Vitorino, R., Guedes, S., Amado, F., Santos, M. & Akimitsu, N. The role of micropeptides in biology. en. *Cellular and Molecular Life Sciences*. ISSN: 1420-682X, 1420-9071. <http://link.springer.com/10.1007/s00018-020-03740-3> (2021) (Jan. 2021).
45. Weatheritt, R. J., Luck, K., Petsalaki, E., Davey, N. E. & Gibson, T. J. The identification of short linear motif-mediated interfaces within the human interactome. en. *Bioinformatics* **28**, 976–982. ISSN: 1460-2059, 1367-4803. <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts072> (2022) (Apr. 2012).



3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. mimicINT is of major interest to explore the human sPEP-ome*

## **SUPPLEMENTARY MATERIAL**



3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. mimicINT is of major interest to explore the human sPEP-ome*

and 10,000 artificial sequences have been generated. SLiMs have thus been identified on the shuffled sequences. Finally, the likelihood of SLiMs to occur serendipitously has been assessed by comparing the number of natural occurrences of SLiMs in the original sequences to their number of occurrences in the shuffled sequences.

In the example, the MOD\_PKA\_2 motif that occurs once in the original sequences, occurs at least as many time in 5 of 10 shuffled sequences (grey circle). The MOD\_NEK2\_1 motif that occurs once in the original sequence, occurs at least as many time in 2 of 10 shuffled sequences (pink circles). Finally, the empirical probability of occurrence of the SLiM is computed using the following formula:

$$pval = \frac{k + 1}{N + 1}$$

with:

- $k$  the number of shuffled sequences in which the SLiM is observed at least as many times as in the original sequence
- $N$  the number of shuffled sequences (10,000)

In the example,

$$pval_{MOD\_PKA\_2} = \frac{5 + 1}{10 + 1} \approx 0.5$$

$$pval_{MOD\_NEK2\_1} = \frac{2 + 1}{10 + 1} \approx 0.3$$

## 6.2 Length of sPEPs depending on the presence of SLiMs or domains

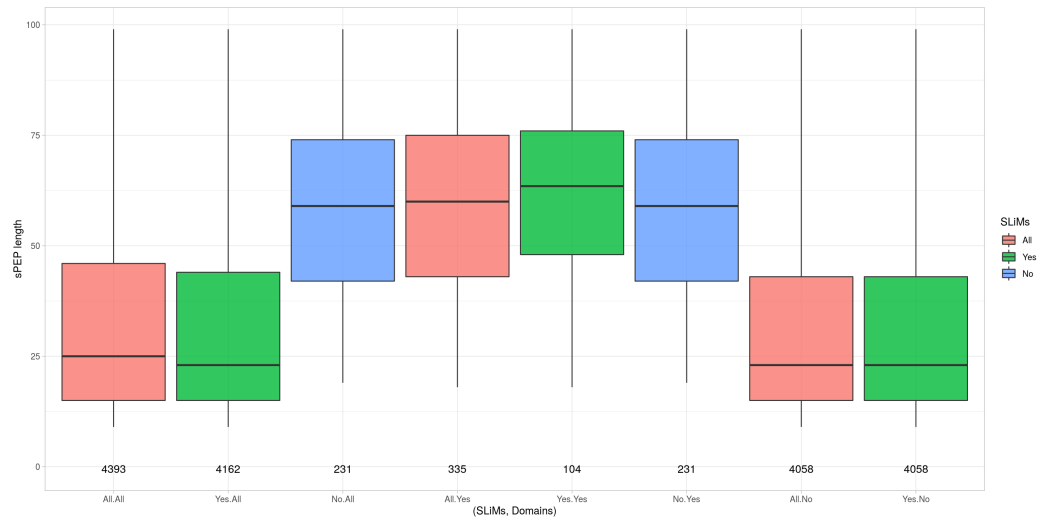


Figure S2: The sPEPs harboring domains are the longest.

3. *sPEP functions in monocytes have been assessed by a system approach based on their interactions with canonical proteins – 3.3. mimicINT is of major interest to explore the human sPEP-ome*

### 6.3 List of supplementary data files

- Fasta file containing the amino acid sequences of the sPEPs identified in Monocytes (from MetamORF)
- Fasta file containing the amino acid sequences of the RefProts expressed in Monocytes (from UniProtKB)

Table S1: Domains used by sPEPs. This file includes the Pfam signatures, the number of occurrences and of interactions mediated by occurrences of the domain as well as the mapping to GO terms (.tsv file)

Table S2: Domains not mediating interactions used by sPEPs. This file includes the Pfam signatures, the number of occurrences and of interactions mediated by occurrences of the domain as well as the mapping to GO terms (.tsv file)

Table S3: SLiMs used by sPEPs. This file includes the ELM identifiers, the number of occurrences and of interactions mediated by occurrences of the SLiM (.tsv file)

Table S4: SLiMs not mediating interactions used by sPEPs. This file includes the ELM identifiers, the number of occurrences and of interactions mediated by occurrences of the SLiM (.tsv file)

Table S5: Gene ontology biological processes annotations of the interacting RefProts (.tsv file)

Table S6: Gene ontology biological processes annotations of the sPEPs (.tsv file)

Table S7: Count of sPEPs for each Gene Ontology biological processes annotation (.tsv file)

Table S8: Count of sPEPs for each Gene Ontology biological processes annotation related to metabolic process (GO:0008152) (.tsv file)

Table S9: Enrichments on RefProts interacting with [sPEPs encoded by genes annotated with a GO term] and annotated by the same GO term for GO:BP terms in the GO generic subset (.tsv file)

Table S10: GO:BP terms for which significant enrichments (FDR  $\leq$  0.05) on the various sets of RefProts (.tsv file)

## 4. Understanding of the translational regulation by uORFs may be improved by mathematical modeling

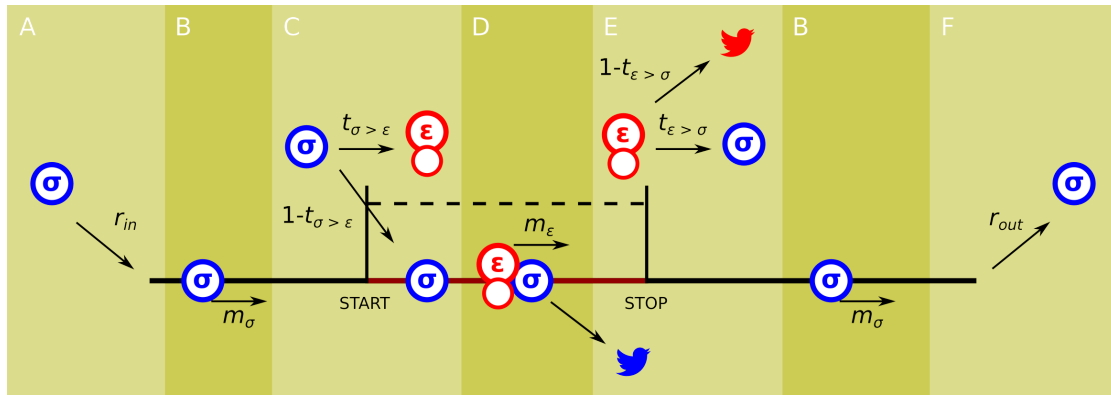
### 4.1. Agent-based modeling may help deciphering complex mechanisms

#### 4.1.1. ICIER, a published TASEP model, partially explains the translational regulation by a single uORF

The third objective of my thesis was to explore the mechanisms of translation by the uORFs. Because of the diversity of uORF organization, the mechanism of uORF-mediated resistance may vary and should be studied individually [7]. In 2018, Andreev *et al.* [6] published [initiation complexes interference with elongating ribosomes \(ICIER\)](#), a pioneer mathematical model to study the regulation of the translation by uORFs. This model is based on the [totally asymmetric simple exclusion process \(TASEP\)](#), a stochastic dynamical system of unidirectional particle movement through a unidimensional lattice where each site can be occupied by no more than one particle, and the probability of particle transition from one site to another is predefined [5, 6]. In 2020, the same team approximated it by a phenomenological deterministic model based on similar assumptions [5]. This last one allowed a rigorous analysis and admitted explicit solutions in limit cases, but failed to explain most of the variability observed in [Ribo-seq](#) data, probably because of its relative simplicity.

The [initiation complexes interference with elongating ribosomes \(ICIER\)](#) model aims at establishing the relationship between the flux of scanning ribosomes loaded at the 5'UTR extremity of a transcript and those reaching the CDS. To do so, it models the flux of scanning ribosome upstream and downstream of a single uORF depending on its features [6]. In ICIER, ribosomes represent the particles and can have two states that give them different dynamic properties: scanning or elongating (*i.e.* translating), with possibilities of transition from one state to the other (Figure 4.1). The model is based on the strong assumption that elongating ribosomes will obstruct the progression of scanning ribosomes, such obstruction which is relieved during stress due to lower initiation at the uORF [6].

4. Understanding of the translational regulation by uORFs may be improved by mathematical modeling – 4.1. Agent-based modeling may help deciphering complex mechanisms



**Figure 4.1.: Principle of the initiation complexes inference with elongating ribosomes (ICIER) model.** Adapted from Andreev *et al.* (2018) [6]. Scanning ribosomes are represented with blue  $\sigma$  and elongating ribosomes with red  $\epsilon$  on the figure. The lattice is shown as a black line. The uORF is represented in dark red. (A)  $r_{in}$  reflects the rate of 43S PIC loaded at the 5' end of the transcript, *i.e.* the ternary complex availability. Normal conditions:  $r_{in} = 0.1$ , absolute stress:  $r_{in} = 0$  (no ribosome loaded). (B)  $m_{\sigma}$  is the probability to move forward for a scanning ribosome, *i.e.* its reflects the speed of the scanning ribosome. If  $m_{\sigma} = 0$ , the scanning ribosome is stalling. (C)  $t_{\sigma > \epsilon}$  is the probability to initiate translation for a scanning ribosome, *i.e.* for a scanning ribosome to turn into an elongating one. Leaky scanning:  $t_{\sigma > \epsilon} = 0$  (no initiation), Non-leaky initiation:  $t_{\sigma > \epsilon} = 1$  (systematic initiation). (D)  $m_{\epsilon}$  is the probability to move forward for an elongating ribosome, *i.e.* it reflects the speed of the elongating ribosome. If  $m_{\epsilon} = 0$ , the elongating ribosome is stalling. Any downstream scanning ribosome that collides with an upstram elongating ribosome dissociates from the transcript (bird symbol). (E)  $t_{\epsilon > \sigma}$  is the probability to reinitiate at the end of the translation, *i.e.* for an elongating ribosome to turn back into a scanning one. No reinitiation:  $t_{\epsilon > \sigma} = 0$  (the ribosome leaves the transcript), systematic reinitiation:  $t_{\epsilon > \sigma} = 1$ . (F)  $r_{out}$  reflects the rate of scanning ribosome reaching the end of the 5'UTR, *i.e.* of ribosomes able to initiate translation at the CDS. If  $r_{out} = 0$ , no scanning ribosome reaches the CDS, meaning there is a total repression of the CDS translation.

Based on a literature review, Andreev *et al.* hypothesized that scanning ribosomes would dissociate from mRNAs when upstream elongating ribosomes collide with them. This hypothesis is based on the assumption that when moving ribosomes collide with downstream ribosome complexes, they may either stay on the transcript or dissociate. Based on existing evidence in the literature, they demonstrated that scanning and elongating ribosomes are both able to queue upstream of an elongating ribosome, whilst scanning ribosome dissociate from mRNA when the collision occurs with an elongating ribosome upstream [6].

Exploiting the ICIER model, Andreev *et al.* were able to check how several important uORFs features are affecting the translation of the CDS, notably the existence of an

4. Understanding of the translational regulation by uORFs may be improved by mathematical modeling – 4.1. Agent-based modeling may help deciphering complex mechanisms

uORF in the 5'UTR, the uORF length ( $2 \leq L \leq 200$ ), movement rates for elongating ribosomes ( $0.2 \leq m_e \leq 0.35$ ), initiation efficiency ( $0.3 \leq t_{\sigma > \epsilon} \leq 0.9$ ), delay at the start codon after translation initiation before to move forward ( $0 \leq \delta_\epsilon \leq 0.2$ ), reinitiation efficiency ( $0 \leq t_{\epsilon > \sigma} \leq 0.055$ ), scanning ( $0 \leq P_\sigma \leq 0.007$ ) and elongating ( $0 \leq P_\epsilon \leq 0.007$ ) spontaneous dissociations as well as scanning ribosome size ( $6 \leq D_\sigma \leq 14$ ). [6].

They notably demonstrated that (upon the hypothesis of non-reinitiation) a single long uORF represses the translation of the CDS, which is de-repressed during stress. Indeed, in absence of uORF and without reinitiation, the  $r_{out}$  correlates nonlinearly with  $r_{in}$ . Interestingly, at a near zero value of  $r_{in}$ , the rates  $r_{in}$  and  $r_{out}$  begin to decrease proportionally, which could be explained by changes in the likelihood of ribosome collisions, as they reduce their flow. Over a certain uORF length, this relationship becomes non-monotonous. In uORF-containing RNAs, levels of  $r_{out}$  are increased with  $r_{in}$  when it is high and this repression increases with uORF length. They explain this phenomenon by the increasing incidence of collisions involving scanning and elongating ribosomes within the uORF and the subsequent dissociation of scanning ribosomes.

They also observed that a small decrease of the elongation rate ( $m_e$ ) in the uORF strongly increases the maximum  $r_{out}$  relatively to its basal level (defining the relative  $r_{out}$  as  $\frac{r_{out}[r_{in}]}{r_{out}[r_{in}=0.1]}$ ), suggesting that stress resistance is increased as the elongation rate decreases. However, they demonstrated also that the more slowly decoded uORFs provide greater resistance to the stress, but at a cost of greater CDS repression under normal conditions. By comparing with the scanning rate, they concluded that there should be an optimal ratio of scanning to elongating ribosome velocities for the uORF to provide stress-resistance. It should be noticed that Andreev et al. considered by default in other simulations a probability of 0.3 that the ribosome moves during a single tact. As average mammalian ribosomes move five codons per second during elongation (a figure that is in accordance with the translocation reactions rate of 2-20 per second reported by Rundlet et al. [112]), a tact in the simulation would correspond to 0.06 s.

The ICIER model demonstrated that increased leakiness of the uORF start elevates the flow of ribosomes downstream of the uORF, a result expected as this is associated with lower dissociation of ribosomes from the transcript at the end of the uORF. This results in a reduced stress resistance, suggesting that uORFs with weaker initiation contexts are less likely to provide stress resistance to the CDS. In addition, an increased time spent at the start codon by the ribosome for starting translation reduces the inhibitory effect of the uORF and thus reduces stress resistance. Andreev et al. explained this observation by the increased distance to the downstream scanning ribosomes and thus decreased chances of collision.

Elevated reinitiation reduces the inhibitory effect of uORFs and their ability to provide stress resistance. They concluded that a single uORF enabling reinitiation to take place does not provide a stress resistance, a mechanism that would be very different than those described for ATF4 and that involves a combination of several uORFs with allowed reinitiation.

#### 4. Understanding of the translational regulation by uORFs may be improved by mathematical modeling – 4.1. Agent-based modeling may help deciphering complex mechanisms

According to Andreev et *al.*, spontaneous dissociation of scanning ribosome is likely to occur because of the unstable link between them and their transcript, although the exact reasons of such dissociation remain to be elucidated. They demonstrated that increased drop-off rates reduce both the inhibitory and stress-protective properties of uORFs.

ICIER is the first advanced model of translational regulation by the uORFs that has been proposed so far (to the best of my knowledge). However, it cannot make accurate predictions of stress-resistance levels for specific mRNAs and struggles to explain the stress-resistance observed in Ribo-seq data (positive correlation between scores of certain uORF features are observed, but they appear weak and not statistically significant). This does not necessarily invalidate the ICIER model but highlights the necessity to pursue our efforts in the development of such models. In addition, it only integrates mRNAs that harbor a single uORF, whilst the archetypal model of ATF4/GCN4 involves two functional uORFs. I hence propose to develop an extended model of ICIER to help tackling these issues.

#### 4.1.2. Many parameters and uORF features may impact the translation and should be considered in future models of translational regulation by uORFs

As discussed, one fundamental additional parameter that is susceptible to deeply affect the regulation of the translation is the number of uORFs harbored by the transcript [32]. Hence, one may think it is of major importance to consider now the possibility to integrate several uORFs (including overlapping and nested ORFs) to new models.

While ICIER already integrates some of the most important features susceptible to affect the translation, namely the uORF length, movement rates for elongating ribosomes, initiation efficiency, delay at the start codon after translation initiation before to move forward, reinitiation efficiency, scanning and elongating spontaneous dissociations and the scanning ribosome size, all of these parameters must be set to a constant value and do not take into consideration the local context of the ribosomes.

As an example, the efficiency of the initiation is known to be affected by the start codon sequence and its local context. Indeed, uORFs starting with non-AUG codons (including UUG and AUA) show a much lower translation initiation efficiency [88], suggesting the importance of the start codon sequence in the initiation of the translation, and consequently the regulation by the uORFs. Hinnebush et *al.* [54] also reported that NUG codons are the most efficient for translation initiation whilst A(A/G)G are the worst. Spealman et *al.* [131] demonstrated that AUG and near-cognate codons (*i.e.* single nucleotide variants of AUG) have higher translation efficiency under certain stress conditions. In addition, it should be highlighted that uORFs using alternative start codons are usually not considered in studies that aim to decipher the regulation



#### 4. Understanding of the translational regulation by uORFs may be improved by mathematical modeling – 4.1. Agent-based modeling may help deciphering complex mechanisms

of the translation, whilst the difference of translational efficiency may be of major importance in the regulation of the translation of the CDS.

In 1986, M. Kozak identified ACCATGG (where a purine (A/G) is in position -3, and the AUG initiation codon is underlined) as the optimal sequence for initiation by eukaryotic ribosomes [69], a sequence now designated as the "Kozak consensus" or "Kozak sequence context". This context has been proved many times since then to enhance the translation of CDS, and it has been reported more recently that the uORF start codon plays a crucial role in controlling its translation [163]. In addition, uORFs whose AUG codons better conform to the Kozak consensus have been reported to repress more the translation of the CDS [54]. It should also be noticed that Kozak sequence context of uORFs has evolved across eukaryotic clades and that uORFs with canonical Kozak sequences context seem to have stronger suppressive effects on the translation than non-canonical ones [163]. Surprisingly, Chew *et al.* [32] demonstrated on contrary that uORFs do not have a distinct initiation sequence context that promote their translation but estimated that nearly 17 % of uORFs have a more favourable initiation context than the median initiation context of CDSs. They also reported that a more favourable initiation context sequence and a less stable secondary structure correlate with an increased translation efficiency (TE) of uORFs [32]. It has also been hypothesized that this context may even have a larger influence on non-canonical start codons, at least when it comes to stabilizing the 43S PIC at the A/G in the -3 position and at the G in the +4 position [54, 109]. Unfavorable sequences near to the start codon can also lead to leaky scanning [54], a mechanism that is favored in CDSs by excessively short 5'UTRs (< 20 nt) [54]. Although the impact of Kozak consensus on the initiation of uORFs is still debated, a growing body of evidence suggests that the initiation codon sequence and its Kozak context are of major importance for the initiation of the translation of uORFs. Hence, these information should ideally be taken into consideration when modeling the translational regulation by uORFs.

Despite having been less studied than start codons, the stop codons may be of major importance, as they are susceptible to influence reinitiation. Some uORFs inhibit downstream translation primarily because ribosomes stall during the elongation or termination, and create a roadblock to scanning ribosomes that bypassed the uORF start codon [54]. In 1987, Kozak demonstrated that the efficiency of reinitiation is progressively improved as the intercistronic sequence is lengthened in eukaryotic ribosomes [68], suggesting an important role for the distance between an uORF and its CDS for the efficiency of reinitiation [32]. Chew *et al.* [32] reported that the efficiency of reinitiation had been observed to decrease as the distance between uORFs and CDSs decreases. On contrary, Couso and Patraquim [36] stated that reinitiation can occur if the initially translated ORF is no longer than 30 aa and if an additional ORF is found approximately 100 to 200 nt downstream of the stop codon of the initially translated ORF. This may be explained by the retention of the eIF3 factor that is facilitated by shorter uORFs and allows for reinitiation to occur [54]. McGillivray *et al.* [84] observed that regulatory uORFs are on average closer to the CDS (203 nt from the CDS start codon in average) and located in shorter 5'UTRs. Chew *et al.* reported also that

4. *Understanding of the translational regulation by uORFs may be improved by mathematical modeling – 4.1. Agent-based modeling may help deciphering complex mechanisms*

**uORF** lengths and the distances between **uORFs** and **CDSs** contributed significantly towards specifying **CDS translation efficiency** [32]. Different translation termination efficiencies have been noted and UAA allows a greater termination efficiency as well as a more rapid ribosomal dissociation from the transcript (UAA is more efficient than UAG being itself more efficient than UGA) [74]. Hence, stronger stop codons in **uORF** are more susceptible to decrease the probability that ribosomes reinitiate translation at a downstream **ORF** [74].

It is also known that elongation rates are not constant along the transcript (5-6 codons per second in average in mouse [122]). In particular, **uORF** mutations introducing suboptimal codons have been shown to slow down translational elongation and impede downstream translation initiation [74]. By the way, the codon usage bias is known to be the major determinant of translation elongation rates. The elongation rates are even generally not evenly distributed on synonymous codons and rare codons are more likely to reduce the elongation rates, as ribosomes require more residence time than commonly used codons, and this may even sometimes causes ribosome stalling [135]. Elongation rate may be also affected by the hydrophobicity and the charge of encoded amino acids [135]. Hence, an ideal model should also include variation in the elongation rate, primarily based on codon rarity, as well as on biophysical properties of the encoded amino acids.

Frameshift, a mechanism by which a ribosome slip back one nucleotide and continues translation in the -1 reading frame; and stop codon readthrough, a mechanism by which a stop codon is ignored by an elongating ribosome, have been described in eukaryotes. Frameshift usually requires two *cis*-acting signals to happen (a heptanucleotide and a short downstream **RNA** structure) [165]. This stresses out the importance to integrate **uORFs** using alternative reading frames in future modeling.

Finally, **mRNAs** enriched with more optimal codons are both more stable and more efficiently translated by the ribosome [74].

To conclude, three main mechanisms of regulations have been highlighted as having a major impact on translation [84]: (i) translation reinitiation: the ribosome is able to resume translation at a downstream **ORF** after the translation of a first **ORF**; (ii) leaky scanning: the ribosome bypasses the translation of an **ORF**; and (iii) ribosome stalling: the ribosome stalls at the start or stop codon of an **ORF** or during elongation.

As demonstrated, these mechanisms are dependent on many features, including notably the **ORF** sizes, the distance between two successive **ORFs**, the **uORF** start codon sequences and their local contexts (Kozak consensus), the stop codon sequences, the translation rates, frameshifts and stop codon readthrough. All these features are thus susceptible to greatly impact the translational efficiency of **ORFs** and **CDSs** and should thus ideally be considered in the implementation of future models.

### 4.1.3. Agent-based models (ABMs) have been used to solve complex questions

Computer modeling can be seen as a mean of dynamic knowledge representation that can form a basis for formal means of testing, evaluating and comparing what is currently known by the scientific community [3]. However, the algorithmic implementation of so many parameters as the ones presented in the previous section is challenging and computational procedural approaches and inductive models usually struggle at modeling such complex systems. Fortunately, agent-based simulations have proved to be efficient when it comes to studying autonomous agents (*e.g.* ribosomes, molecules, cells, organisms, individuals etc.) with their environment (*e.g.* transcript, cell, organism, building etc.). They provide an adequate tool to capture the complexity and dynamics of large systems [19, 26]. Agent-based modeling is a simulation technique which replicates decision-making entities, called agents [19]. Agent-based models (ABMs) are rule-based, discrete-event and discrete-time computational methods that use computational objects that focus on the rules and interactions among the individual components (the agents) of a system. They are of particular interest in the field of system biology because of their ability to encompass multiple scales of process and spatial considerations [3, 15]. However, it should be noticed that ABMs are not inductive models, *i.e.* they are not based on patterns of data, but instead intend to reconstruct mechanisms of observed patterns of data, starting with simple rules for behaviors and based on known or presumptive mechanisms [3].

Such models have notably been successfully used in the fields of sociology, economy, finance, management sciences, robotics, security, building industry, town planning and more recently anthropology and biology (including ecology, cellular and molecular biology as well as medicine) to study the participation of individuals' behaviors to consequences at the (eco-)system level. They have even been used by companies for defining investment strategies and during decision-making processes [3, 19, 26, 59]. ABMs are based on the paradigm that individuals behave according to internal laws as well as in response to their understanding of the (local) environment. An *et al.* [3] describe ABMs as having the following properties: they have the ability to integrate easily space (usually using grids), they utilize parallelism (differing local conditions lead to different behavioral trajectories of the individual agents), they integrate stochasticity (behaviors can be based on probabilities), they have a modular structure (the behavior of the model is dictated by the rules of its agents), they reproduce emergent properties (as a consequence of previous points, they generate systemic dynamics that could not have been reasonably inferred from examination of the rules of the agents alone) and they can be constructed in the absence of complete knowledge (the rules are defined as simple and verifiable as possible). Janssen *et al.* stress out the emergence of patterns, structures and behaviours that were not explicitly programmed into the models as a consequence of agent interactions [59] (*the whole is more than the sum of the system's parts* [108]). Finally, they represent the advantage to be usually more intuitive to non-mathematicians than alternative modeling such as

4. Understanding of the translational regulation by uORFs may be improved by mathematical modeling – 4.1. Agent-based modeling may help deciphering complex mechanisms

ordinary differential equations or partial differential equations and their variants [3], and to allow integration of many more parameters [15]. In addition, such paradigm allows to design expressive and realistic agents, but it has been reported to be difficult to implement for non-computer scientists [26]. For instance, ABMs have already been successfully implemented for studying intracellular trafficking, viral infections, circulation of inflammatory cells in guts, systemic inflammation and sepsis, tumor growth, angiogenesis, infectious diseases spreading or displacements of corpses by ants (unpublished data and) [3, 63].

In ABMs, each agent individually assesses its situation and its environment and makes decisions on the basis of internal rules (that may be updated as they evolve or adapt to their environment) [19, 63]. A classic paradigm to formalise the internal architecture of such complex agents is the BDI (Beliefs - Desires - Intentions) explored by Bratman in 1987 [19, 26, 143]. The BDI approach represents the way agents can do complex reasoning. It aims at disambiguate various concepts (those of belief, desire and intention) and the logical relationships between them. A classic framework of BDI is the procedural reasoning system (PRS). This last includes three main processes: the perception (in which the agent acquires information from the environment), the central interpreter (which helps the agent to deliberate its goals and then to select the available actions) and the execution of intention (which represents the agent reactions). Several extensions have been developed, such as allowing the model to add to agents a set of beliefs (information it gets by perception or communications) and intentions (what it wants to execute), and ways to manage the two sets. Such approaches are very powerful, but remain computer scientist-oriented, as they require high programming skills to develop bridges between the framework and the platforms, and to write agents' behaviours without a dedicated modeling language [26]. Last, the scalability of the tools must be taken in consideration, as hundreds of agents (and resulting interactions and environmental changes) must be manageable by the program.

Caillou et al. [26] clarify that, in the BDI approach, beliefs ("what it thinks") correspond to the internal knowledge the agent has about the world. Desires are the objectives the agent would like to accomplish ("what it wants"). Like the Belief base, the Desire base is updated during the simulation and desires can be related by hierarchical links, when a desire is created as an intermediary objective. Desires are also ranked by priority, and these priorities may evolve during the simulation. Finally, intentions are what the agent has chosen to do ("what it is doing"). The current intention will determine the selected plan and intentions can sometimes be put in stand-by. The agent behaves according to two steps: its perception and the setting of a plan. The perception is a function called at each iteration, where the agent can update its bases of beliefs and desires. The plans are behaviors defined to reach specific desires, that can be instantaneous or persistent, and that can be ranked by priority [26].

It is important to note that, according to An et al. [3], ABMs are usually not appropriate if the starting point is a mass of raw data. Cap-dependent translation is a

#### 4. Understanding of the translational regulation by uORFs may be improved by mathematical modeling – 4.1. Agent-based modeling may help deciphering complex mechanisms

complex mechanism which involves numerous actors and events. Here, I propose to combine the strengths of **ABMs** with the availability of large-scale data about the **sORFs** registered in MetamORF. The rationale is to model translational mechanisms by taking advantage of **ABMs** to study a complex system based on experimental literature and the existing **ICIER** model; but to integrate also actual human transcripts with their **sORFs** in the simulation (instead of using "artificial" transcripts). This led me to develop **agent-based modeling of uORF cis regulatory functions informed by experimental data (ABMCisReg)**, an agent-based model of **uORF cis** regulatory functions informed by experimental data.

#### 4.1.4. Agent-based modeling allowed to implement a new model of translational regulation by the uORFs

##### 4.1.4.1. ABMCisReg is an ABM allowing to easily model uORF cis regulation and taking into account novel features

In **ABMCisReg**, ribosomes were represented as agents, whilst transcripts (along with their **uORFs** and **CDSs**) were representing their environment. Using a **BDI** paradigm as describe above, I was able to model the complex interactions between agents and their environments (*e.g.* recognition of an **uORF** start codon by a ribosome) as well as among agents (*e.g.* collision between an upstream elongating ribosome and a downstream scanning one). In the frame of a **BDI** model, the ribosome's beliefs are the following assessments (that can be answered with booleans): there is a start codon at the current location, the stop codon of the **ORF** being translated has been reached, there is already a ribosome occupying the next position and the **CDS** start codon has been reached or is located upstream. Ribosome's desires can be considered as either translate (with the sub-desires: initiate translation, elongate and terminate translation), search for a start codon, leave transcript or queue. Finally, the agent may have the following intentions to accomplish its tasks: fix the **5'UTR** of a transcript, move forward, move backward, stall at position, initiate the translation (*i.e.* turn into elongating ribosome), terminate the translation, reinitiate (*i.e.* turn back into a scanning ribosome) and leave the transcript. The Figure 4.2 provides the decision tree that is used by the agents to make decisions.

It should be noticed that in **ABMCisReg**, the system is updated using an asynchronous process, from 3' to 5' end (*i.e.* from right to left). This means that the agents located the nearest to the 3' end (*i.e.* the most on the right on the lattice) perform their action first (and consequently may update the environment before an upstream agent takes information from its environment). For instance, considering agents located on two consecutive places on the one-dimensional grid, if the downstream agent moves forward (*i.e.* towards the right), then the place will be free for the upstream agent during the same pseudo-time step.

Specific **ABM** software environments and toolkits have been developed (such as

4. Understanding of the translational regulation by uORFs may be improved by mathematical modeling – 4.1. Agent-based modeling may help deciphering complex mechanisms

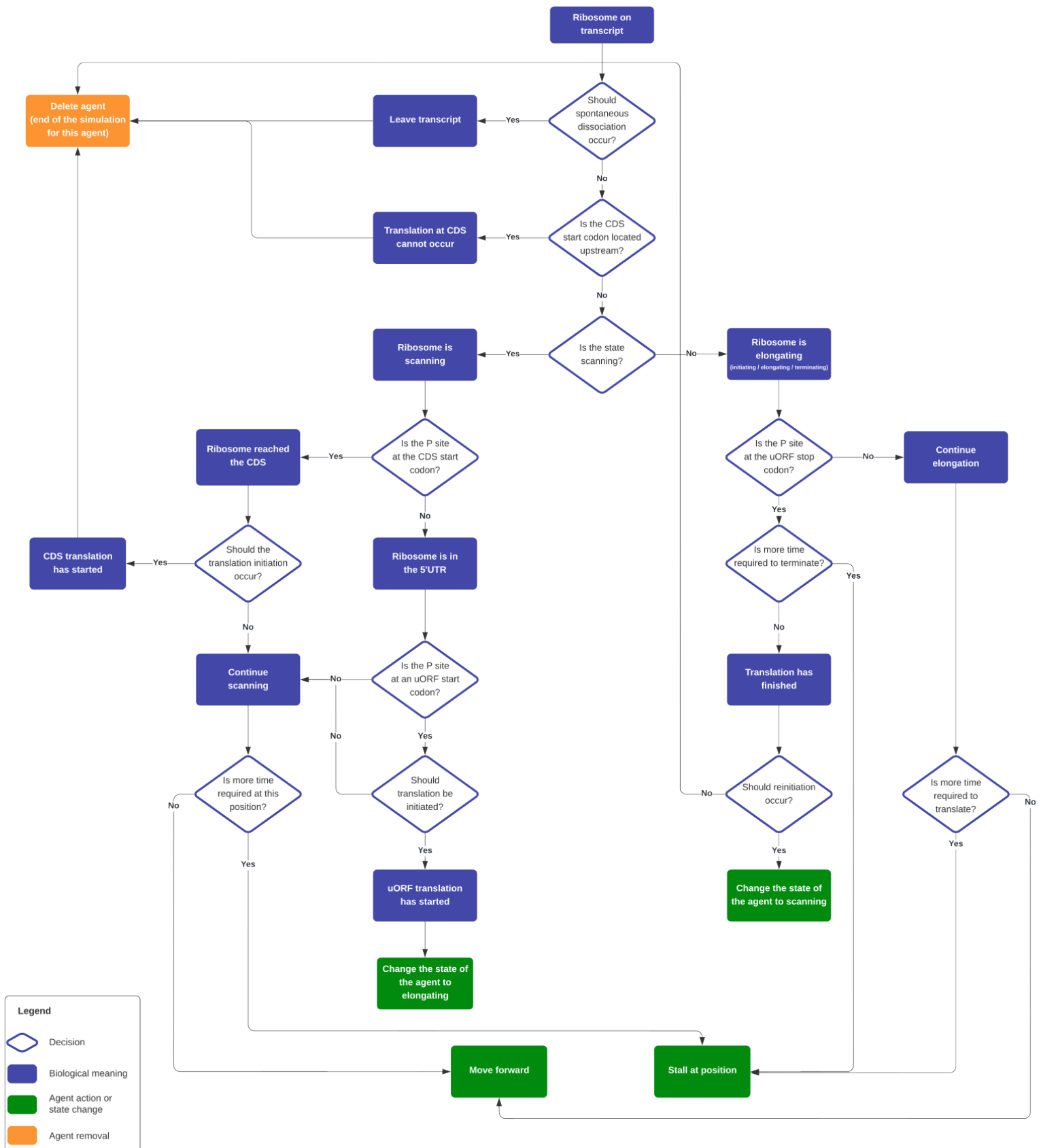


Figure 4.2.: Decision tree of the agent-based modeling of uORF *cis* regulatory functions informed by experimental data (ABMCisReg).



#### 4. Understanding of the translational regulation by uORFs may be improved by mathematical modeling – 4.1. Agent-based modeling may help deciphering complex mechanisms

Repast, Swarm, MASON, NetLogo) and offer user-friendly environments for the development of [ABMs](#) by non-computer scientists [3, 19, 26]. However, they show a limited adaptability and fail to develop systems as complex as the cytoplasmic translation, in particular when one is willing to make it adaptable for grid search<sup>1</sup>. Hence, [ABMCisReg](#) has been developed using object-oriented programming in Python (v3.5) to tackle these issues. Docker and Singularity environments have been used in order to ensure reproducibility and to facilitate deployment on high-performance clusters. Statistical analyses have been performed with R and Snakemake was used to facilitate grid search.

##### 4.1.4.2. [ABMCisReg](#) could be improved in the future by implementing novel features

Despite the implementation of many novel features in our [multi-agents system \(MAS\)](#), it should be noticed that some features have not yet been integrated to our model at the current stage.

As an example, elongation rates are still constants in our modeling, still this is something that can be easily improved. Tian et al. [135] recently developed RiboMIMO, a deep learning based method that models the translation elongation rates of full-length transcripts. Hence, I suggest to implement the outputs of this tool into [ABMCisReg](#) in the future. As variations in elongation rates are susceptible to change the collisions between elongating and scanning ribosomes, this may have a major impact on the simulations. Pavlov et al. [100] recently developed a model that accounts for the local codon context-dependent variation of peptide elongation times and [RPF](#) generation biases. They demonstrated that a local context of five codons (including those at the A, P and E sites) accounted for the ribosomal dwell time on each A-site codon of the transcriptome. This finding could also be easily integrated in [ABMCisReg](#) in the future.

It is important to keep in mind that [RNA](#) secondary structure may affect the translation [32], a parameter that has not been taken in consideration in this study. As an example, the initiation at a start codon located in a suboptimal context (*i.e.* not a Kozak sequence context) is facilitated by the downstream presence of a secondary structure moderately stable (*e.g.* a stem-loop), making slower the initiation complex [67]. Strong stem-loops just downstream of the start codon are also susceptible to stall the scanning ribosome, increasing its stalling time and thus the probability of leaky scanning through near-cognates or AUG triplets in poor contexts [54]. Recent findings also suggest that secondary structures of [RNAs](#) affect elongation rates [135]. As a consequence, integration of secondary structures in [ABMCisReg](#) should be considered in the future. However, if stem-loops can now be predicted by softwares such

---

<sup>1</sup>Grid search is a technique used to determine the optimal hyperparameters for a model (*e.g.* to find the best values for probabilities). It performs by running the model with all possible combinations of parameters (provided as sets of discrete values instead of an unique one) and by checking how well the model fits to the data (*e.g.* using metrics such as area under the curve).

4. Understanding of the translational regulation by uORFs may be improved by mathematical modeling – 4.1. Agent-based modeling may help deciphering complex mechanisms

as SPOT-RNA [125], integration of such data in the model is still hard, as the effect of the secondary structure on scanning and elongating rates has not been yet fully elucidated.

Additionally, Schott et al. [122] demonstrated that the loading process of ribosomes on novel transcript is susceptible to be progressive, *i.e.* that the transcript accumulates progressively ribosomes after its export in the cytoplasm. This suggests that mRNAs may benefit from recycling of ribosomal subunits and translation factors concentrated in their vicinity. They also demonstrated that a newly transcribed mRNA cannot reach its full potential after the first round of translation, but needs to build up a pool of ribosomal subunits and translation factors that are concentrated in the vicinity of the mRNA molecule [122]. As our model is based on a progressive loading of ribosomes on the mRNA, this constitute an important limitation, and suggest we consider all transcript as newly synthesized. A future improvement of ABMCisReg could consist in using a finite pool of ribosomes in the vicinity, and ensure the conservation law (*i.e.* a scanning ribosome leaving the transcript because of a collision, after the termination of the translation or spontaneously, would be available for fixing again the transcript).

Finally, pepto-switches are mechanisms inhibiting downstream coding sequence expression by blocking the ribosomes through direct or indirect small molecular activation or interaction of the nascent peptide with the ribosomal complexes [147]. Such features are impossible to implement in models at the current stage, as it would require to know exactly the function of all uORF-encoded peptides.

We also explicitly decided to omit the mechanisms of scanning-free translation (usually expected to occur on transcript with longer 5'UTRs [11]). However, implementation of such translational process, for example through the integration of IRESs, could change our understanding of the translational regulation. Indeed, IRESs have been shown to be particularly involved in stress and apoptosis signalling [138]; and databases such as the IRES Atlas [157] already register their location on the transcriptome.

To conclude, the activity of uORFs varies considerably across cell types and cell conditions [84], which complicates the development of global models of regulation of the translation by uORFs, as different stress conditions are susceptible to trigger different translational responses. It should also be kept in mind that processes of translation initiation, elongation and termination at translated uORFs are maintained by a selective pressure [74], and many of the processes and features evoked above are likely to be also under selective pressure. Hence, I suggest that when the exploitation of ABMCisReg will be advanced enough to identify clearer mechanisms of regulation of the translation by the uORFs, it would be of particular interest to check if such mechanisms can be also identified in other species. In particular, it should be noticed that MetamORF provides information about sORFs identified in the mouse genome, in the exact same format as for the human. This should make easier the application of



4. *Understanding of the translational regulation by uORFs may be improved by mathematical modeling – 4.1. Agent-based modeling may help deciphering complex mechanisms*

the [ABMCisReg](#) model to this particular species at first.

## 5. Concluding remarks, limitations and perspectives

The size and complexity of most eukaryotic proteomes have thus probably been greatly underestimated so far [117], assuming that the complexity of an organism is dictated by the increase of its proteome's diversity. However, if the number of sPEPs was underestimated for historical reasons, the existence of artifacts in recent methods probably caused the number of sPEPs to be overestimated [133]. Indeed, many sORFs were identified during the past years, and MetamORF gathers information about 664,771 unique sORFs for *H. sapiens*. This repository of sORFs identified by complementary methods has the advantage (i) to provide a repository of unique sORFs identified in humans and mouse as comprehensive as possible and (ii) to give the opportunity to the end-user biologists to select only the sORFs that have been identified by several complementary methods (computational prediction, Ribo-seq and/or MS) and/or identified in several distinct original studies.

Using a system approach, I was able to identify some of the key functions fulfilled by putative sPEPs and to perform a large-scale annotation of peptides that may be encoded by sORFs in human monocytes. To the best of my knowledge, this large-scale interactome of sPEPs with canonical proteins is the first to be inferred. In addition to the analyses I run, I expect it to be a resource of interest for further investigation. By looking at the most commonly used interfaces of interactions and taking advantage of the topology of the network for clustering and sPEP annotation, I provided clues that sPEPs are likely to be involved in metabolic processes. The results suggest they are able to take part to many pathways, in particular related to the metabolism of the cell. Interestingly, annotations have been computed for some intergenic sORFs identified by Ribo-seq, which suggest that even this class of sORFs may have *trans* functions (assuming they are actually translated into stable peptides). It is clear that future efforts are required to prove that these peptides accumulate in the cell at significant levels and more detailed functional studies, in particular experimental low-scale studies, will be needed to validate the predicted functions. However, regarding this huge pool of novel peptides, the big number of interactions predicted and the fact that functions have already been demonstrated by multiple studies for some sPEPs, we may reasonably conclude that sPEPs (and consequently sORFs) are actual functional elements of the genome. Although one may still argue about the real proportion being actually functional, it is now for sure that studying sPEPs is relevant and should probably bring a full set of interesting discoveries, with potential applications. Because much more is known about canonical proteins (such as the SLiMs or domains mediating

## 5. Concluding remarks, limitations and perspectives

interaction), it makes sense to take advantage of this knowledge to study this novel class of peptides. However, one must remain cautious, as it is also not unlikely that some properties are specific to short peptides or even specifically **sPEPs**, and thus not shared with canonical proteins. At the moment, such specificities have not yet been identified, but if it came to be the case, then these properties would help scientists making the difference between canonical proteins and **sPEPs**, other than based on historical and semantic reasons.

One important topic which has been omitted in this thesis is the mechanisms by which **sPEPs** are degraded. Life time of proteins and peptides is of major importance in the regulation of homeostasis, as the degradation of proteins usually help to end some particular pathways or signals. Indeed, given that many **sPEPs** experimentally identified so far are thought to function as regulators induced by specific conditions, we may expect that mechanisms that downregulate their levels or activities when they are no longer needed exist. These may notably involve binding of small molecules, amino acid modifications or regulated proteolysis for instance [48].

Despite not being able to explain the process of regulation of the translation by **sORFs**, the agent-based system I developed (**ABMCisReg**) should provide a solid and easily-adaptable computational tool for studying it in the future. This tool mainly remains to be exploited, and I expect that the concurrence of experts in the fields of **sORF** biology, translational regulation and ribosome biology will help mature this model and find the most appropriate ranges of values for the parameters tested.

However, one must keep in mind that translational changes go hand in hand with regulation of the transcription, **mRNA** export and stability, protein stability and degradation, all of which determine the final protein output [122]. Disentangling the contribution of each individual process has always been one of the greatest challenge of biology, and the discovery of **sORF** regulatory functions brought a novel layer of complexity.

Clearly, mechanisms of regulation of the translation by the **uORFs** remain cryptic, and even the **ATF4**-like mechanism is largely debated and does not fit well to most transcripts. It is clear that the number of **uORFs** does not success alone to explain such complex regulation. In addition, they are so many parameters that are likely to be important for this regulation that it is totally unlikely that a model will perfectly explain the stress-resistant behavior of some **CDS** before a long time. Because **uORFs** are likely to take part in the the regulation of protein levels during changes in the cellular identity along development trajectories, it could also be of interest to look at other conditions than stress/non-stress conditions, considering that the availability of the **ternary complex** may also be affected by other parameters than the availability of **eIF2 $\alpha$** .

I must also emphasize that methylation levels of the transcripts were not discussed here, whilst it is known that methylation of the **RNAs** may slow down the scanning rate, probably by affecting the translation of their **ORFs**. As an example, **uORF2** of the **ATF4 mRNA** has been reported to undergo specific demethylation following activation

## 5. Concluding remarks, limitations and perspectives

of the **ISR** [109, 115], a process that probably participates to the regulation of the translation. Another point is that only cap-dependent mechanisms of translation have been considered at the moment in this model. In the lack of evidence, we may hypothesize that **uORFs** may also play a role in the regulation of cap-independent translation processes. Additionally, other **5'UTR** regulatory elements have been described, despite better characterization is still required, in particular in eukaryotes. We may notably cite the existence of **IREs** as well as of **5' terminal oligopyrimidine (5TOP)** motif that allows **mTOR**-dependent stimulation of the expression of proteins of the translational machinery [54]. Viral **mRNAs** have also been reported to harbor stretches of unstructured nucleotides in their **5'UTR** that can bypass the requirement for the **m<sup>7</sup>-methylguanosine (m<sup>7</sup>G)** cap and the **eIF4F** initiation factor [54]. For instance, **m<sup>6</sup>A** modifications of **mRNAs** have been reported under cellular stress and shown to lead to efficient translation under conditions of suppressed cap-dependent translation [115].

In addition, termination at an **uORF** stop codon can elicit the same **mRNA** destabilization evoked by the **nonsense-mediated decay (NMD)** pathway at premature termination codons in **ORFs** [54], a **mRNA** degradation pathway that eliminates transcripts containing premature termination codons and which is notably inactivated during the **ISR** [119]. **NMD** has not been discussed in this manuscript, whilst **sORF** probably play a role in activating this mechanism. As an example, **uORFs** in plants have been shown to trigger **NMD** in a size-dependent manner. The **uORFs** encoding **sPEPs** longer than 50 **aa** activate **NMD** responses, whereas shorter **uORFs** do not activate such responses. However, not all **uORF**-containing **RNAs** are sensitive to **NMD** and the features that distinguish those triggering **NMD** responses remain to be determined [97].

Finally, other classes of **sORFs** have been omitted, whilst we now that **3'UTR** sequences may change the structure of **RNAs** and consequently the fixation or translocation rate of ribosomes. Because no model can reasonably account for so many parameters, some particular cases will necessary not be explainable, and it is only reasonable to look first for models that success at explaining the regulation for the largest proportion of **RNAs** (or on contrary of specific cases).

So far, most studies on **sORFs** focused either on their *cis* or *trans* function. There is an urgent need to check if these roles are connected, as a dual role for a single **sORF** cannot be excluded [25, 163]. In particular, it has been demonstrated that regulatory nascent peptides (identified in bacteria and eukaryotes), encoded by **uORFs**, have the capacity of arresting their own translation, either at the elongation or the termination step, by interacting with the ribosomal components [116, 167]. Translated **uORFs** could also generate functional peptides that directly or indirectly regulate expression of the **CDS** [132]. My findings suggest also that **sPEPs** are involved in **RNA** metabolism, which makes them likely to take part in the regulation of the translation. It is thus likely that (some) are playing both *cis* and *trans* regulatory functions, that may be connected or not, and play a major role in the homeostasis of the cell. Because I noticed that 6 % of the **sPEPs** are targeting their near-cognate **CDS** (*i.e.* the one

## 5. Concluding remarks, limitations and perspectives

harbored by their own transcript), I think this would be of major interest to perform low-scale, advanced characterization of these particular sORFs at first. In addition, I suggest that extensive characterization of the functions of sORFs encoded by the same transcript and expressed at meaningful levels would be also of great importance to identify if operon-like systems exist in eukaryotes.

Ultimately, experimental evidence and full functional characterization is the only way to ascertain the translation and function of each individual sORF [10], a tremendous task that will necessarily require to pursue our efforts on the study of sORFs and their peptides in the future.

Last but not least, because of their involvement in fundamental cellular processes and in the etiology of many diseases, I personally think that many biotechnological applications related to sORFs and sPEPs are coming in a near future.

Gene editing methods (CRISPR-Cas9) have already been used successfully for the development of stress-tolerant crops, by editing uORFs in order to control the translation of downstream ORFs. As another example, editing of an uORF of *LsGCP2* which encodes an enzyme involved in vitamin C biosynthesis in plants has been reported to enhance oxidative stress tolerance and elevated ascorbate levels [137].

Beaudoin *et al.* [17] also reported the presence of many sORFs on mRNA vaccines that have been recently developed in response to Covid-19 pandemic. Because the number and nature of these sequences highly variate between the wild-type sequence and those of mRNA vaccines (Moderna mRNA-1273 and Pfizer BNT162b2), we may also wonder the impact on human health of the sORFs newly introduced in these vaccines and their eventual sPEPs. We notably started a project in collaboration with B. Nal-Rogier to address this question, by using a system approach similar to the one described in this thesis (sPEPRI prediction with mimicINT etc.). Because these non canonical sORFs vary between the vaccines, they may theoretically encode different sPEPs, and thus interactions with canonical proteins may be lost, gained or get their affinity modified. I would guess that these difference among vaccines (and between the natural and the vaccine sequences) may (partially) explain differences of efficiency as well as side effects and should thus be carefully considered in drug-development in the future. In parallel, recent Ribo-seq data applied to SARS-CoV-2 found evidence of 23 novel proteins, beyond the 37 already annotated for the virus [133], suggesting we may have missed many host-pathogen protein-protein interactions so far.

To conclude, sORFs editing and targeting may present application of particular interest in many fields of the biology, including notably agronomy and health industry [137], fields that cannot be dissociated, in particular regarding the emerging concept of *One Health*. Because they represent a full novel class of potential drug targets and seem to be involved in many (unrelated) diseases, I am personally convinced that sORFs and their peptides will no longer be ignored for long, and that new classes of drugs are going to result from their extensive study.

## A. Article: In depth exploration of the alternative proteome of *Drosophila melanogaster*

B. Fabre, S. Plaza and their colleagues recently optimized MS-based approaches to identify sPEPs in *D. melanogaster*. In the frame of a collaborative project and using the same computational tools as the ones described in chapter 3, I helped for the functional characterization of 401 (yet unannotated) sPEPs by looking at the SLiMs and domains they harbor. We notably noticed that most sPEPs contain disorder regions. In addition, the majority of SLiMs retrieved belong to the post-translational modification sites (MOD), ligand binding sites (LIG) and docking sites (DOC) class types. The SLiMs the most represented suggested a possible role of sPEPs in cell cycle and autophagy. Surprisingly, the SLiM class type targeting sites for subcellular localization (TRG) were only identified in one single sPEP produced from an uORF. We also demonstrated that sPEPs produced from dORFs are less susceptible than other to carry particular functions based on domain prediction, and we suggested that the ORF itself might be involved in the regulation of the translation, something that would be interesting to explore in the near future but that is out of the scope of my thesis.

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, Lilley KS, Plaza S (2022). In depth exploration of the alternative proteome of *Drosophila melanogaster*. *Frontiers in Cell and Developmental Biology*, 10:901351. eCollection 2022.



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

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

<sup>1</sup>Laboratoire de Recherche en Sciences Végétales, UMR5546, Université de Toulouse, UPS, INP, CNRS, Auzeville-Tolosane, France, <sup>2</sup>Cambridge Centre for Proteomics, Cambridge Systems Biology Centre and Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom, <sup>3</sup>Aix-Marseille Université, INSERM, TAGC, Turing Centre for Living Systems, Marseille, France, <sup>4</sup>Fédération de Recherche (FR3450), Agrobiosciences, Interactions et Biodiversité (AIB), CNRS, Toulouse, France, <sup>5</sup>Institut de Pharmacologie et de Biologie Structurale (IPBS), Université de Toulouse, CNRS, UPS, Toulouse, France, <sup>6</sup>Infrastructure Nationale de Protéomique, ProFI, FR 2048, Toulouse, France, <sup>7</sup>Cambridge Systems Biology Centre and Department of Genetics, University of Cambridge, Cambridge, United Kingdom, <sup>8</sup>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,  
Bulet-Schiltz O, Russell S,  
Combiér 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  $\mu$ l equivalent of embryo per replicate), adult flies (10 adult flies per replicate), or S2 cell pellets ( $5 \times 10^8$  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  $\mu$ g 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  $\mu$ l 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.
- Embryos (100  $\mu$ l 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  $\mu$ g 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.
  - Embryos (100  $\mu$ l 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<sub>2</sub> 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  $\times$  10<sup>8</sup> 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.
  - 5  $\times$  10<sup>8</sup> 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.
  - 5  $\times$  10<sup>8</sup> 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. Coverslips 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.

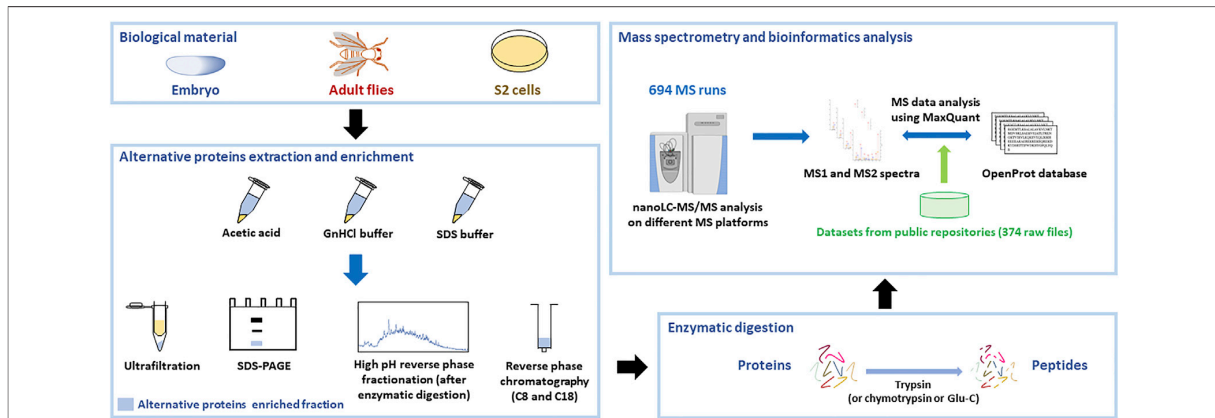
For each class of motif, and for each class of AltProt (ncRNA, isoform, 5'UTR, CDS, and 3'UTR), enrichment and depletion in AltProt with at least one motif of the class among the AltProt of the class have been assessed, using one-sided Fisher's exact tests. The *p*-values computed have then been adjusted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

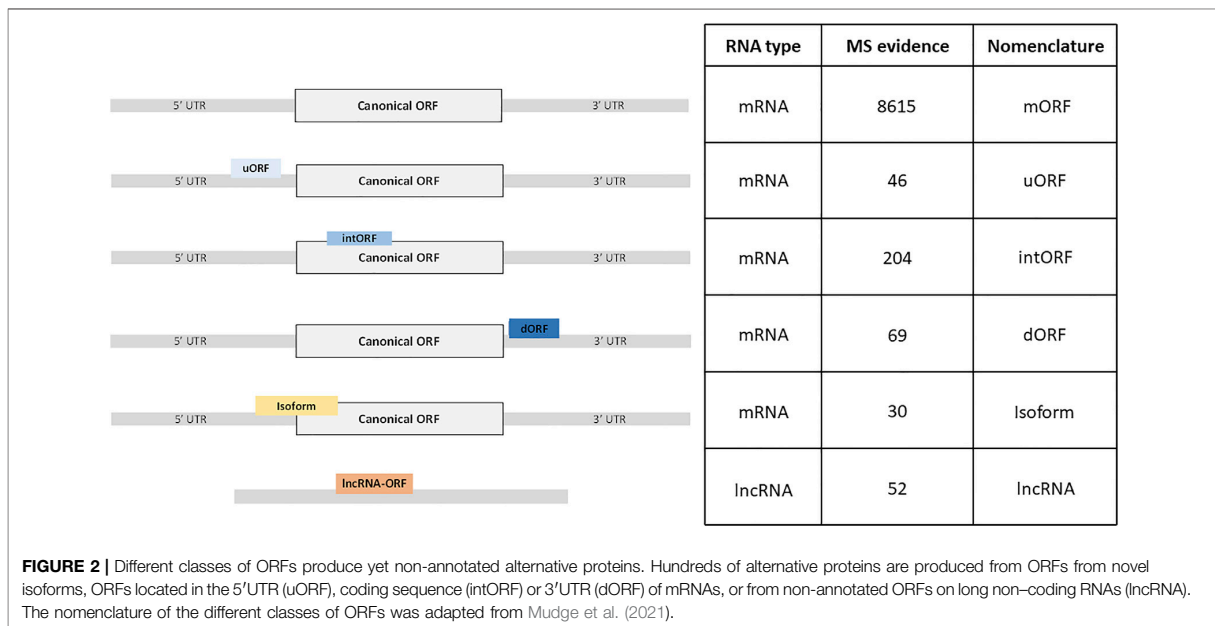
# A. Article: In depth exploration of the alternative proteome of *Drosophila melanogaster*

Fabre et al.

Alternative Proteome of *Drosophila melanogaster*



**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.



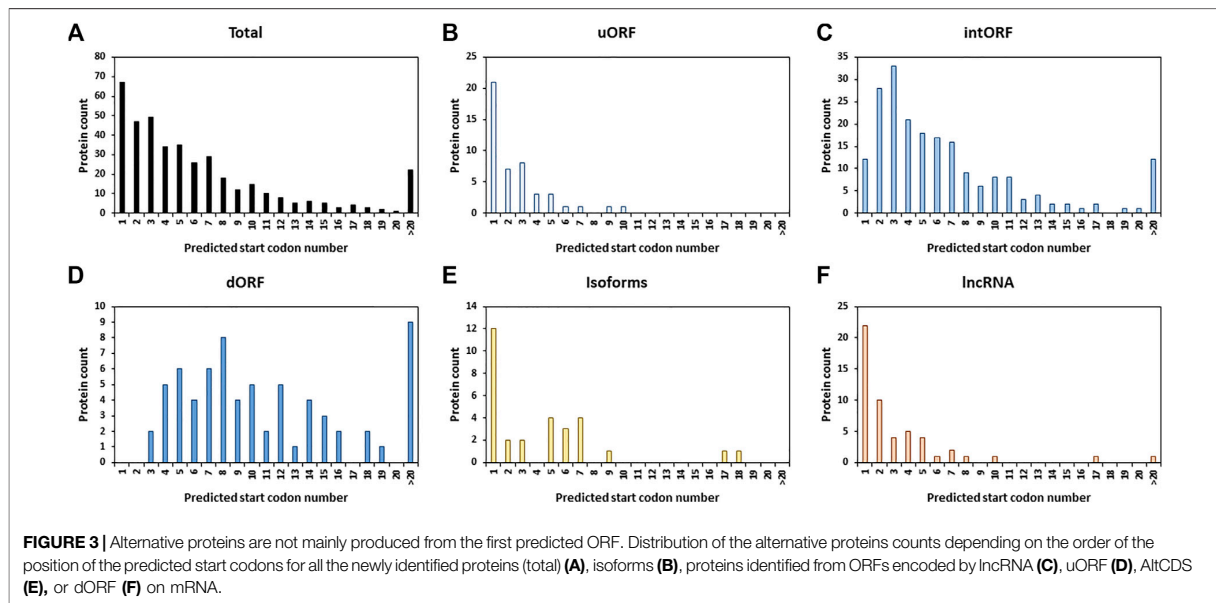
**FIGURE 2 |** Different classes of ORFs produce yet non-annotated alternative proteins. Hundreds of alternative proteins are produced from ORFs from novel isoforms, ORFs located in the 5'UTR (uORF), coding sequence (intORF) or 3'UTR (dORF) of mRNAs, or from non-annotated ORFs on long non-coding RNAs (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

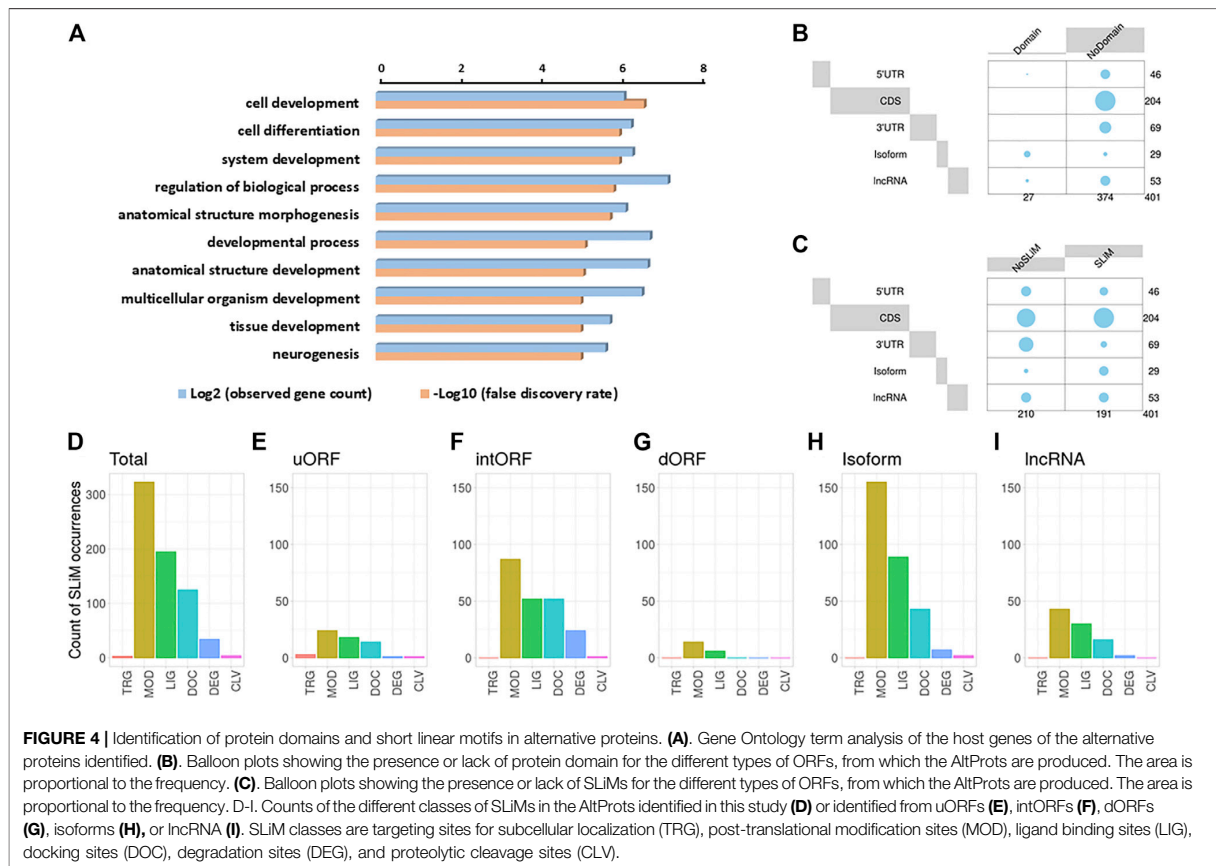
by mass spectrometry (Ma et al., 2016; Cardon et al., 2020). Extensive fractionation, using high pH reverse-phase chromatography, as well as specific enrichment of short proteins, through SDS-PAGE, ultrafiltration, acid precipitation, and reverse-phase chromatography, was employed to retrieve AltProts 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 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



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 ( $FDR = 2.4 \times 10^{-7}$ ) and cell differentiation ( $FDR = 9.6 \times 10^{-7}$ ), 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 =  $3.5 \cdot 10^{-4}$ , and odds ratio = 5.41), ligand binding sites (LIG) (especially enriched in isoforms, Benjamini-Hochberg adjusted  $p$ -value =  $1.42 \cdot 10^{-7}$ , 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\_PIK\_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

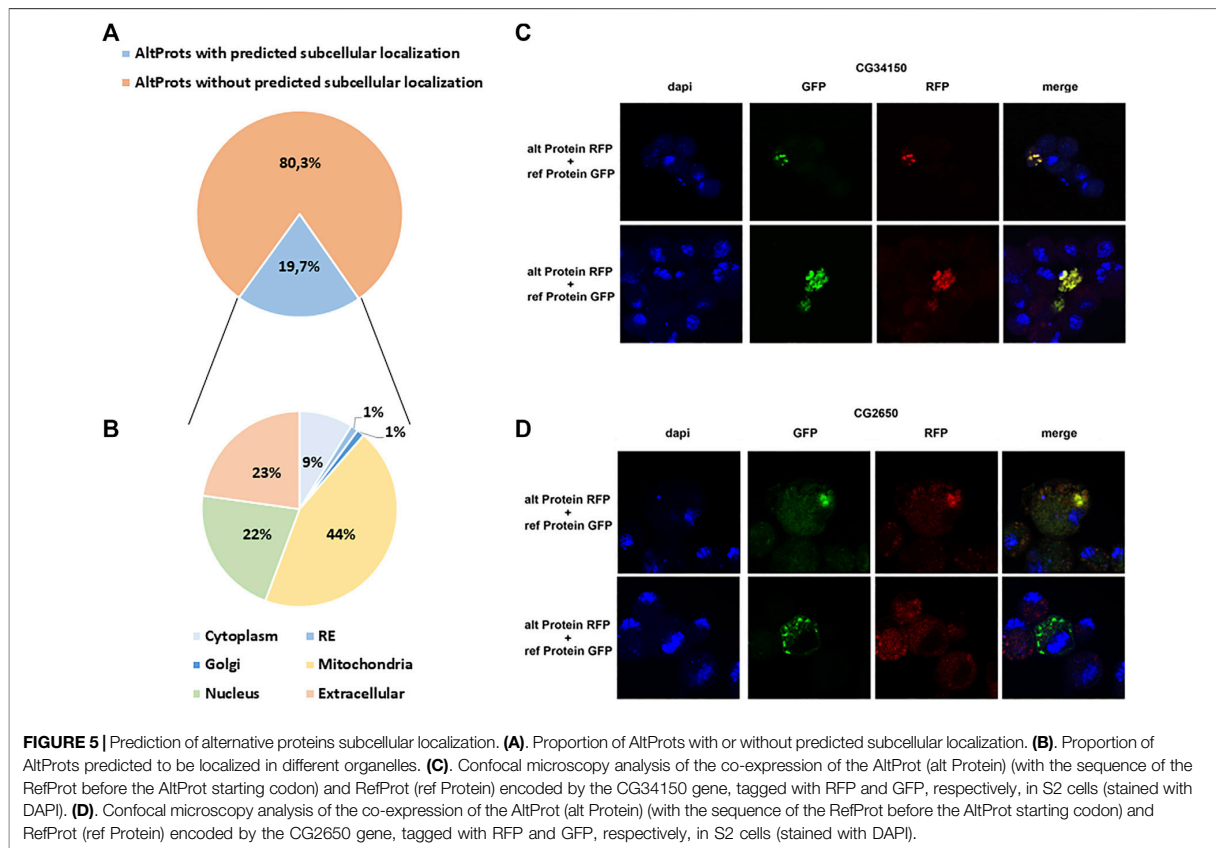
(55.9%) tend to have disordered regions compared to lncRNA (38.5%) or dORFs (26.1%) (Supplementary Figure S12). This is in agreement with a recent report on plants, which also predicted that numerous non-annotated short proteins might contain disordered regions, transmembrane domains, or signal peptides (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).



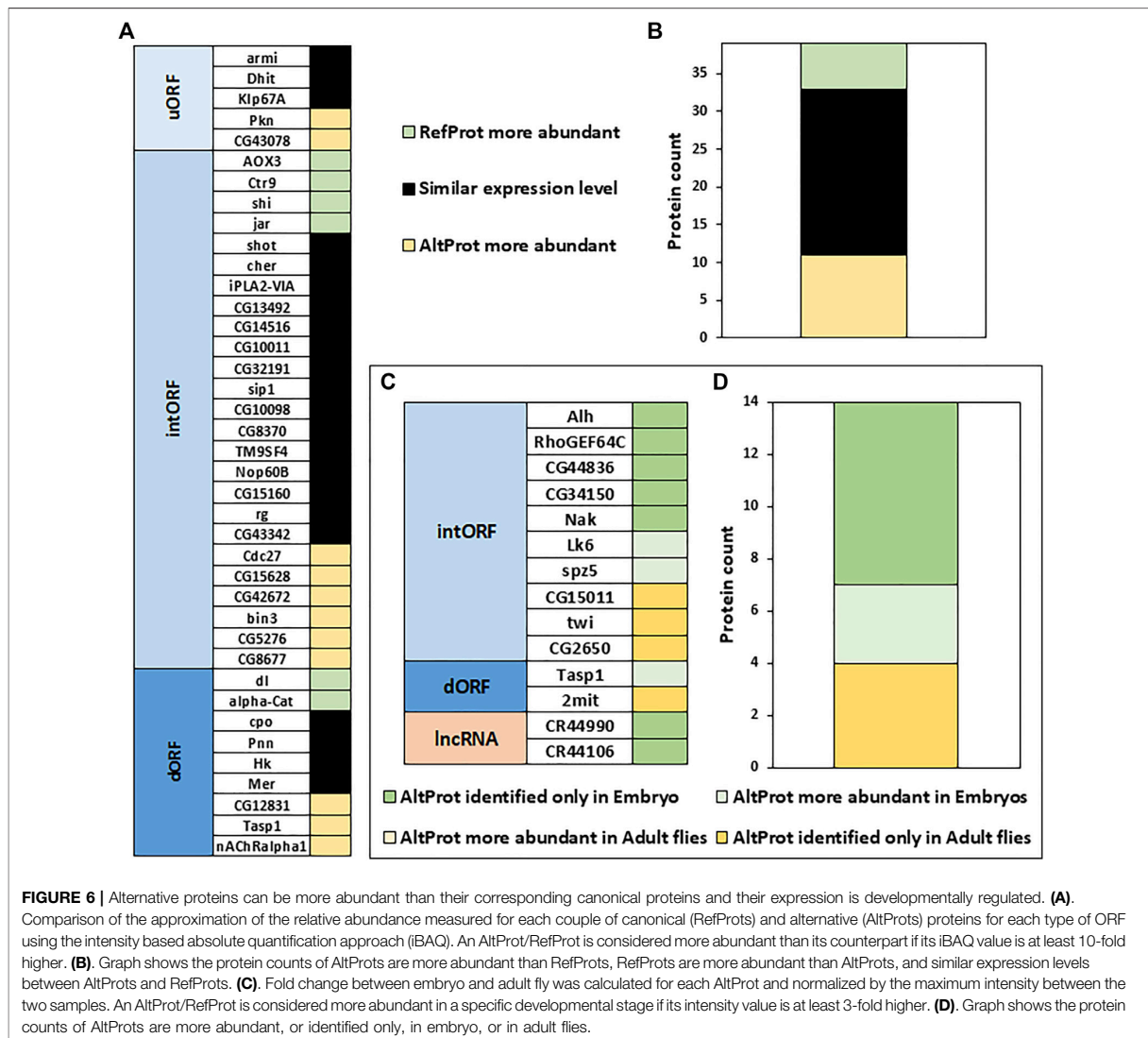
Comparing the intensities measured for peptides from AltProts and RefProts from total lysates and high pH reverse-phase fractionation (no specific AltProts enrichment, see Material and Methods section protocol 2 as well as data from Müller et al. (2020), the peptides from the latter were slightly more 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 than their corresponding RefProts (**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.



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.

## 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

## 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>

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

# A. Article: In depth exploration of the alternative proteome of *Drosophila melanogaster*

Fabre et al.

Alternative Proteome of *Drosophila melanogaster*

- 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., Kandasamy, 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., Salzter, 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., Lubner, 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., Salzter, 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, Á. (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., Combiere, 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., Kočevár 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
- Hrabec, 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., Müller, 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

# A. Article: In depth exploration of the alternative proteome of *Drosophila melanogaster*

Fabre et al.

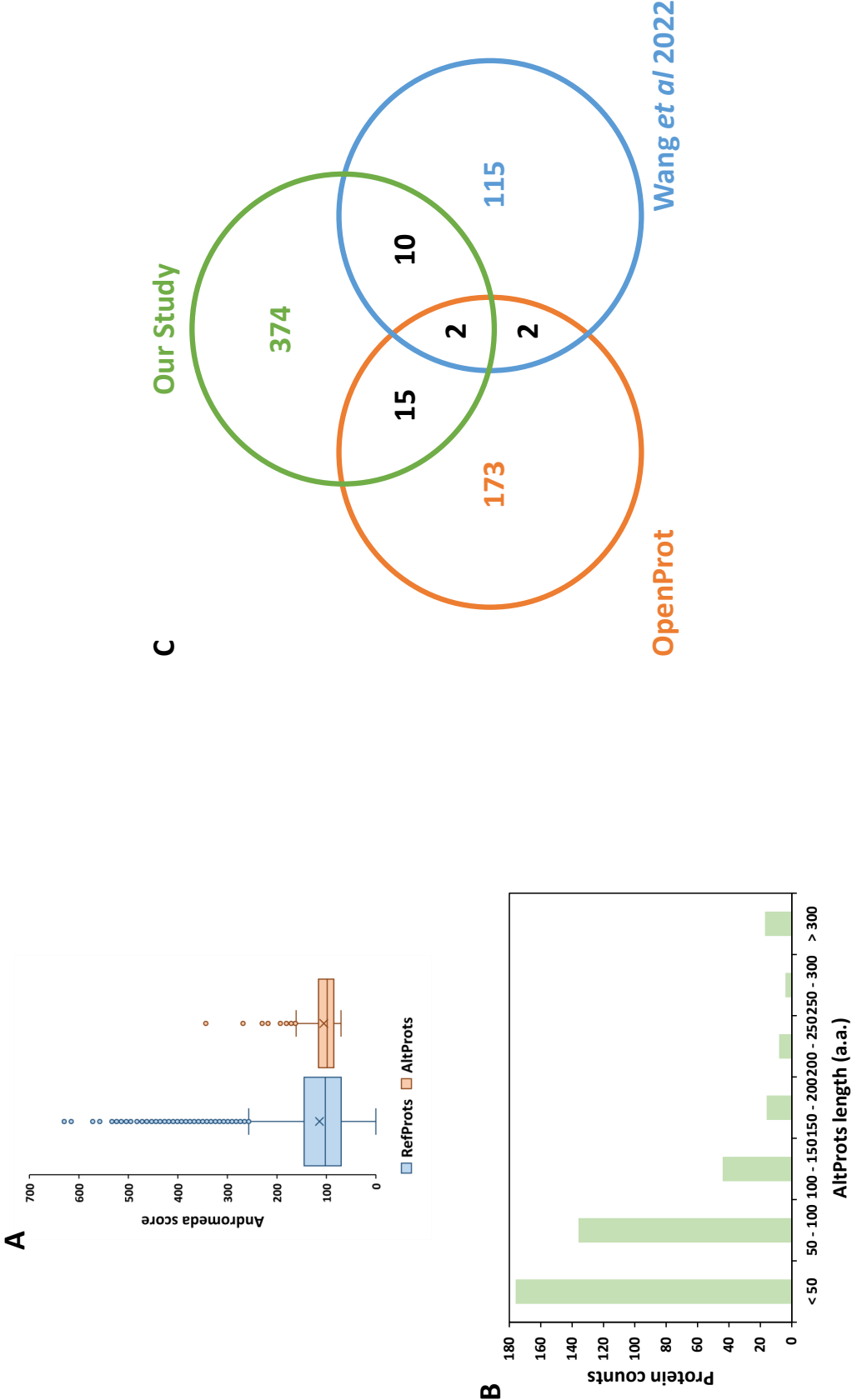
Alternative Proteome of *Drosophila melanogaster*

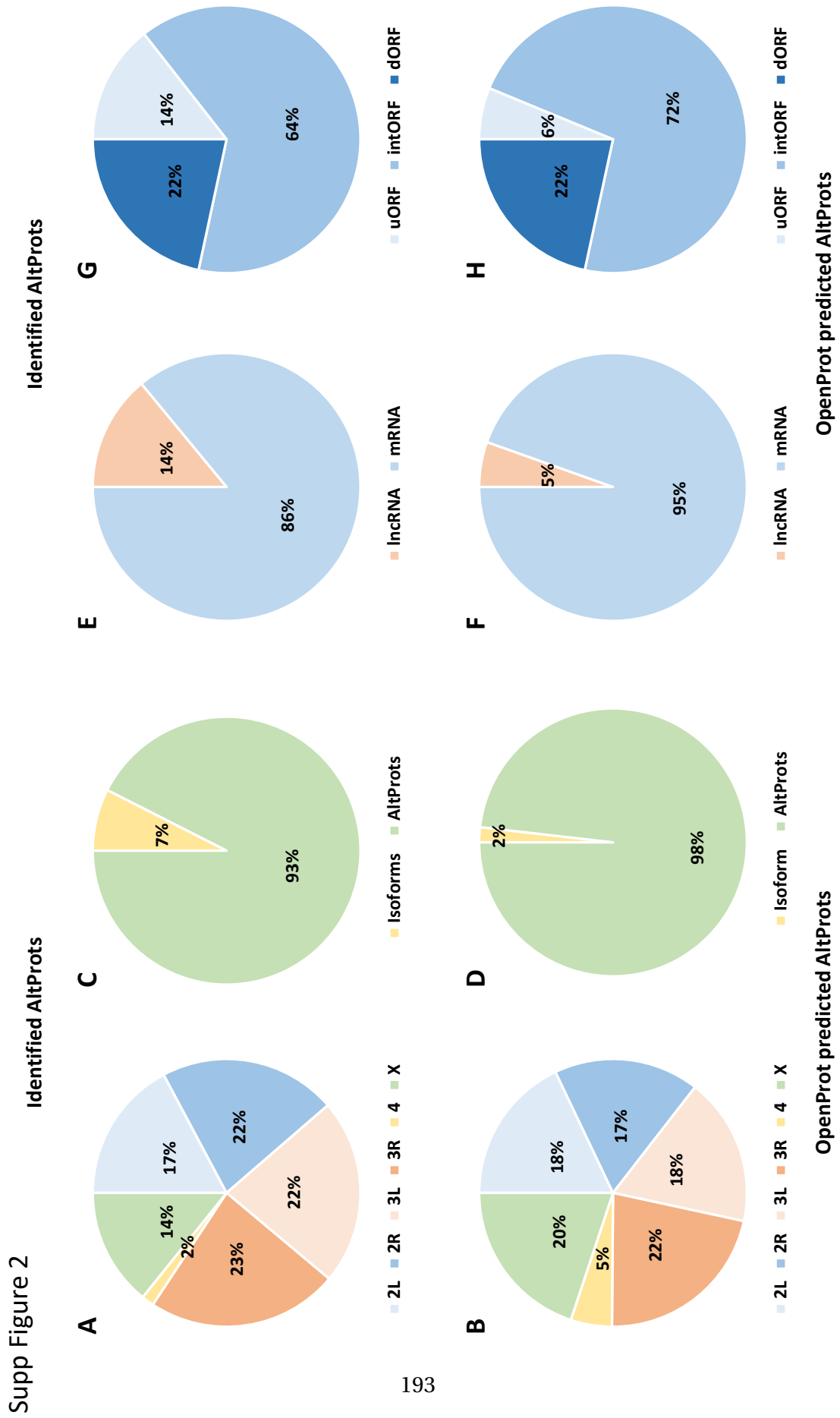
- Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., et al. (2021). Pfam: The Protein Families Database in 2021. *Nucleic Acids Res.* 49, D412–D419. doi:10.1093/nar/gkaa913
- Montigny, A., Tavormina, P., Duboe, C., San Clémente, H., Aguilar, M., Valenti, P., 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/emboj.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.
- Copyright © 2022 Fabre, Choteau, Duboé, Pichereaux, Montigny, Korona, Deery, Camus, Brun, Burlet-Schiltz, Russell, Combier, Lilley and Plaza. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

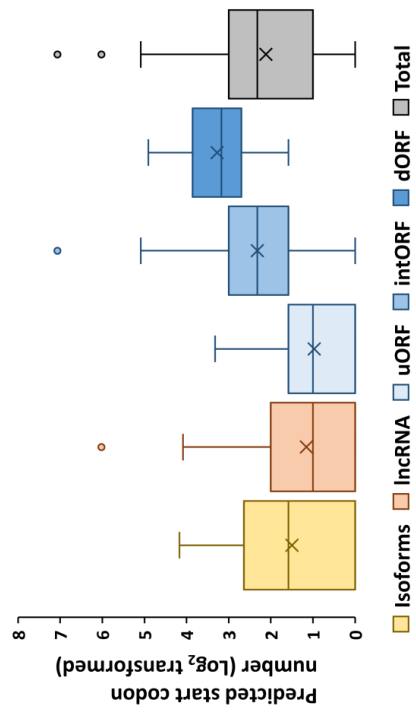
*A. Article: In depth exploration of the alternative proteome of Drosophila melanogaster*

**SUPPLEMENTARY MATERIAL**

Supp Figure 1

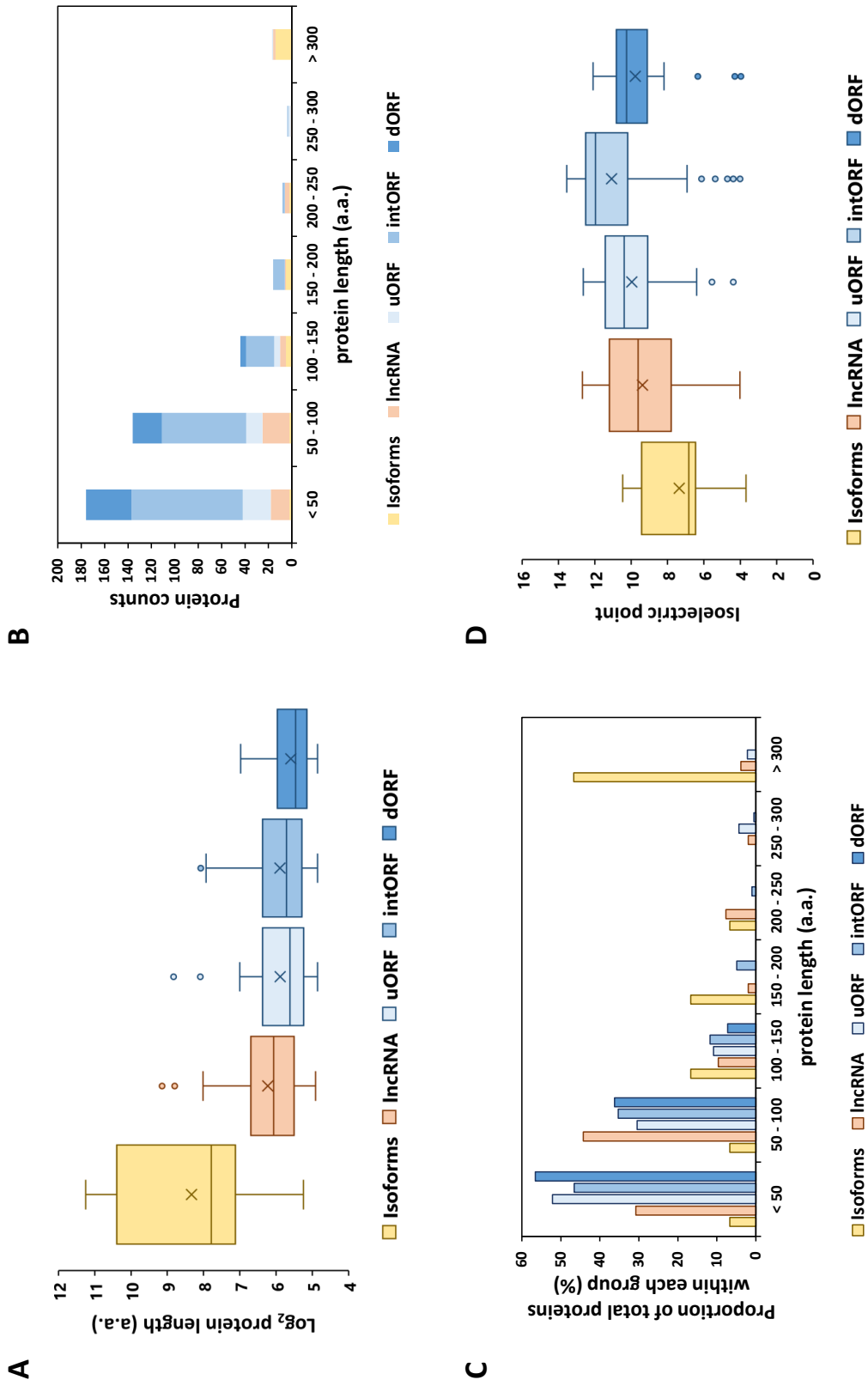




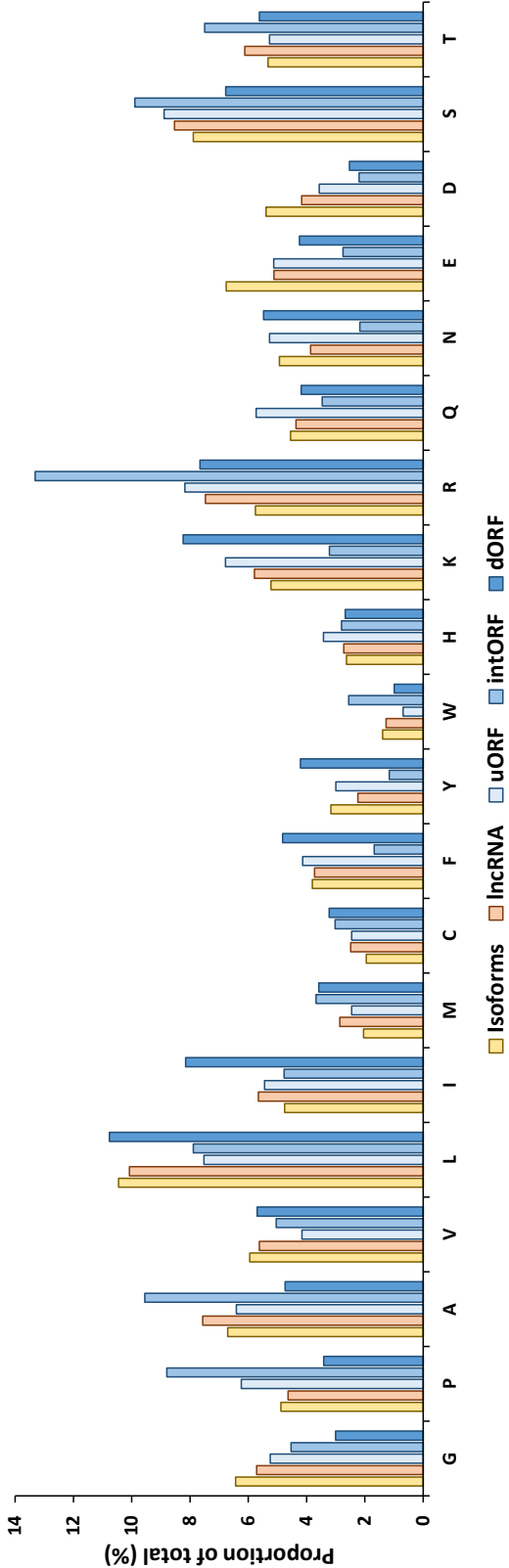


Supp Figure 3

Supp Figure 4



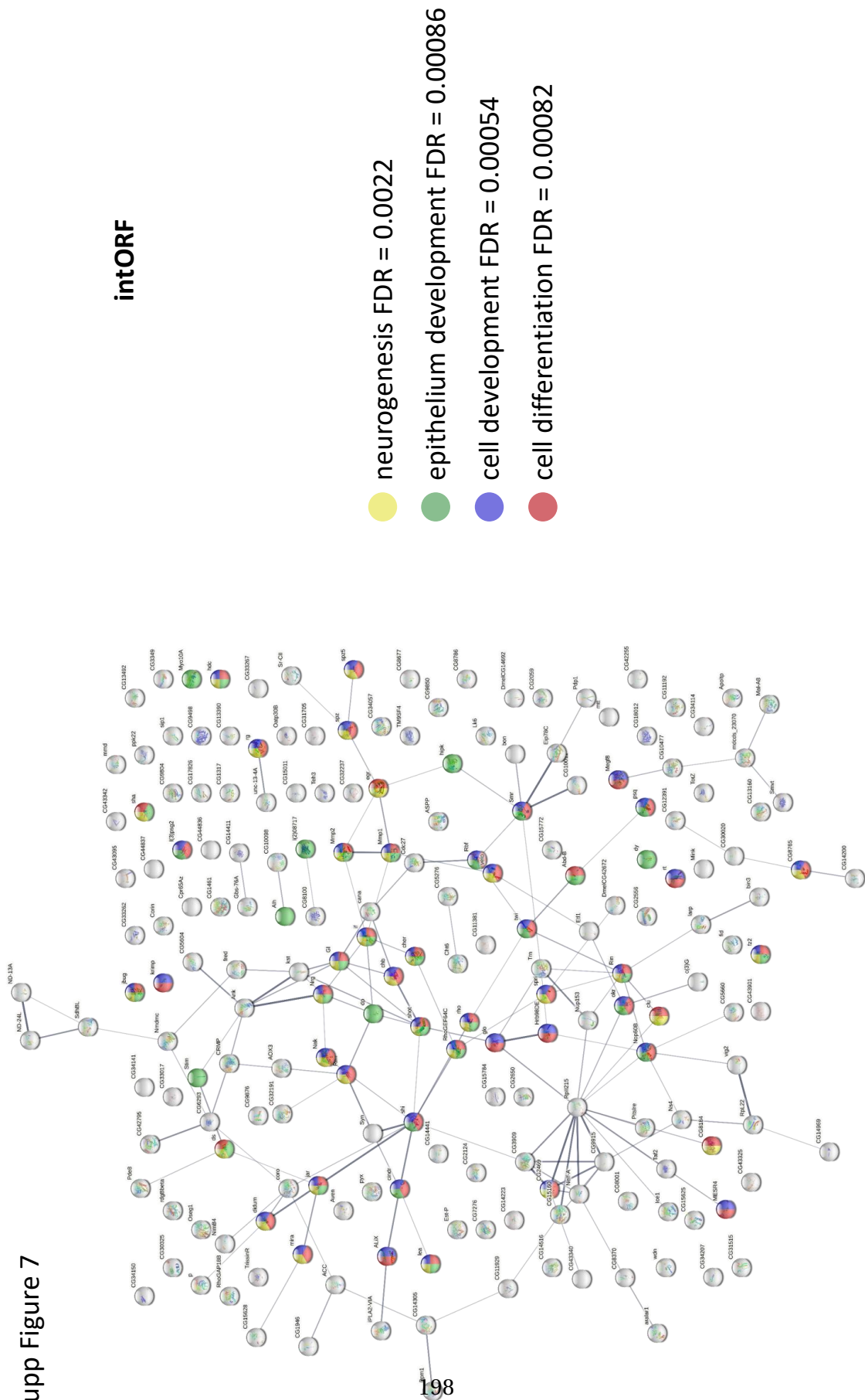




Supp Figure 5

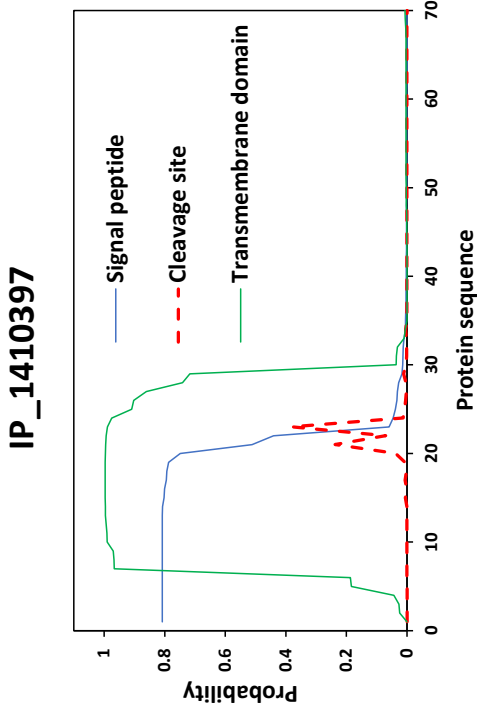


Supp Figure 7



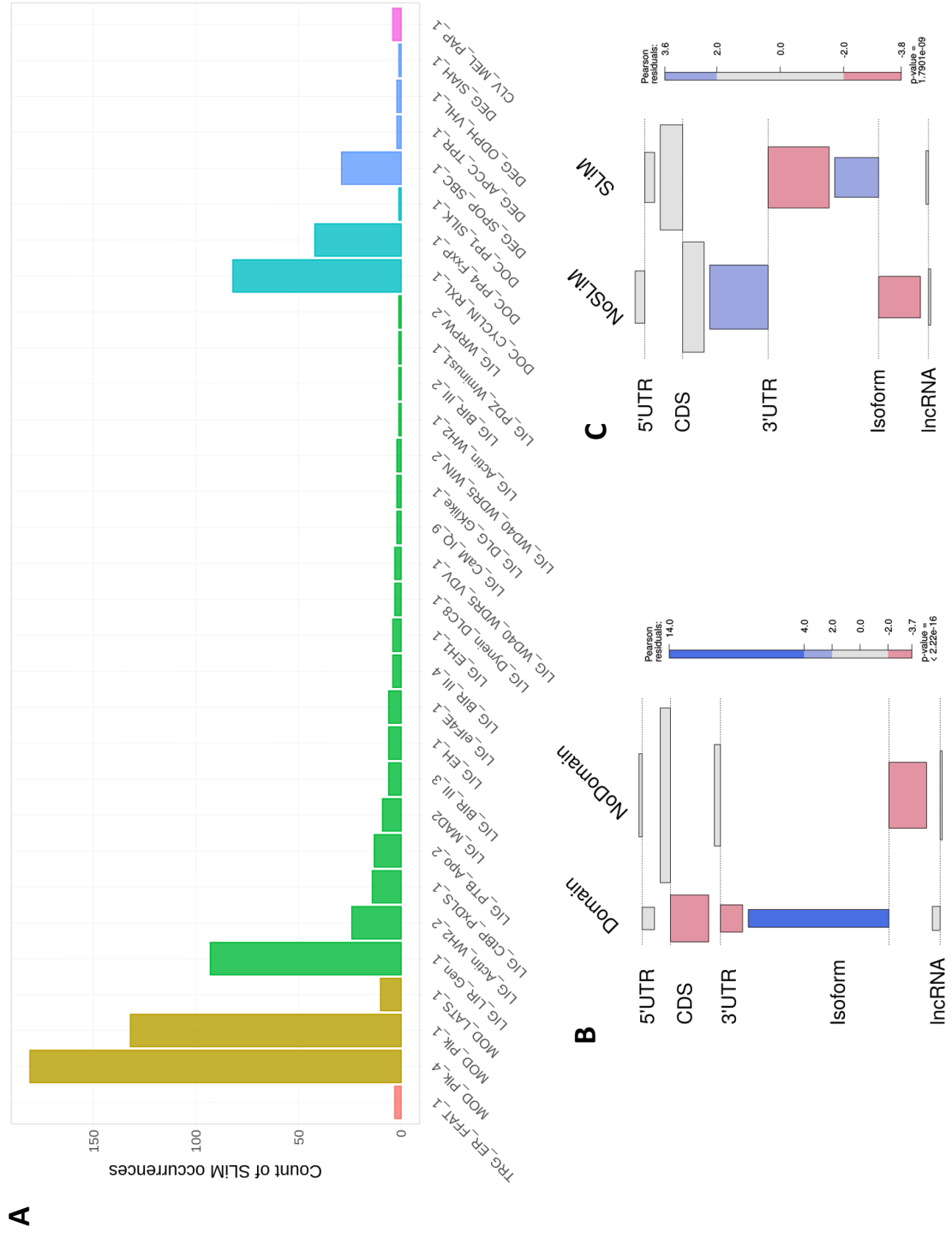


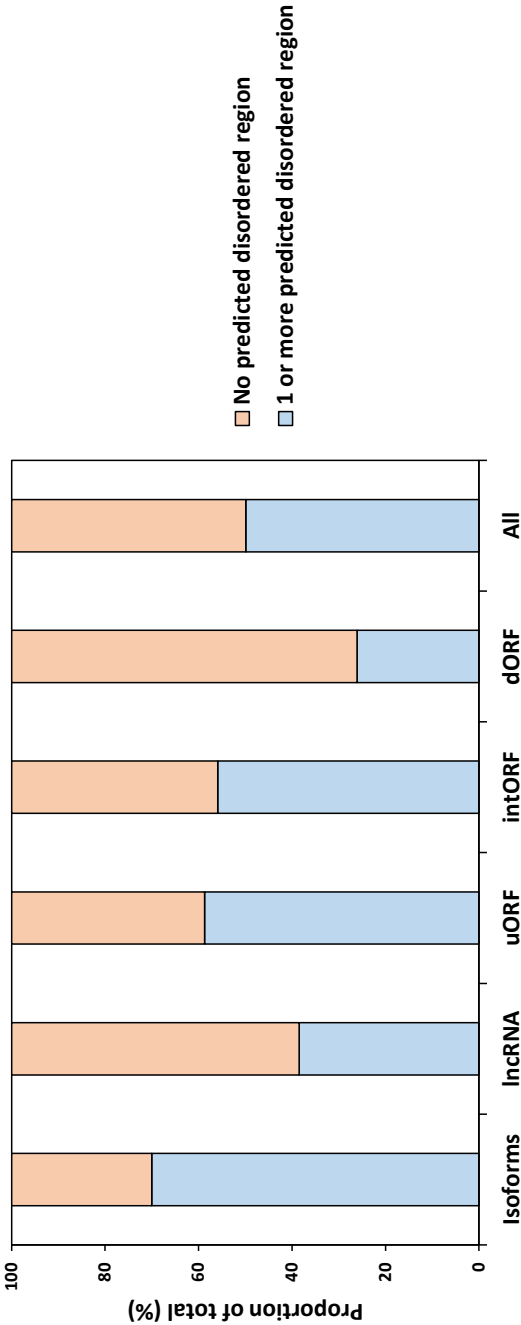




Supp Figure 10

Supp Figure 11

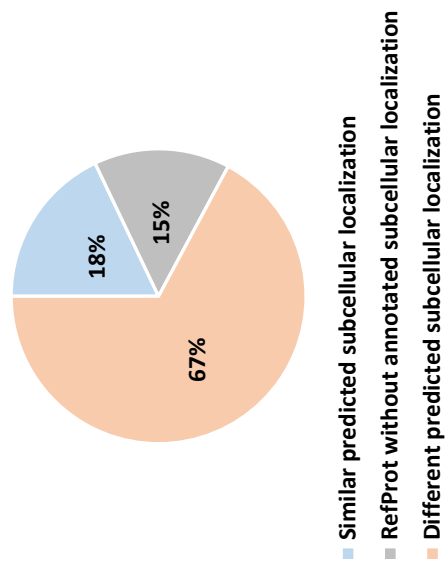


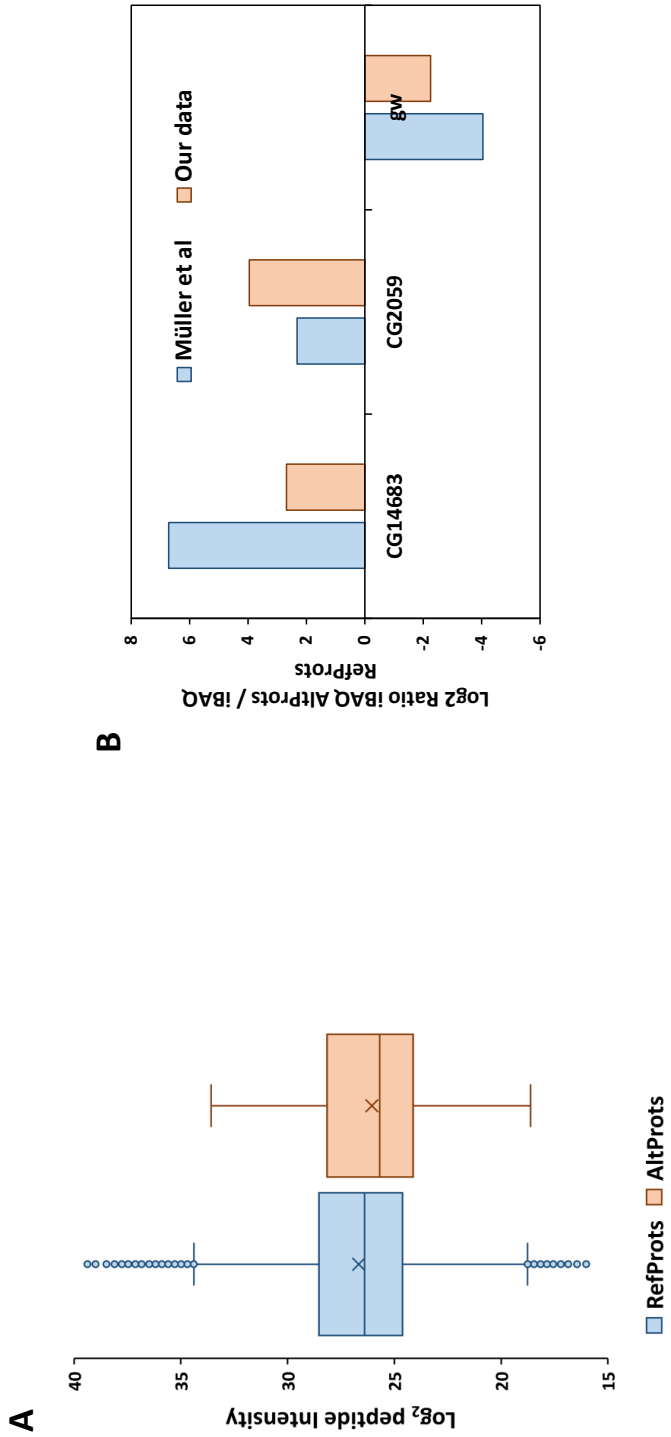


Supp Figure 12

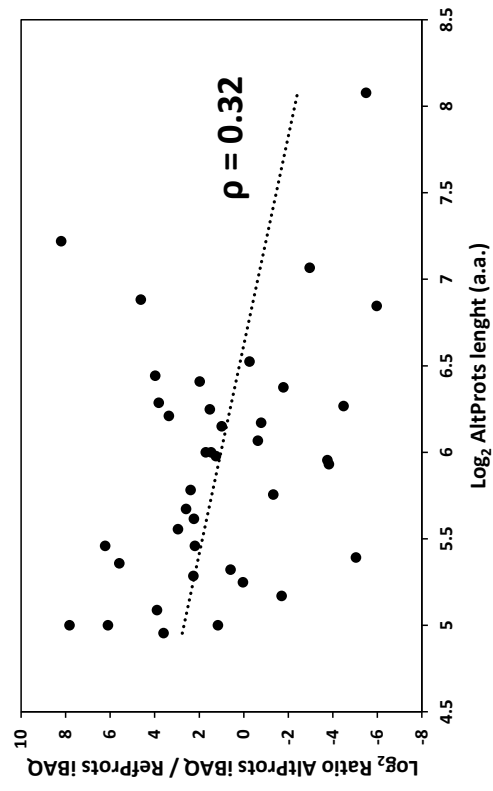


Supp Figure 13





Supp Figure 14



Supp Figure 15

## A. Article: In depth exploration of the alternative proteome of *Drosophila melanogaster*

### Supplementary Figure legends:

**Supplementary Figure 1: Metrics of the different alternative proteins identified in *Drosophila melanogaster*.** A. Andromeda score distribution for peptide spectrum matches of RefProts identified under default MaxQuant settings and AltProts using optimized parameters. B. Distribution of the amino acid length of the AltProts/Isoforms identified in this study. C. Venn diagram representing the overlap of AltProts/Isoforms identified in this study compared to AltProts/Isoforms with MS evidence in OpenProt and the 129 AltProts (out of the 410 small proteins they identified, 281 being already annotated in UniProtKB) identified by Wang *et al.* 2022.

**Supplementary Figure 2: Proportions of the different types of AltProts identified and predicted in *Drosophila melanogaster*.** Distribution of the newly identified proteins depending on their chromosomal location (A), if they are AltPorts or new Isoforms (C), encoded by mRNA or lncRNA (E) and the location of their corresponding ORFs on mRNA (G). The similar distributions were obtained for predicted AltProts/Isoforms from OpenProt (respectively, B, D, F and H).

**Supplementary Figure 3: Distribution of the alternative proteins and isoforms start codon positions ( $\text{Log}_2$  transformed) depending on the types of ORFs they are produced from.**

**Supplementary Figure 4: Alternative proteins produced from different classes of ORFs have different chemical properties.** A. Distribution of the number of newly identified proteins in each type of ORF class and depending on their length in amino acids (a.a.). B-C. Distribution of the amino acid length of the AltProts/Isoforms identified for each type of ORF (B) and normalized by the total protein counts within each group (C). D. Repartition of the isoelectric point measured for the proteins identified in each type of ORF class.

**Supplementary Figure 5: Amino acids proportions obtained from the sequences of the proteins identified in each type of ORF class.**

**Supplementary Figure 6: Gene Ontology term analysis of the host genes of the alternative proteins and isoforms identified in this study using STRING v11.5.**

**Supplementary Figure 7: Gene Ontology term analysis of the host genes of the alternative proteins identified from intORFs using STRING v11.5.**

**Supplementary Figure 8: Gene Ontology term analysis of the host genes of the alternative proteins identified from dORFs using STRING v11.5.**

*A. Article: In depth exploration of the alternative proteome of Drosophila melanogaster*

**Supplementary Figure 9: Gene Ontology term analysis of the host genes of new proteins isoforms and alternative proteins identified from uORFs using STRING v11.5.**

**Supplementary Figure 10: Probability of the presence of signal peptide and transmembrane domain within the first 70 amino acids of the AltProt IP\_1410397 as predicted by SignalP - 5.0 and TMHMM - 2.0, respectively.**

**Supplementary Figure 11: Predicted domains and SLiMs on AltProts.** A. Counts of the different SLiMs motifs in the AltProts identified in this study. B. Association plots showing the dependency between the presence or lack of protein domains for the different types of ORFs the AltProts are produced from. Blue color represents positive association and pink color represents negative association between the presence or absence of protein domains and the type of AltProt. C. Association plots showing the dependency between the presence or lack of SLiMs for the different types of ORFs the AltProts are produced from. Blue color represents positive association and pink color represents negative association between the presence or absence of SLiMs and the type of AltProt.

**Supplementary Figure 12: Distribution of the AltProts with at least one predicted disordered region predicted by IUPred2A and depending on the types of ORFs.**

**Supplementary Figure 13: Proportion of AltProts with predicted subcellular localization similar or different compared to known subcellular localization of their corresponding RefProts.**

**Supplementary Figure 14: A. Comparison of the distribution of peptide intensities ( $\text{Log}_2$  transformed) measured for RefProts and AltProts in protocol 2 from the material and methods section as well as data from Müller *et al.* B. Graph representing the  $\text{Log}_2$  ratio between the iBAQ value of AltProt and corresponding RefProts measured in our study and Müller *et al* for the *CG14683*, *CG2059* and *gw* genes.**

**Supplementary Figure 15: Graph representing the correlation (measured using the Pearson correlation coefficient) between the  $\text{Log}_2$  ratio of the iBAQ value of AltProt and corresponding RefProts and the AltProts amino acid length ( $\text{Log}_2$  transformed).**

## B. Article: Proteostasis in dendritic cells is controlled by the PERK signaling axis independently of ATF4










In this project, our team notably demonstrated that **dendritic cells (DCs)** are displaying high level of phosphorylation of **eIF2 $\alpha$** , mostly induced by an **endoplasmic reticulum** stress. However, I showed in this study that steady state **DCs** do not display a chronic **integrated stress response**-like response. I took advantage of available transcriptomic data and **ATF4**-dependent, **CHOP**-dependent and chronic **ISR**-related gene signatures to perform **gene set enrichment analysis (GSEA)**. No significant enrichment has been found in **DCs** for any of the gene lists, suggesting that neither **CHOP** nor **ATF4** signature can be preferentially detected in **DCs** compared to other immunological cell types. In addition, it is to note that this study highlights the important role played by **PERK** and **GADD34** in the regulation of translational responses of **DCs** submitted to an **ER** stress.

Mendes A, Gigan JP, Rodriguez Rodrigues C, **Choteau SA**, Sanseau D, Barros D, Almeida C, Camosseto V, Chasson L, Paton AW, Paton JC, Argüello RJ, Lennon-Duménil A, Gatti E, Pierre P (2020). Proteostasis in dendritic cells is controlled by the **PERK** signaling axis independently of **ATF4**. *Life Science Alliance*, 10.26508/lsa.202000865, 4(2):e202000865.

Supplementary data are available online at <https://doi.org/10.26508/lsa.202000865>.



# Proteostasis in dendritic cells is controlled by the PERK signaling axis independently of ATF4

Andreia Mendes<sup>1,2,3,\*</sup> , Julien P Gigan<sup>1,\*</sup> , Christian Rodriguez Rodrigues<sup>1</sup> , Sébastien A Choteau<sup>1,6</sup> , Doriane Sanseau<sup>4</sup>, Daniela Barros<sup>1,2,3</sup> , Catarina Almeida<sup>2,3</sup> , Voahirana Camosseto<sup>1,3,4</sup>, Lionel Chasson<sup>1</sup>, Adrienne W Paton<sup>5</sup>, James C Paton<sup>5</sup>, Rafael J Argüello<sup>1,4</sup> , Ana-Maria Lennon-Dumênil<sup>4</sup>, Evelina Gatti<sup>1,2,3,4</sup> , Philippe Pierre<sup>1,2,3,4</sup> 

**In stressed cells, phosphorylation of eukaryotic initiation factor 2 $\alpha$  (eIF2 $\alpha$ ) controls transcriptome-wide changes in mRNA translation and gene expression known as the integrated stress response. We show here that DCs are characterized by high eIF2 $\alpha$  phosphorylation, mostly caused by the activation of the ER kinase PERK (EIF2AK3). Despite high p-eIF2 $\alpha$  levels, DCs display active protein synthesis and no signs of a chronic integrated stress response. This biochemical specificity prevents translation arrest and expression of the transcription factor ATF4 during ER-stress induction by the subtilase cytotoxin (SubAB). PERK inactivation, increases globally protein synthesis levels and regulates IFN- $\beta$  expression, while impairing LPS-stimulated DC migration. Although the loss of PERK activity does not impact DC development, the cross talk existing between actin cytoskeleton dynamics; PERK and eIF2 $\alpha$  phosphorylation is likely important to adapt DC homeostasis to the variations imposed by the immune contexts.**

DOI [10.26508/lsa.202000865](https://doi.org/10.26508/lsa.202000865) | Received 29 July 2020 | Revised 10 December 2020 | Accepted 10 December 2020 | Published online 21 December 2020

## Introduction

DCs are key regulators of both protective immune responses and tolerance to self-antigens (Dalod et al, 2014). DCs are professional APCs, equipped with pattern recognition receptors (PRRs), capable of recognizing microbe-associated molecular patterns (MAMPs) (Akira et al, 2006) and enhance their immunostimulatory activity (Steinman, 2007). MAMPs detection by DCs triggers the process of maturation/activation, which culminates in the unique capacity of priming naïve T cells in lymphoid organs. LPS detection by TLR4 promotes DCs maturation by triggering a series of signaling cascades resulting in secretion of polarizing and inflammatory cytokines, up-regulation of co-stimulatory molecules, as well as

enhanced antigen processing and presentation (Mellman, 2013). All these functions are accompanied by major remodeling of membrane trafficking and actin organization to favor both antigen capture and migration to the lymph nodes (West et al, 2004; Chabaud et al, 2015; Arguello et al, 2016; Bretou et al, 2017).

Upon activation by MAMPs, like LPS, a large augmentation of protein synthesis, representing a two to fivefold increase above resting state, occurs in DCs. This is required for the up-regulation of co-stimulatory molecules at the cell surface and acquires T-cell immunostimulatory function (Lelouard et al, 2007; Reverendo et al, 2019). The phosphorylation of eukaryotic initiation factor 2 (eIF2) is a central hub for regulating protein synthesis during stress. In homeostatic conditions, eIF2 mediates the assembly of the mRNA translation initiation complex and regulates start codon recognition. During stress, phosphorylation of the  $\alpha$  subunit of eIF2 (eIF2 $\alpha$ ) on serine 51 is mediated by a group of four eIF2 $\alpha$  kinases (EIF2AK1-4), which specifically senses physiological imbalance (Arguello et al, 2016; Costa-Mattioli & Walter, 2020). Phosphorylation of eIF2 $\alpha$  converts eIF2 into an inhibitor of the GDP-GTP guanine exchange factor eIF2B, impairing the GDP-GTP recycling required to form new translation initiation complexes (Yamasaki & Anderson, 2008). Consequently, increased eIF2 $\alpha$  phosphorylation impacts cells in two main ways: (i) By reducing the rate of translation initiation and thus global protein synthesis levels; (ii) By favoring the translation of the activating transcription factor 4 (ATF4) (Han et al, 2013; Fusakio et al, 2016) which in turn activates the transcription of genes involved in the integrated stress response (ISR) (Costa-Mattioli & Walter, 2020).

The ISR protects cells from amino acid deprivation, oxidative, mitochondrial stress or viral infections, and is also incorporated as a branch of the ER unfolded protein response (UPR) upon PERK-activation. The ISR comprises a negative feedback loop that causes eIF2 $\alpha$  dephosphorylation, through the induction of GADD34 (also known as PPP1R15a), a phosphatase 1 (PP1c) co-factor (Novoa et al, 2001; Harding et al, 2009). Dephosphorylation of p-eIF2 $\alpha$  by GADD34/PP1c complexes, and

<sup>1</sup>Aix Marseille Université, Centre National de la Recherche Scientifique (CNRS), Institut National de la Santé et de la Recherche Médicale (INSERM), Centre d'Immunologie de Marseille Luminy (CIML), CENTURI, Marseille, France <sup>2</sup>Department of Medical Sciences, Institute for Research in Biomedicine (iBiMED) and Ilidio Pinho Foundation, University of Aveiro, Aveiro, Portugal <sup>3</sup>International Associated Laboratory (LIA) CNRS "Mistra", Marseille, France <sup>4</sup>INSERM U932, Institut Curie, ANR-10-IDEX-0001-02 PSL\* and ANR-11-LABX-0043, Paris, France <sup>5</sup>Department of Molecular and Biomedical Science, Research Centre for Infectious Diseases, University of Adelaide, Adelaide, Australia <sup>6</sup>Aix-Marseille Université, INSERM, Theories and Approaches of Genomic Complexity (TAGC), CENTURI, Marseille, France

Correspondence: pierre@ciml.univ-mrs.fr; gatti@ciml.univ-mrs.fr  
\*Andreia Mendes and Julien P Gigan contributed equally to this work

## B. Article: Proteostasis in dendritic cells is controlled by the PERK signaling axis independently of ATF4



associated protein synthesis restoration, signal ISR termination, and return to cellular homeostasis (Novoa et al, 2001). If stress persists, long-term ATF4 expression promotes programmed cell death, through the induction of the pro-apoptotic transcription factor CHOP (Marciniak et al, 2004). ATF4 also regulates the expression of Rho GTPases and can control cell motility (Pasini et al, 2016), whereas globular actin is part of the PP1c/GADD34 complex and provides additional targeting specificity for dephosphorylating p-eIF2 $\alpha$  (Chambers et al, 2015; Chen et al, 2015).

The ISR can enter in a cross talk with specialized MAMPs sensing pathways, which turns on or amplifies inflammatory cytokines production in different cell types including DCs (Claudio et al, 2013; Reverendo et al, 2018). TLR activation in macrophages undergoing an ISR suppress CHOP induction and protein synthesis inhibition, preventing apoptosis in activated cells (Woo et al, 2009). Moreover, ATF4 binds interferon regulatory factor-7 (IRF7) and prevents type-I IFN transcription (Liang et al, 2011). Several key innate immunity signaling cascades are also believed to be dependent for their signalosome assembly on the chaperone HSPB8 and the eIF2 $\alpha$  kinase heme-regulated inhibitor (HRI/EIF2AK1) (Pierre, 2019). Microbe-activated HRI was shown to mediate phosphorylation of eIF2 $\alpha$  and increase ATF4-dependent expression of HSPB8, thus amplifying signal transduction and inflammatory cytokines transcription in macrophages (Abdel-Nour et al, 2019).

We show here that DCs from spleen or derived from Fms-related tyrosine kinase 3 ligand (Flt3-L) treated-BM cultures display high levels of phosphorylated eIF2 $\alpha$ . Using Cre/lox recombination to generate mice specifically lacking GADD34 (PPP1R15a) or PERK (EIF2AK3) activity in DCs, we demonstrate that PERK-dependent eIF2 $\alpha$  phosphorylation is acquired during BMDC differentiation in vitro. PERK drives high eIF2 $\alpha$  phosphorylation in steady-state DCs with a low impact on protein synthesis levels. We found that mRNA translation in DCs, differently to what has been shown during chronic ISR (cISR) (Guan et al, 2017), is mediated despite high p-eIF2 $\alpha$  levels by an eIF4F-dependent mechanism. These features endow DC with increased resistance to acute ER stress, preventing ATF4 induction in response to stressors such as the bacterial subtilase cytotoxin (SubAB). We also found that LPS-activated primary DCs rely on PERK and eIF2 $\alpha$  phosphorylation to amplify type-I IFN expression, but, conversely to macrophages, not to promote pro-inflammatory cytokines transcription nor IL-1 $\beta$  secretion (Abdel-Nour et al, 2019; Chiritoiu et al, 2019). GADD34 antagonizes PERK activity to maintain functional protein synthesis levels in non-activated DCs and upon stimulation with LPS, contributing directly to DC function by modulating IFN- $\beta$  expression. PERK activity impacts positively DC migration speed, correlating with the regulation of p-eIF2 $\alpha$  levels by the synergistic action of GADD34 and actin cytoskeleton reorganization. Thus, DCs require PERK and GADD34 activity to coordinate protein synthesis, activation, type-I IFN production and migration capacity in response to MAMPs and adapt their biochemical functions to the variations encountered in their external environment.

## Results

### Steady-state DCs display high levels of eIF2 $\alpha$ phosphorylation

Physiological levels of phosphorylated eIF2 $\alpha$  (p-eIF2 $\alpha$ ) were monitored in mouse spleen sections by immunohistochemistry. All

CD11c<sup>+</sup> DC subsets expressing either CD8 $\alpha$  (cDC1), CD11b (cDC2), or B220 (plasmacytoid DC, pDC), displayed high levels of eIF2 $\alpha$  phosphorylation (Fig 1A and B), strongly contrasting with other splenocytes, such as B cells (Fig 1B lower panel). Splenocytes isolation and flow cytometry based-quantification of eIF2 $\alpha$  phosphorylation confirmed that DC subsets display higher levels of p-eIF2 $\alpha$  than T (CD3<sup>+</sup>/CD4<sup>+</sup> or /CD8<sup>+</sup>) or B cells (Fig 1C). We next evaluated p-eIF2 $\alpha$  in BM-derived DCs differentiated in presence of Flt3-Ligand (Flt3-L BMDC), encompassing the major cDC1, cDC2, and pDC subsets in different proportions (circa 30%, 60%, and 10%, respectively) with phenotypes equivalent to those of spleen DC subsets (Brasel et al, 2000). Cell sorting and analysis of the different populations by immunoblot confirmed that all DC subsets display higher eIF2 $\alpha$  phosphorylation in comparison with isolated primary CD8<sup>+</sup> T cells or MEFs stimulated or not with the ER-stress inducing drug thapsigargin for 2 h (Fig 1D). Quantification of p-eIF2 $\alpha$ /eIF2 $\alpha$  ratios indicated that steady-state DCs display two to four times more p-eIF2 $\alpha$ , than stressed MEFs, with the cDC1 population displaying the highest ratio of phosphorylation (Fig 1D). We next evaluated when eIF2 $\alpha$  phosphorylation was acquired during DC differentiation in vitro. Daily analysis of differentiating Flt3-L BMDCs established that high p-eIF2 $\alpha$  levels appear from 4 d of culture (Fig 1E), confirming that eIF2 $\alpha$  phosphorylation is an integral part of Flt3-L induced DC differentiation.

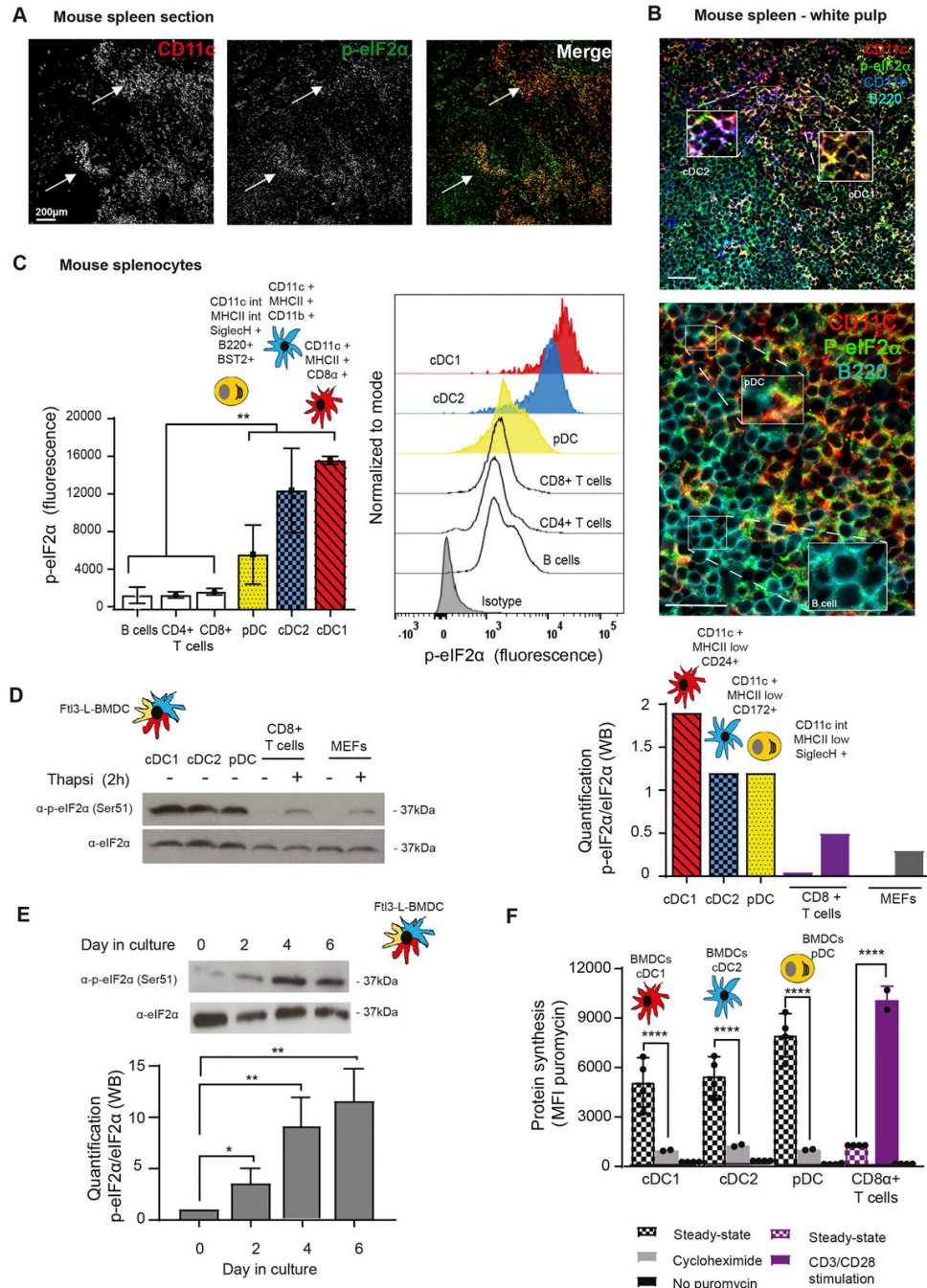
Given the dominant negative effect of p-eIF2 $\alpha$  on translation initiation, we monitored protein synthesis in the different DC subsets. CD8<sup>+</sup> T cells were used as a reference because in these cells, p-eIF2 $\alpha$  is barely detectable. We used puromycilation and detection by flow cytometry (flow) to measure protein synthesis level in splenocytes populations (Schmidt et al, 2009; Arguello et al, 2018). Despite higher eIF2 $\alpha$  phosphorylation levels in all resting DC subsets, mRNA translation is five to eight times higher than in resting CD8<sup>+</sup> T cells and close to the levels reached by these cells upon CD3/CD28 stimulation (Fig 1F). We next monitored protein synthesis every 2 d of culture to establish precisely the influence of eIF2 $\alpha$  phosphorylation during DC differentiation in vitro. We applied flow cytometry and dimensionality reduction using t-distributed stochastic neighbor embedding (tSNE) to visualize DC differentiation and protein synthesis activity within the different subpopulations over time (Fig 2A). We confirmed that protein synthesis levels steadily increased with the appearance of all three DC subsets, this despite high eIF2 $\alpha$  phosphorylation. Noteworthy, the cDC1 population that displays the most elevated level of eIF2 $\alpha$  phosphorylation is the DC subset endowed with the highest level of protein synthesis. These observations suggest that steady-state DCs have adapted their translation machinery to overcome the dominant negative effect on translation initiation of eIF2 $\alpha$  phosphorylation on Ser51, which is associated with the acquisition of the DC phenotype.

### eIF2B and eIF2A expression is up-regulated upon DC differentiation

P-eIF2 $\alpha$  inhibits translation initiation by forming a stable inhibitory complex that reduces the guanidine exchange factor activity of eIF2B. eIF2B is an enzymatic complex with a  $\gamma$ 2 $\epsilon$ 2 sub-units core with levels generally lower than those of its substrate eIF2. Thus, a



*B. Article: Proteostasis in dendritic cells is controlled by the PERK signaling axis independently of ATF4*



**Figure 1. Steady-state FIt3-L BMDCs and splenic DCs display remarkably high levels of eIF2α without inhibition of translation.** (A) Immunohistochemistry of mouse spleen with staining for CD11c (red) and p-eIF2α (green). Scale bar: 200 μm, magnification: 10×. Single color images are shown and merged picture (right row), high level of p-eIF2α staining is mostly found co-localizing in cells positive for CD11c+ (DCs, white arrowheads). (B) Immunohistochemistry of mouse spleen in the white pulp for CD11c (red), p-eIF2α (green), CD11b (blue), and B220 (turquoise). Scale bars: 50 μm, magnification: 40×. In the upper panel, magnified areas show p-eIF2 detection in cDC2 (CD11c+/CD11b+) and cDC1 (CD11c+/CD11b-). In the lower panel, magnified areas show p-eIF2 detection in pDCs (B220+/CD11c+) and in B cells (B220+ and CD11c-). (C) Relative p-eIF2α levels measured by flow in different mouse spleen populations. Statistical analysis was performed by Mann-Whitney

## B. Article: Proteostasis in dendritic cells is controlled by the PERK signaling axis independently of ATF4



partial eIF2 $\alpha$  phosphorylation is sufficient to attenuate protein synthesis initiation in most cells (Adomavicius et al, 2019). We monitored, during Flt3-L BMDC differentiation, the expression of the different eIF2B components. From day 2 in culture, all eIF2B subunits levels were increased transcriptionally, and for eIF2B $\epsilon$ , translationally as well (Fig 2B and C). A similar observation was done with eIF2A, a factor involved, in place of eIF2, in the translation of specialized cellular or viral mRNAs (Kim et al, 2011; Starck et al, 2016). The progression of the ratio eIF2B $\epsilon$ , and potentially of eIF2A, over eIF2 $\alpha$  expression and phosphorylation during differentiation (Fig 2C) could therefore explain the progressive acquisition by BMDCs of significant protein synthesis levels despite abundant eIF2 $\alpha$  phosphorylation.

### Flt3-L BMDCs activation by LPS promotes eIF2 $\alpha$ and eEF2 dephosphorylation

Complementary to p-eIF2 $\alpha$ , phosphorylated translation elongation factor 2 (eEF2) is a major repressor of translation in adverse growth conditions, such as starvation, or accumulation of misfolded proteins in the ER (Ryazanov, 2002; Lazarus et al, 2017). Like for p-eIF2 $\alpha$  (Fig 3A), the high levels of p-eEF2 present in steady-state DCs (Arguello et al, 2018) were gradually decreased during *Escherichia coli* LPS stimulation of TLR4-expressing cDC2 (Fig 3B). eEF2 phosphorylation therefore parallels what is observed for eIF2 $\alpha$  in cDC1 and cDC2 (Fig 3A) and could be involved in the control of protein synthesis levels upon DC activation. Given the rapidity and intensity of eIF2 $\alpha$  and eEF2 dephosphorylation upon activation, we applied to cDC2 the SunRISE technique, a method for monitoring translation elongation intensity using flow (Arguello et al, 2018). cDC2 displayed a striking augmentation of translation intensity upon LPS activation compared with the steady-state situation (T = 0 s), quasi doubling its level in 6 h (Fig 3C and D). Polysomes elongation speed, indicated by the rate of puromycin staining decay after harringtonine treatment (slope), was also increased (x2) by LPS (Fig 3D). eIF2 $\alpha$  and eEF2 dephosphorylation are correlated with increased mRNA translation initiation and elongation allowing protein synthesis to reach its maximum concomitantly to the acquisition by DC of their full immune-stimulatory capacities (Lelouard et al, 2007).

### PPP1R15a (GADD34) controls eIF2 $\alpha$ dephosphorylation in activated DCs

The inducible PP1c co-factor PPP1R15a, known as GADD34, is key in mediating p-eIF2 $\alpha$  dephosphorylation in the resolution phase of the ISR during the UPR (Novoa et al, 2001, 2003). Interestingly, GADD34 induction was reported in inflammatory situations or upon

MAMPs stimulation of different immune cell subsets (Clavarino et al, 2012b, 2016; Ito et al, 2015). In MEF, GADD34 expression is necessary for the production of IFN- $\beta$  upon concomitant sensing of cytosolic dsRNA by RIG I-like-helicases and activation of protein kinase RNA-activated (PKR)-dependent phosphorylation of eIF2 $\alpha$  (Clavarino et al, 2012a).

To further explore the importance of GADD34 in the control eIF2 $\alpha$  pathway in DC, we generated a novel transgenic mouse model with floxed alleles for *Ppp1r15a/Gadd34*. This modification in the *Ppp1r15a* gene allows, upon Cre recombinase expression, the deletion of the third exon that codes for the C-terminal PP1 interacting domain of GADD34. This deletion creates a null phenotype for GADD34-dependent eIF2 $\alpha$  dephosphorylation (Harding et al, 2009) (Fig S1A). *Ppp1r15a*<sup>loxp/loxp</sup> C57/BL6 mouse was crossed with an Itgax-cre deleter strain (Caton et al, 2007) to specifically inactivate GADD34 activity in CD11c-expressing cells, including all DC subsets. Despite inducing a light splenomegaly, GADD34 inactivation had no obvious consequences for splenocyte development in vitro and in vivo (Fig S2). Flt3-L BMDCs derived from WT and Itgax-cre/*Ppp1r15a*<sup>loxp/loxp</sup> (GADD34 $\Delta$ C) mice were LPS-activated prior detection of different translation factors by immunoblot (Fig 4A). GADD34 inactivation prevented LPS-dependent eIF2 $\alpha$  dephosphorylation; however, phosphorylation levels of the activator  $\beta$  subunit of eIF2 (eIF2 $\beta$ ), eEF2, and ribosomal S6 protein remained unchanged, underlining GADD34 specificity for eIF2 $\alpha$  (Fig 4A). eIF2 $\beta$  phosphorylation is known to counteract p-eIF2 $\alpha$  negative effect and promotes mRNA translation (Gandin et al, 2016). However, in our experimental setting, it was neither impacted by LPS activation nor by the loss of GADD34 activity. eIF2 $\beta$  is, therefore, unlikely to interfere with eIF2 $\alpha$  regulation in DCs. Functional deletion of GADD34 inhibited translation initiation in both steady-state and LPS-activated cDC2 (Fig 4B), and also reduced translating polysomes speed in non-stimulated cells. GADD34 expression seems, therefore, to prevent protein synthesis inhibition linked to abundant eIF2 $\alpha$  phosphorylation in steady-state DCs. The amount of eIF2B present in the DC seems, however, sufficient to maintain a lower but still active protein synthesis despite GADD34 inactivation and increased p-eIF2 $\alpha$  (Fig 4A).

In MEF, whereas induction of GADD34 transcription during ER stress is ATF4 dependent (Walter & Ron, 2011), expression of GADD34 upon viral sensors activation is interferon regulatory factor 3 (IRF3) dependent (Dalet et al, 2017). We, therefore, inhibited the TANK-binding kinase 1 (TBK1)/IKK $\epsilon$ /IRF3 signaling axis to investigate if it is also responsible of GADD34 induction in LPS-activated DCs. Treatment with the TBK1 inhibitor (MRT67307, TBKin) (Clark et al, 2011) prevented LPS-dependent induction of GADD34 mRNA (Fig 4C). *Ppp1r15a/GADD34* transcription is, therefore, also partially dependent on the TBK1/IKK $\epsilon$  signaling cascade in DC and not only on

test. \*\**P* < 0.01. (D) Levels of p-eIF2 $\alpha$  and total eIF2 $\alpha$  were measured in DC populations by immunoblot. Sorted steady-state Flt3-L BMDCs were compared with MEFs and freshly isolated CD8 $\alpha$ <sup>+</sup> T cells stimulated or not with thapsigargin (Tg) for 2 h (200 nM). Ratio of p-eIF2 $\alpha$ /eIF2 $\alpha$  is quantified in the graph of the lower panel. (E) Levels of p-eIF2 $\alpha$  and total eIF2 $\alpha$  were measured in bulk Flt3-L BMDCs during different days of BM differentiation in vitro. (F) Levels of protein synthesis were measured by puromycylation and intracellular flow cytometry detection in different subsets of Flt3-L BMDCs and in CD8<sup>+</sup> splenic T cells. Cells were incubated with puromycin 10 min before harvesting and when indicated, cycloheximide (CHX, 10  $\mu$ M) was added 5 min before puromycin. Steady-state Flt3-L BMDCs were directly compared with CD8<sup>+</sup> splenic T cells either steady-state or stimulated overnight with anti-CD3 (10  $\mu$ g/ml) and anti CD28 (5  $\mu$ g/ml). Samples without previous incorporation of puromycin were used as control. All data are representative of n = 3 independent experiments. Data in (F) represent mean fluorescence intensity  $\pm$  SD of three independent experiments. Statistical analysis was performed using unpaired t test (\*\*\*\**P* < 0.0001).

## B. Article: Proteostasis in dendritic cells is controlled by the PERK signaling axis independently of ATF4



the ATF4-dependent transcriptional axis. Importantly, protein synthesis was reduced in GADD34ΔC cells (Fig 4B) and eIF2α phosphorylation increased upon TBK1 inhibition in resting cells (Fig 4D). However, GADD34 mRNA expression levels were too low to define if basal IKKε/TBK1/IRF3 signaling activity could promote GADD34 mRNA expression in steady-state BMDCs. These results suggest nevertheless that GADD34 transcription and translation is regulated in both steady-state and activated DCs by the IKKε/TBK1/IRF3 signaling cascade (Reid et al, 2016).

### PERK mediates eIF2α phosphorylation in steady-state DCs

We next investigated the consequences of inactivating known eIF2α kinases in steady-state Flt3-L BMDCs (Krishna & Kumar, 2018). We tested pharmacological and genetic inactivation of PKR (EIF2AK2) and GCN2 (EIF2AK4) (Fig S3), without observing any major disturbances in eIF2α phosphorylation levels. We next turned toward the ER-stress kinase PERK (EIF2AK3) by crossing PERK<sup>loxp/loxp</sup> mice with the Itgax-cre strain (Caton et al, 2007) allowing for the deletion of the exons 7–9, coding for the kinase domain (PERKΔK) in most CD11c-expressing cells (Fig S1B). PERK protein synthesis levels were enriched in WT CD11c+ splenic DC compared with other splenocytes (Fig 5A). PERK expression was efficiently abrogated in Flt3-L BMDCs and to a relatively lesser extent in spleen DCs isolated from animals bearing the floxed-PERK alleles (Fig 5A). PERK inactivation did not impair DC development *in vitro* nor *in vivo* (Fig S4) but decreased p-eIF2α levels by 60% in steady-state DCs (Fig 5B), whereas p-eEF2 levels remained unchanged (Fig 5C). Interestingly, LPS stimulation induced eIF2α dephosphorylation although PERK levels were increased upon activation of WT Flt3-L BMDCs (Figs 4A and 5A). PERKΔK DCs did not display any additional decrease in p-eIF2α levels, suggesting that GADD34/PP1c activity requires functional PERK activity or high p-eIF2α levels to be implemented in DCs. Conversely to GADD34-deficient cells, PERK deletion increased translation initiation and elongation rate as measured by SunRISE in both steady-state and LPS stimulated cDC2 (Fig 5D). PERK is the EIF2AK responsible for most eIF2α phosphorylation in Flt3-L BMDCs DCs and mirrors GADD34 activity to regulate active protein synthesis at steady-state and during DC activation.

### DCs are insensitive to ISR induction by subtilase cytotoxin (SubAB)

PERK is activated during DC development leading to intense eIF2α phosphorylation at steady state, whereas these cells avoid translational arrest, by expressing GADD34 and eIF2B, among other potential compensatory biochemical mechanisms. We initiated a search to identify the cause of PERK activation in DCs by testing if ER stress triggers IRE1α and PERK activation during DC differentiation. We monitored the splicing of XBP1 mRNA that reflects IRE1α pathway activation (Walter & Ron, 2011) and found limited accumulation of the spliced form of the XBP1 mRNA (sXBP1) in the bulk of differentiating DCs (Fig S5A). mRNA expression of other major transcription factors induced during the UPR (Walter & Ron, 2011; Han et al, 2013), such as ATF4 and ATF6, was found moderately increased at day 7, whereas CHOP transcription remained

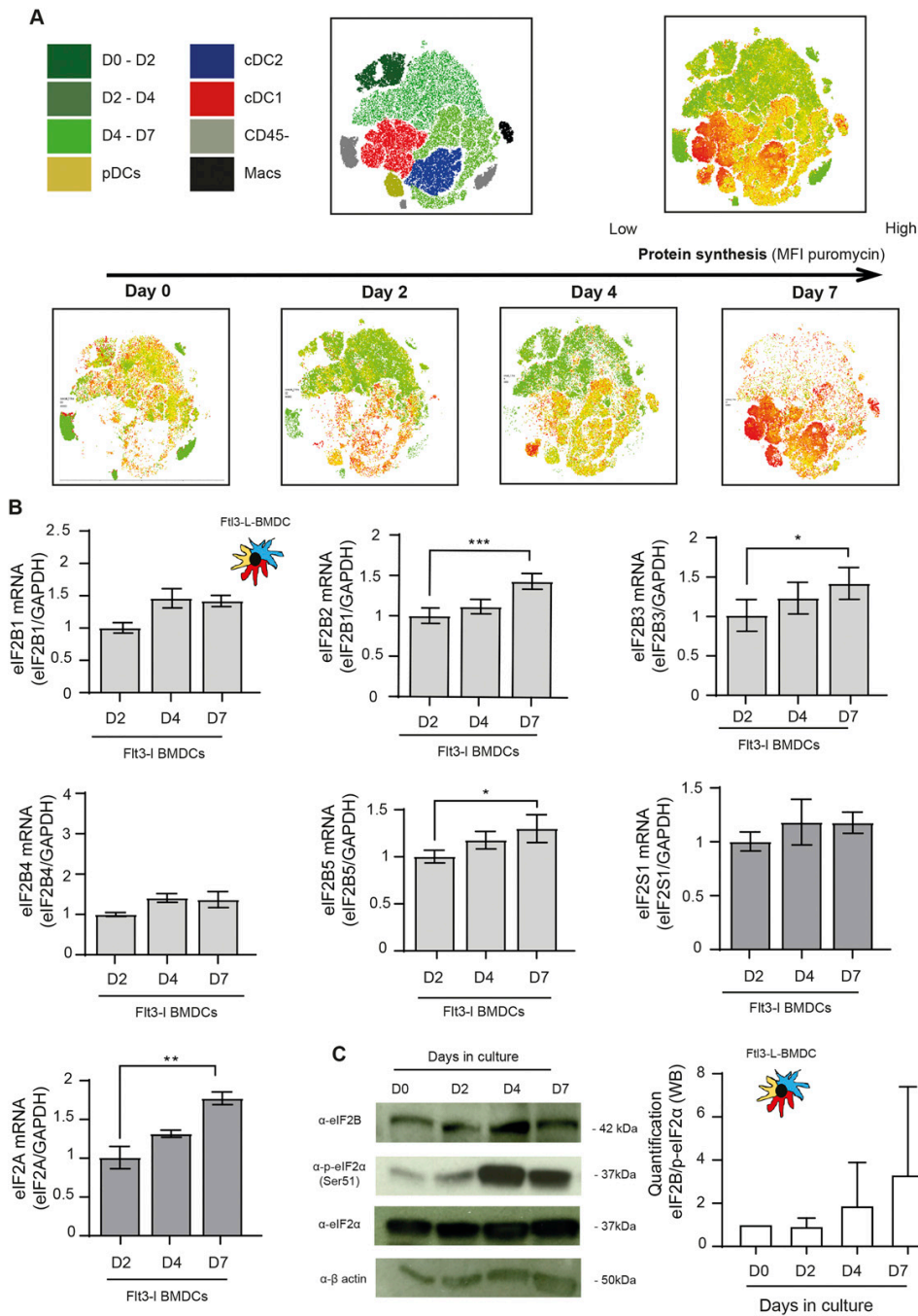
unaffected during differentiation. Because the induction of these factors is mostly regulated at the posttranscriptional level, we wondered if the constant PERK activity observed in DCs could induce a chronic ATF4-dependent ISR in these cells. Recently, translational and transcriptional programs that allow adaptation to chronic ER stress have been described by Guan et al (2017). This cISR operates via PERK-dependent mechanisms, which allow simultaneous activation of stress-sensing and adaptive responses while allowing recovery of protein synthesis.

We took advantage of available transcriptomic data (GSE9810, GSE2389) and of ATF4/CHOP-dependent gene (Han et al, 2013) to perform a Gene Set Enrichment Analysis (GSEA) and define the level of common gene expression found in DCs and potentially shared with an artificially induced acute or cISR. GSEA was followed by multiple testing correction (Subramanian et al, 2005) using the BubbleGUM software, which allows statistical assessment and visualization of changes in the expression of a pre-defined set of genes in different conditions (Spinelli et al, 2015). GSEA revealed no significant enrichment of ATF4- and CHOP-dependent genes expression in the DC transcriptome (false discovery rate [FDR] > 0.25) (Fig S6A and B). Acute ISR- and cISR-dependent transcriptions, respectively, obtained after 1- or 16-h treatments with thapsigargin (Guan et al, 2017) were also compared with splenocyte transcriptomes. Again, no significant gene enrichment could be detected during these analyses (Fig S6C and D) (FDR > 0.25).

PERK- and p-eIF2α-mediated translational reprogramming during cISR appears to bypass cap-mediated translation (Guan et al, 2017). We, therefore, tested if protein synthesis in Flt3-L BMDC was independent of 5' mRNA cap binding eIF4F complex, composed of eIF4A, eIF4E, and eIF4G. We used 4EGI-1, an inhibitor of eIF4F assembly (Moerke et al, 2007), and ROCA, an eIF4A inhibitor (Iwasaki et al, 2019), to treat WT and PERK-deficient DCs and confirm the dependency of their protein synthesis on eIF4F activity. Both compounds had a profound inhibitory effect on DCs translational activity (80% of reduction), irrespective of their subsets or activation state (Fig S5B and C). This level of inhibition indicates that DC mostly depend on eIF4F-dependent cap-mediated translation, thus again contrasting from cells undergoing cISR (Guan et al, 2017). DCs have therefore adapted to the consequences of high eIF2α phosphorylation to allow for translation of their specialized transcriptome, without induction of acute or cISR and ATF4-dependent transcriptional programs.

The lack of ATF4-dependent gene signatures in DCs when compared with other CD45<sup>+</sup> cell types made us to wonder whether the high p-eIF2α levels observed at steady state could interfere with DC capacity to respond to ER-stress. ER-chaperone BiP (HSPA5, heat shock protein family A (Hsp70) member 5) is a key component of the UPR. Accumulation of misfolded protein in the ER-lumen causes BiP dissociation from IRE1α and PERK to induce their dimerization and initiate the different signaling cascades controlling the UPR. Flt3-L BMDCs were exposed to subtilase cytotoxin (SubAB), a bacterial AB5 toxin, which by proteolytic cleavage of BiP induces a strong UPR, including PERK-dependent eIF2α phosphorylation (Paton et al, 2006). When WT and PERKΔK DCs were submitted to SubAB treatment, a modest PERK-dependent phosphorylation of eIF2α was observed in WT cells

*B. Article: Proteostasis in dendritic cells is controlled by the PERK signaling axis independently of ATF4*

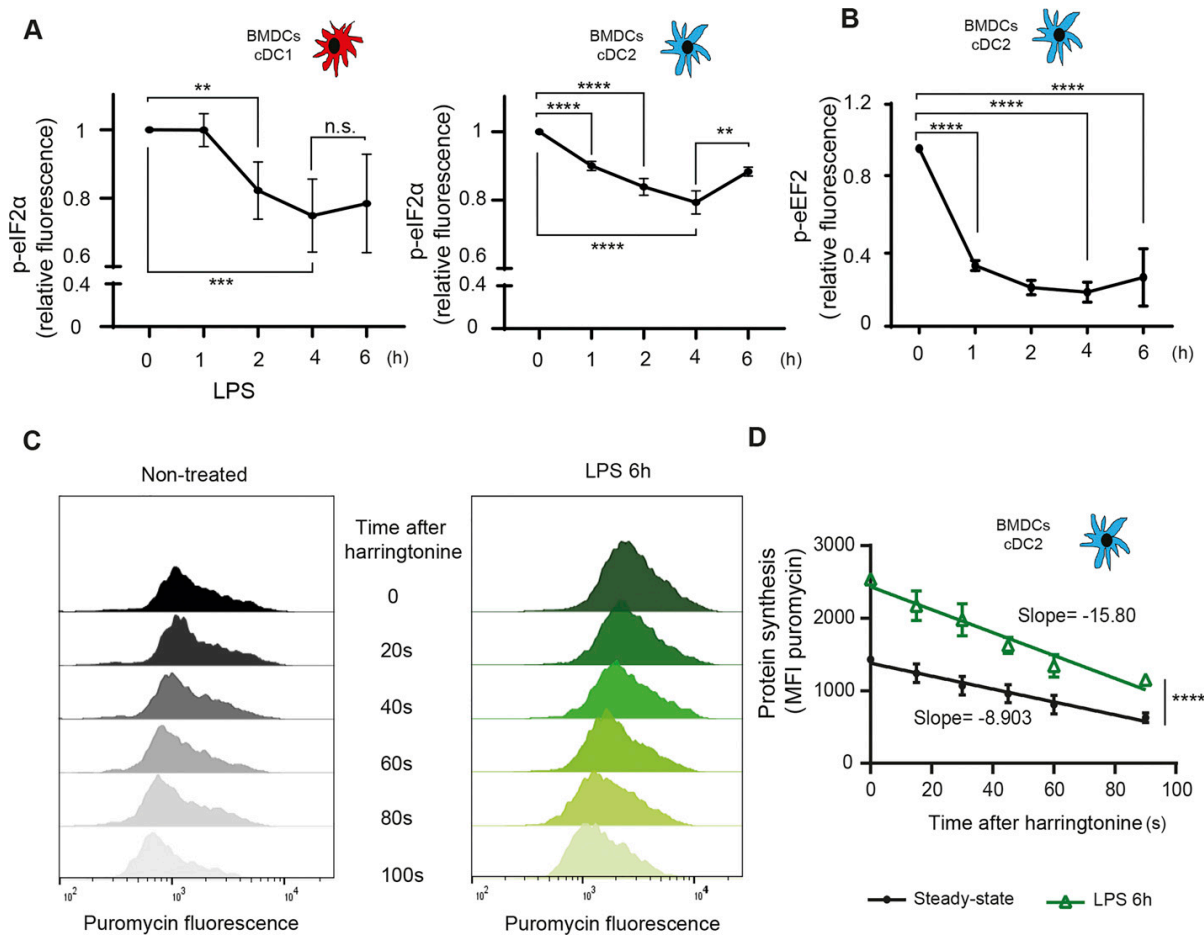


**Figure 2. Protein synthesis is increased during in BMDCs differentiation.**

**(A)** Levels of protein synthesis were measured every 2 d by flow cytometry during Fit3-L BMDCs differentiation in vitro (0, 2, 4, and 7 d). Cells were incubated during 10 min with puromycin, intracellularly stained with an  $\alpha$ -puromycin antibody prior analysis. The same dimensionality reduction using t-distributed stochastic neighbor embedding was applied to all samples. Macrophages (Macs) in black are gated as CD45<sup>+</sup>, CD11c<sup>+</sup>, CD11b<sup>+</sup>, F4/80<sup>+</sup>, and CD64<sup>+</sup> cells; cDC1 in red express CD45<sup>+</sup>, CD11c<sup>+</sup>, MCHII<sup>+</sup>, and CD24<sup>+</sup>; cDC2 in purple express CD45<sup>+</sup>, CD11c<sup>+</sup>, MHC II<sup>+</sup>, CD11b<sup>+</sup>, and Sirp $\alpha$ <sup>+</sup>; pDC in yellow express CD11c<sup>int</sup> and Siglec H<sup>+</sup>; cells negative for CD45 in gray are considered as non-immune. **(B)** mRNA levels of eIF2B $\epsilon$  (B5), eIF2B $\gamma$  (B3), eIF2B $\alpha$  (B1), eIF2B $\beta$  (B2), eIF2B $\delta$  (B4), eIF2 $\alpha$  (eIF2S1), and eIF2A measured by qRT-PCR in bulk Fit3-L BMDCs at



*B. Article: Proteostasis in dendritic cells is controlled by the PERK signaling axis independently of ATF4*



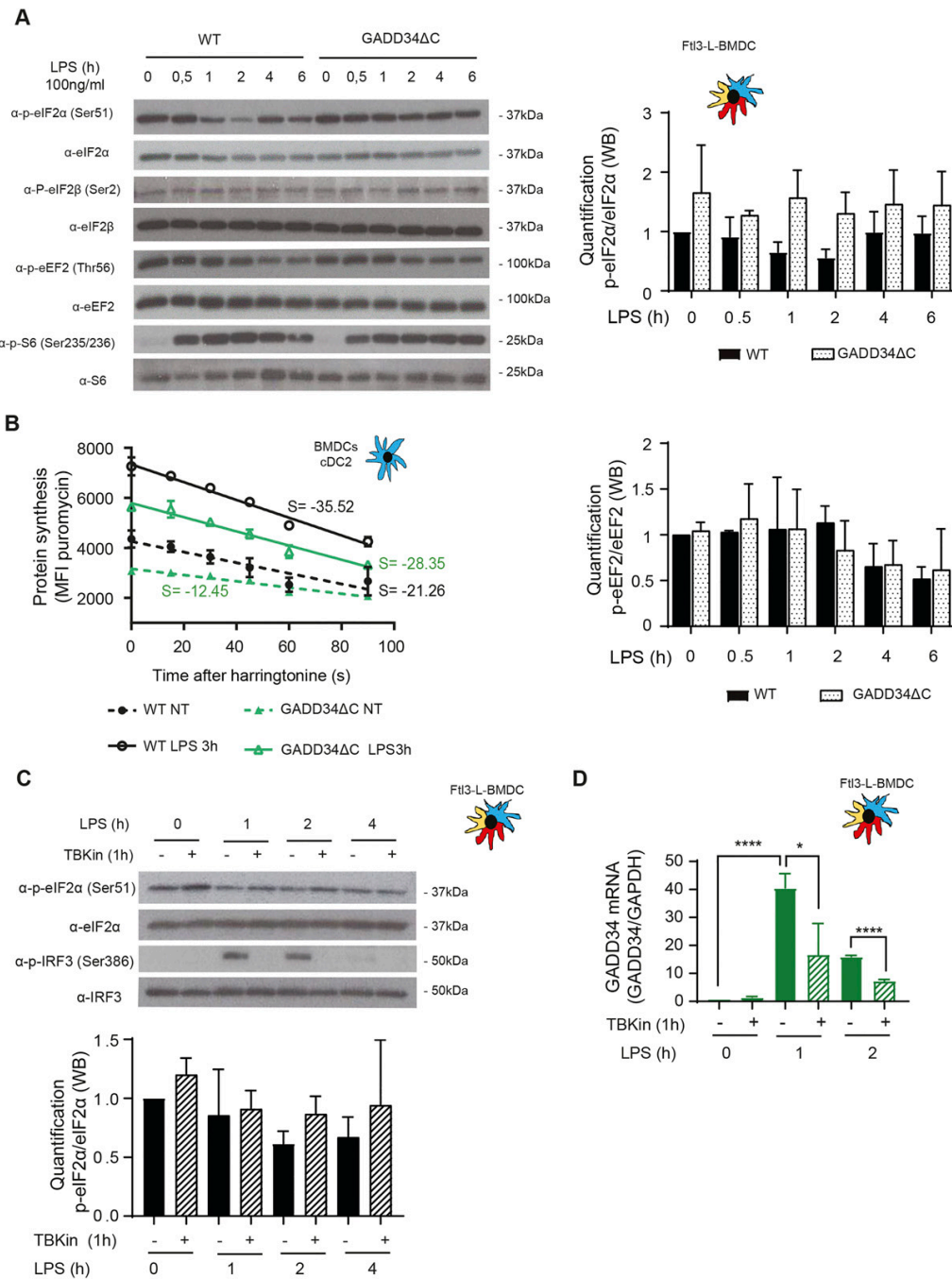
**Figure 3. P-eIF2 $\alpha$  and p-eEF2 levels are down-regulated upon LPS stimulation.** Flt3-L BMDCs were stimulated with LPS (100 ng/ml) for indicated hours. **(A, B)** Monitoring of p-eIF2 $\alpha$  and (B) p-eEF2 by intracellular flow in cDC1 and cDC2. **(C)** Flow detection of puromycin incorporation was performed on the cDC2 population. Total puromycin mean fluorescence intensity between steady-state and LPS activated cDC2 is shown for different time of harringtonine treatment. **(D)** Mean fluorescence intensity was plotted as a decay slope and establish the speed of translation elongation. All data are representative of  $n = 3$  independent experiments. Data represent mean  $\pm$  SD of three independent experiments. **(A, B, C, D)** Statistical analysis was performed using Dunnett's multiple comparison (A, B, C, D) Mann-Whitney test (\*\* $P < 0.01$  and \*\*\*\* $P < 0.0001$ ).

(Fig 6A), with limited consequences on translation (Fig 6B). In contrast to SubAB, thapsigargin treatment arrested translation more efficiently and triggered stronger eIF2 $\alpha$  phosphorylation by PERK, but also by a different EIF2AK because p-eIF2 $\alpha$  levels were also increased in PERK $\Delta$ K cells. Little ATF4 could be detected in the cytosolic or nuclear fractions of control or toxin-treated DC (Fig 6C), reflecting the modest induction of eIF2 $\alpha$  phosphorylation (Fig 6D), and confirming the limited impact of SubAB on ISR induction. The efficacy of SubAB treatment was tested in MEFs, in which ATF4 was strongly induced by the toxin and absent from control ATF4-/-

cell (Fig 6C). DCs are therefore unable to induce the ISR upon SubAB treatment, despite the activation of other UPR branches, as demonstrated by augmented IRE1 $\alpha$  activity (Fig 6E), that is responsible for XBP-1 splicing and translation reduction through IRE1-dependent decay of mRNA (RIDD) (Tavernier et al, 2017). BiP mRNA levels were moderately augmented during DC differentiation. However, the similar expression levels observed for DCs and MEFs (Fig 6F) suggest that BiP transcriptional regulation is not involved in the DC resistance to SubAB. The relatively high PERK and GADD34 activity observed in steady-state DCs, together with

indicated days of differentiation and compared with control MEFs. **(C)** Levels of eIF2B $\epsilon$ , P-eIF2 $\alpha$ , total eIF2 $\alpha$ , and  $\beta$ -actin were measured in bulk Flt3-L BMDCs at indicated days of differentiation in vitro. Quantification of the ratio eIF2B $\epsilon$ /P-eIF2 $\alpha$  is shown on the right. All data are representative of  $n = 3$  independent experiments. Data in (B) represent Mean  $\pm$  SD of three independent experiments. Statistical analysis was performed using Dunnett's multiple comparison (\* $P < 0.05$ , \*\* $P < 0.01$ , and \*\*\*\* $P < 0.001$ ).

*B. Article: Proteostasis in dendritic cells is controlled by the PERK signaling axis independently of ATF4*



**Figure 4. GADD34 mediates eIF2α dephosphorylation and controls translation in DCs.**

WT and GADD34ΔC BMDCs were stimulated with LPS (100 ng/ml) for the indicated times. **(A)** Levels of p-eIF2α, total eIF2α, p-eIF2β, total eIF2β, p-eEF2, total eEF2, P-S6, and total S6 were detected by immunoblot (top left) and quantification is shown in the different panels. **(B)** The speed of translation elongation was measured by SunRISE after 3 h of incubation with LPS. Harringtonine (2 μg/ml) was added at different times up to 90 s prior incubation with puromycin during 10 min. Flow intracellular staining was performed in cDC2 using an α-puromycin antibody. The total decay of puromycin mean fluorescence intensity between WT and GADD34-deficient DCs in steady-state and upon activation indicates that translation initiation and elongation speed is decreased in GADD34-deficient DCs. Fli3-L BMDCs were pretreated with 2 μM of the TBK1 inhibitor

## B. Article: Proteostasis in dendritic cells is controlled by the PERK signaling axis independently of ATF4



high eIF2B expression, and potentially eIF2A, would prevent a full ISR induction, including translation arrest, ATF4 synthesis and associated transcriptional response.

### Importance of the ISR for PRRs signaling

In macrophages, p-eIF2 $\alpha$ - and ATF4-dependent expression of HSPB8 is required for the assembly of PRR signaling adapters, such as mitochondrial antiviral-signaling protein (MAVS), or TIR domain-containing adapter protein inducing interferon- $\alpha$  (TRIF), but not of myeloid differentiation primary response gene 88 (MyD88). HSPB8 seems necessary for TRIF and MAVS to be incorporated into protein aggregates that constitute signalosomes for different innate immunity signaling pathways triggered by MAMPs (Abdel-Nour et al, 2019). Given the lack of ATF4 synthesis and the resistance of active DCs to mount an acute ISR, we investigated their capacity to produce pro-inflammatory cytokines and type-I IFN in a perturbed eIF2 $\alpha$ -phosphorylation context. Importantly, we have previously shown that GADD34-deficient BM-derived and spleen DCs, have a reduced capacity to produce IFN- $\beta$  (Clavarino et al, 2012b; Perego et al, 2018). IFN- $\beta$  and IL-6 mRNA expression were therefore quantified after 4 h of LPS activation of WT and PERK $\Delta$ K DCs. IL-6 transcription (Fig S7A) and secretion (Fig 7A) remained unchanged. However, IFN- $\beta$  secretion was reduced by half in PERK $\Delta$ K DCs (Fig 7A), whereas its transcription remained surprisingly unaffected in mutant cells (Fig S7A), suggesting that PERK activity is necessary for normal synthesis and secretion of IFN- $\beta$  independently of the activation of the TRIF-dependent pathway downstream of TLR4.

We further investigated the importance of eIF2 $\alpha$ -phosphorylation with respect to PERK activity using a pharmacological ISR inhibitor (ISRIB) (Sidrauski et al, 2015), which prevents inhibition of eIF2B by p-eIF2 $\alpha$  and prevents translation inhibition, as shown here for MEFs in different ISR-inducing conditions (Fig 7B). ISRIB should prevent the induction of the ISR in activated DCs and interfere, as reported for macrophages (Abdel-Nour et al, 2019), with TRIF signaling and down-stream IFN- $\beta$  expression. In our experimental system, we used LPS and polyinosinic:polycytidylic acid (poly(I:C)) to stimulate, respectively, TRIF-dependent TLR4 and TLR3 (Fitzgerald & Kagan, 2020). IFN- $\beta$  mRNA induction upon stimulation of DCs with either LPS (Figs 7C and S7B) or poly(I:C) (Fig S7B) was not impaired by ISRIB. IL-6 transcription which is believed to be mostly Myd88-dependent was moderately decreased by ISRIB in LPS-activated DCs and more acutely in BMDM (Fig S7B and C), confirming that cell-specific mechanisms control the transcription of the IL-6 family of cytokines during the ISR (Sanchez et al, 2019). Importantly, TRIF-dependent expression of IFN- $\beta$  upon LPS activation of BMDM was not impacted by ISRIB treatment, whereas comparatively, poly I:C activation of these cells was too inefficient to obtain statistically reliable data (Fig S7C). We next tested if ISRIB treatment augments IFN- $\beta$ , IL-6, IL-10, and TNF secretion after 4 h of LPS stimulation in BMDM (Fig 7D). This unchanged (DC) or augmented (BMDM) IFN- $\beta$

production observed in the presence of ISRIB confirms that TRIF-dependent signaling does not require acute ISR induction nor ATF4-dependent transcription to promote signalosomes assembly and cytokines expression in DC and probably also macrophages. ISRIB facilitates, however, the production of several cytokines upon activation, confirming that eIF2 $\alpha$  phosphorylation decreases the efficacy of cytokines mRNAs translation upon MAMPs detection.

Given the impact of ISRIB on cytokine secretion, we decided to analyze further the response to LPS in cells inactivated for PERK pharmacologically. Cytokines expression was monitored in LPS-activated DCs in presence of the PERK inhibitor GSK2656157 (Axten et al, 2013) (Fig 7C). GSK2656157 treatment decreased both IFN- $\beta$  and IL-6 secretion by 30–50% (Fig 7A). Over 4 h of treatment, no significant changes in IL-6 mRNA expression was observed, whereas IFN- $\beta$  transcription was reduced (Fig S7D). These cytokines seem therefore differently affected by alterations in DCs of PERK activity and of eIF2 $\alpha$  phosphorylation. IL-6 transcription is sensitive to ISRIB and requires eIF2 $\alpha$  phosphorylation. IFN- $\beta$  transcription and secretion seems, however, dependent on PERK activity, but surprisingly not on eIF2 $\alpha$  phosphorylation nor the ISR. We extended our analysis to IL-10 and TNF secretion upon GSK2656157 treatment of LPS-stimulated Flt3L-BMDCs (Fig 8A). These cytokines expression remained, however, unaffected by ISRIB, but like for IFN- $\beta$  and IL-6, their translation was reduced upon PERK inhibition, suggesting a key role for PERK in promoting cytokines translation in activated DCs.

PERK was recently proposed to control the caspase-1-dependent proteolysis of pro- to mature IL-1 $\beta$  to allow its secretion (Chiritoiu et al, 2019). Given the contrasting effects of PERK inactivation on IFN- $\beta$  and IL-6 expression, we examined how WT and PERK $\Delta$ K-DCs co-stimulated with LPS and ATP were promoting the conversion and secretion of mature IL-1 $\beta$  (Fig 8B). Surprisingly, we did not observe any impairment of IL-1 $\beta$  expression in PERK $\Delta$ K-DCs, but rather an increase by 25% of both IL-1 $\beta$  mRNA transcription and mature IL-1 $\beta$  secretion compared with WT cells (Fig 8B). IL-1 $\beta$  secretion was also monitored in presence of ISRIB and GSK2656157 (Fig 8C), which moderately reduced IL-1 $\beta$  mRNA transcription only in LPS and ATP activating conditions, this without incidence on mature IL-1 $\beta$  secretion. These results suggest that acute pharmacological PERK inactivation has little effect on IL-1 $\beta$  processing and secretion, whereas long-term inactivation favors this secretion potentially by decreasing the inflammasome activation threshold, as previously observed in autophagy deficient macrophages or DCs (Terawaki et al, 2015). PERK inactivation in DCs is therefore not detrimental to IL-1 $\beta$  processing but favors its production and secretion, which could in turn increase *IL1B* mRNA transcription in a feed-back positive loop (Ceppi et al, 2009).

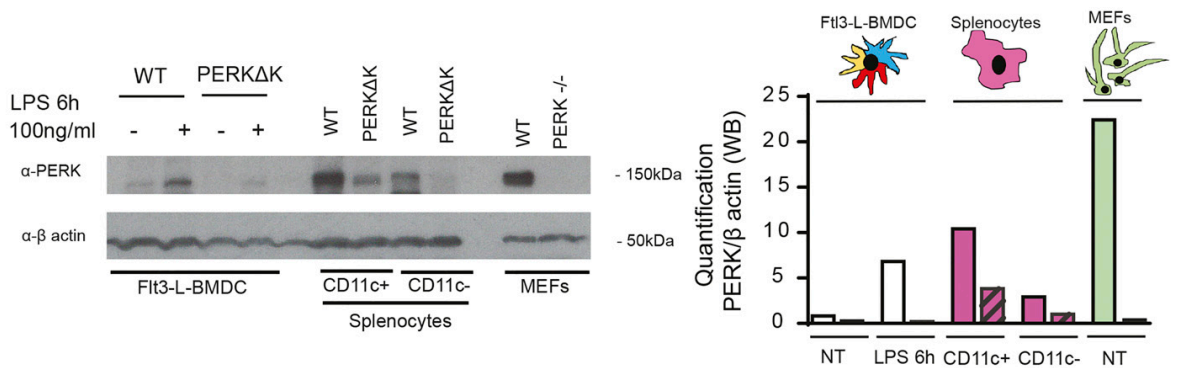
### Antigen presentation in PERK-deficient DCs

Given the impact of PERK inactivation in DC capacity to secrete type-I IFN, we decided to investigate how it could also interfere with the

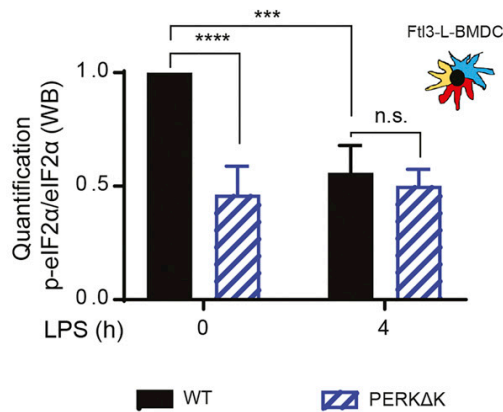
(MRT67307) for 1 h, before stimulation with LPS (100 ng/ml) for indicated times. (C) mRNA levels of GADD34 were measured by qRT-PCR and normalized to the housekeeping gene (GAPDH) level. (D) Immunoblot detection of p-eIF2 $\alpha$ , total eIF2 $\alpha$ , P-IRF3, and total IRF3. Quantification is represented on the right. (A, B, C) Statistical analysis was performed using the Wilcoxon test (A, C, B), and Mann-Whitney test (\* $P$  < 0.05 and \*\*\*\* $P$  < 0.0001). (D) All data are representative of  $n$  = 3 independent experiments except (D),  $n$  = 2.

B. Article: Proteostasis in dendritic cells is controlled by the PERK signaling axis independently of ATF4

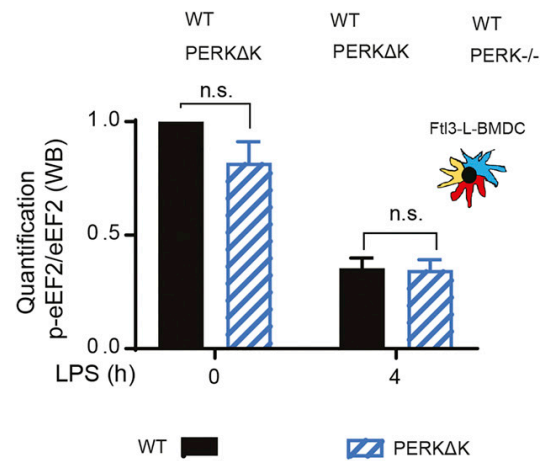
A



B



C



D

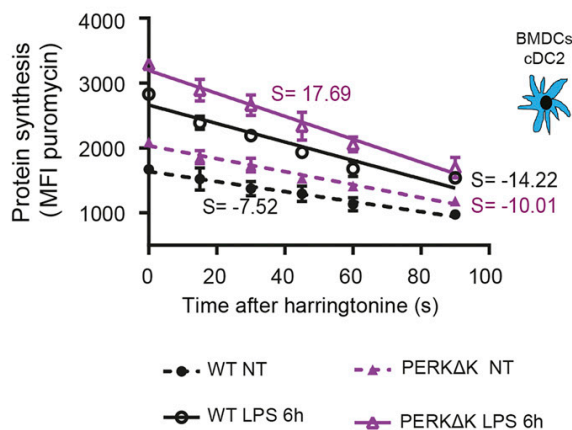


Figure 5. PERK is activated in steady-state DCs.

(A) Immunoblot detection of PERK and β-actin. Quantification is shown on the right. WT and PERKΔK FIt3-L BMDC treated or not with LPS (100 ng/ml) for 6 h were compared with CD11c+ and CD11c- fractions of splenocytes. WT and PERK-/- MEFs were used as control. (B, C) WT and FIt3-L PERKΔK BMDCs were stimulated or not with LPS (100 ng/ml) for 4 h. (B) Quantification of p-eIF2α/eIF2α ratio (immunoblot). (C) Quantification of p-eEF2/eEF2 ratio (immunoblot). (D) The speed of translation elongation was measured using SunRISE in FIt3-L cDC2 after 4 h of incubation with LPS. The total decay of puromycin mean fluorescence intensity between WT and PERKΔK in steady-state and upon activation is shown. All data are representative of n = 3 independent experiments. (B, C, D) Statistical analysis was performed using Wilcoxon test (B, C) and Mann-Whitney test (D). Data in (D) represent mean fluorescence intensity ± SD of three independent experiments (\*\*P < 0.01, \*\*\*P < 0.001, and \*\*\*\*P < 0.0001).



## B. Article: Proteostasis in dendritic cells is controlled by the PERK signaling axis independently of ATF4



processing and presentation of exogenous antigens. Surface levels of MHC II and CD86 of WT and PERKΔK BMDC were monitored by flow to establish the capacity of DC1 and DC2 subsets to activate in response to LPS (Fig S8A). Our analysis indicated that WT and PERKΔK BMDC responded equally well to LPS stimulation and did not display differences in surface MHC II nor CD86 levels that could suggest an impairment in their presentation capacity to T cells. We next incubated Flt3-L BMDCs for 8 h with increasing concentrations of hen egg lysozyme (HEL) in presence or not of the PERK inhibitor GSK2656157, prior assaying processing and presentation by measuring CD69 surface up-regulation and the production of IL-2 by the 3A9 (HEL 48-62 on I-Ak) specific T hybridoma (Fig S8B). We found that PERK inhibition had no effect on the efficiency of MHC II restricted processing and presentation of soluble antigens in vitro and consequently did not interfere with the transport of MHC II molecules.

### PERK and actin polymerization coordinates p-eIF2α levels and migration in DCs

Recently, the importance of globular actin in the formation of a tripartite holophosphatase complex assembled with GADD34 and PP1c to dephosphorylate eIF2α was revealed (Chambers et al, 2015; Crespillo-Casado et al, 2017, 2018). Given the unusual regulation of eIF2α phosphorylation in DCs, actin organization could impact this pathway in a different setting than artificial ER stress induction. Actin depolymerizing and polymerizing drugs, respectively, Latrunculin A (Lat A) and Jasplakinolide (Jaspk) had opposite effects on eIF2α phosphorylation in Flt3-L BMDCs (Fig 9A). Globular actin accumulation induced by Lat A was strongly correlated with eIF2α dephosphorylation (Fig 9A and B), whereas actin polymerization induced by Jaspk resulted in a massive increase in eIF2α phosphorylation, together with a reduction in protein synthesis (Fig 9A and C). We next tested the impact of the two drugs on WT and PERKΔK Flt3-L BMDCs activated or not by LPS (Fig 9D). LPS activation or Lat A treatment resulted in the same levels of eIF2α dephosphorylation (Fig 9D). In contrast, Jaspk dominated LPS effect and strongly increased p-eIF2α levels in all conditions tested. PERK inactivation decreased the levels of p-eIF2α, but had no obvious consequences on the efficacy of the drugs because both induced similar responses in WT and PERKΔK cells.

We tested the impact of actin remodeling and translation regulation on the acquisition by DC of their immune-stimulatory phenotypes. Surface MHC II and co-stimulatory molecule CD86 expression were up-regulated by LPS stimulation but remained unaffected by Jaspk treatment (Fig S9A). Similarly, transcription levels of key cytokines such as IL-6 and IFN-β were insensitive to this actin polymerizing drug (Fig S9B and C). However, whereas IL-6 secretion remained identical, IFN-β levels were found reduced by Jaspk treatment, as expected from a situation in which GADD34 activity is reduced (Clavarino et al, 2012b). Interestingly, the IL-6 gene was described to bear an upstream uORF-dependent translational regulation, which could allow IL-6 mRNA translation upon high eIF2α-phosphorylation conditions induced by Jaspk treatment (Sanchez et al, 2019). Taken together, these observations suggest that extensive actin polymerization in DCs increases strongly eIF2α phosphorylation, affecting protein synthesis and

ultimately controlling translationally specific cytokines expression, similarly to what has been observed with Cdc42 or Wiskott-Aldrich syndrome protein mutants (Pulecio et al, 2010; Prete et al, 2013). These results further suggest that actin dynamics and its effect on eIF2α phosphorylation could be key regulating factors for the homeostasis and translation of specific mRNA encoding for proteins generally secreted in a polarized fashion, such as type I IFNs, or associated with cell migration (Pulecio et al, 2010; Prete et al, 2013). Finally, given the interplay between eIF2α phosphorylation and actin polymerization, we wondered whether PERK-deficient cells could display some migratory deficits. We used micro-fabricated channels, which mimic the confined geometry of the interstitial space in tissues (Heuze et al, 2013; Bretou et al, 2017), to find that PERK-deficient cells were not able to increase their migration speed in response to LPS (Fig 9E), confirming the link between eIF2α phosphorylation and actin dynamics. PERK activity is, therefore, necessary for DCs to acquire normal immune-stimulatory and migratory activities, presumably by coordinating protein synthesis and translation specific mRNAs with actin polymerization.

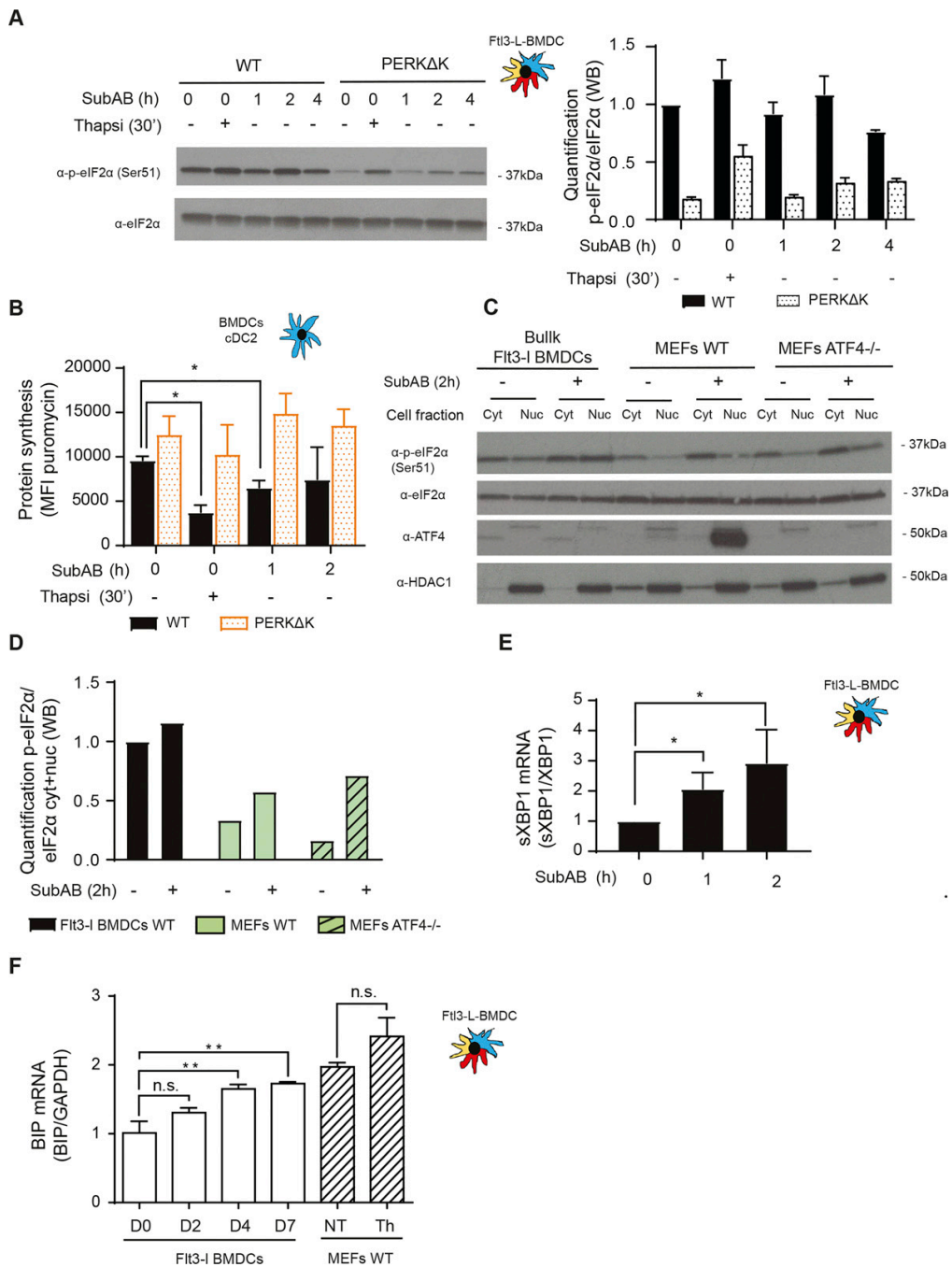
## Discussion

We have previously proposed the existence of a strong causative link between cell activation by TLR ligands and eIF2α phosphorylation, notably by virtue of strong GADD34 expression in most transcriptomics studies performed on PAMP-activated DCs (Clavarino et al, 2012b; Claudio et al, 2013; Reverendo et al, 2018). Our present work suggests that steady-state DCs activate PERK-mediated eIF2α phosphorylation to acquire their functional properties during differentiation (Fig S8), but distinctly from known ISR programs, normally induced upon acute or chronic ER stress (Han et al, 2013; Guan et al, 2017).

To our knowledge, the level of p-eIF2α observed in primary DC both in vivo and in vitro are unique in their amplitude. Although as judged comparatively from experiments performed with artificial induction of the different EIF2KAs, such p-eIF2α levels should be inhibitory for global protein synthesis (Dalet et al, 2017). DCs have acquired biochemical resistance, like high expression of eIF2B and eIF2A, to compensate for the consequences of this developmental PERK activation and to undergo high eIF2α phosphorylation, whereas maintaining normal proteostasis.

ATF4's role in controlling *Ppp1r15a/GADD34* mRNA expression and eIF2α dephosphorylation to restore protein synthesis during stress has been extensively studied (Novoa et al, 2001). Despite high eIF2α phosphorylation, the active translation observed in DCs does not seem to allow ATF4 synthesis and consequently the activation of a bona fide ISR in these cells. In contrast, GADD34 is functional in non-activated DCs with IKKε/TBK1 activity required for its mRNA transcription. This dependency of *Ppp1r15a/GADD34* transcription on IKKε/TBK1 confirms that the *PPP1R15a* gene belongs to a group of genes directly induced by TLR or RLR signaling, as previously suggested by genomic analysis of viral or poly (I:C)-stimulated cells (Freaney et al, 2013; Lazear et al, 2013; Dalet et al, 2017). GADD34 protein expression is undetectable in DCs, without prior treatment with proteasome inhibitors (Clavarino et al, 2012b), which is

*B. Article: Proteostasis in dendritic cells is controlled by the PERK signaling axis independently of ATF4*



**Figure 6. Flt3-L BMDCs are resistant to subtilase cytotoxin-induced integrated stress response.** WT and PERKΔK Flt3-L BMDCs were stimulated with thapsigargin (200 nM) and subtilase cytotoxin (SubAB, 250 ng/ml) for the indicated times. **(A)** Levels of p-eIF2α and total eIF2α detected by immunoblot (left) and quantified (right). **(B)** Levels of protein synthesis measured by flow using puromycylation detection in cDC2. Cells were incubated with puromycin 10 min before harvesting. The graph shows the total puromycin mean fluorescence intensity levels. **(C)** WT Flt3-L BMDCs, WT and ATF4<sup>-/-</sup> MEFs were stimulated with subtilase cytotoxin (SubAB, 250 ng/ml) for 2 h. Both cytoplasmic (cyt) and nuclear (nuc) fractions were analyzed. Levels of p-eIF2α, total eIF2α, ATF4, and HDAC1 (nuclear loading control) were revealed by immunoblot. **(D)** p-eIF2α quantification is represented in (D). **(E)** WT Flt3-L BMDCs were stimulated with subtilase cytotoxin (SubAB, 250 ng/ml) for the indicated times. mRNA levels of spliced XBP1 were measured by qRT-PCR in bulk Flt3-L BMDCs. Raw data were normalized to total XBP1

## B. Article: Proteostasis in dendritic cells is controlled by the PERK signaling axis independently of ATF4



presumably indicative of an extremely fragile equilibrium between its translation and active degradation. GADD34 mRNA translation, like that of ATF4, is controlled through 5' upstream ORFs regulation, which are normally bypassed upon general translation arrest to favor the synthesis of these specific ISR molecules (Palam et al, 2011). Interestingly, GADD34 mRNA has been recently shown to be also actively translated in unstressed MEFs, albeit at much lower levels than upon ER stress (Reid et al, 2016). Thus, in steady-state DCs, GADD34 synthesis likely occurs in a high eIF2 $\alpha$ -phosphorylation context, despite relatively normal level of translation, whereas that of ATF4 does not. The difference in the 5' upstream ORFs organization of the mRNAs coding for these two molecules (Palam et al, 2011; Andreev et al, 2015), could explain this difference, and how GADD34/PP1c activity contributes to the maintenance of protein synthesis activity by counteracting PERK in steady-state DCs.

Importantly, the PERK/eIF2 $\alpha$ /GADD34 molecular trio sets the physiological range for potential protein synthesis initiation available in the different DC activation stage; however, the upward progression triggered by LPS stimulation, from one level of protein synthesis to the next, does not depend on this biochemical axis and is likely regulated by other protein synthesis regulation pathways, like the mTORC1 or casein kinase 2 pathways (Lelouard et al, 2007; Reverendo et al, 2019) (Fig S10). Other mechanisms that contribute to escape PERK-mediated eIF2 $\alpha$  phosphorylation, including the eIF2B-independent and eIF3-dependent pathway recently described to rescue translation during chronic ER stress (Guan et al, 2017), do not seem to be used in the DC context. Given the amount of eIF2A and eIF2B expressed by differentiated DCs, these factors are likely to be sufficient to counteract excessive eIF2 $\alpha$  phosphorylation and maintain protein synthesis level in DCs. This activity could be equivalent to the TLR-dependent activation of eIF2B through PP2A-mediated dephosphorylation of the eIF2B $\epsilon$ -subunit, which prevents translation arrest in tunicamycin treated macrophages (Woo et al, 2012).

These DC-specific mechanisms prevent the induction of the ISR by the AB5 subtilase cytotoxin, which targets the ER chaperone BiP. Independently of demonstrating that induction of acute eIF2 $\alpha$  phosphorylation by thapsigargin is not solely dependent on PERK activation, our observations suggest that DCs could escape EIF2KA-dependent translation arrest during exposure to different metabolic insults relevant to the immune context. These situations can include viral infection (Clavarino et al, 2012b), exposure to high levels of fatty acids during pathogenesis, oxidative stress during inflammation or amino acids starvation, mediated by amino acid-degrading enzymes, such as arginase 1 or IDO, which are induced during infection or cancer development (Munn et al, 2004; Claudio et al, 2013). Importantly we could also show, which the ISR induction is not necessary for DCs to drive the transcription of pro-inflammatory cytokines in response to TRIF or MAVs dependent-signaling (Abdel-Nour et al, 2019), nor the secretion of IL-1 $\beta$  (Chiritoiu et al, 2019).

PERK activation is therefore required to regulate mRNA translation during DC differentiation and potentially also GADD34 synthesis, which not only provides a negative feed-back to PERK, but

also is required for normal DC activation and cytokines expression (Clavarino et al, 2012b; Perego et al, 2018). Although a role for the ISR has been suggested to favor the survival of tissue associated DCs (Tavernier et al, 2017), we could not detect any particular phenotype impairing the development of DCs in the spleen of PERK-deficient animals. A close examination of the functional capacity of DC in vitro showed that although soluble antigen presentation was not affected by PERK inactivation, it induced nevertheless an alteration of IFN- $\beta$  and cytokines production as well as of DC migratory capacity upon MAMPs activation. Interestingly, ROCK-induced actomyosin contractility in transformed fibroblasts enhances signaling through PERK and ATF4 (Boyle et al, 2020), whereas PERK itself has been shown to interact with filamin-A and to participate to F-actin remodeling in MEFs (van Vliet et al, 2017), suggesting that the migratory deficit observed in PERK-deficient DCs could be also dependent on these interactions. This finding echoes with the existence of a cross-talk between actin skeleton organization and the main molecular actors involved in the ISR (Chambers et al, 2015; Chen et al, 2015). We confirmed that globular actin synergizes with the PP1c to dephosphorylate p-eIF2 $\alpha$ , suggesting that the PERK/GADD34 pathway could play an important role in regulating translation in response to actin dynamics and possibly in coordinating migration or interactions with T cells.

DCs therefore represent a model of choice for studying this possibility, given their developmental regulation of eIF2 $\alpha$  phosphorylation and their requirement for actin dependent-phagocytosis and migration to perform their immune-stimulatory function. The activation of PERK/GADD34 pathway in steady-state DCs also underlines the importance of these molecules in homeostatic condition, independently of obvious acute ER stress, for the acquisition of specialized function. Clearly, the use of PERK-deficient cells versus pharmacological inhibition creates some discrepancies on how several biochemical functions in DCs are truly affected directly by PERK loss or reduced eIF2 $\alpha$  phosphorylation. Our findings open nevertheless new pharmacological perspectives for therapeutic immune intervention by targeting PERK, GADD34, or eIF2 $\alpha$  phosphorylation.

## Materials and Methods

### Cell culture

BM was collected from 6- to 9-wk-old female mice and differentiated in DCs or macrophages during 7 d. The culture was kept at 37°C, with 5% CO $_2$  in Roswell Park Memorial Institute medium (RPMI) (GIBCO), 10% FCS (Sigma-Aldrich), 100 U/ml penicillin, 100 U/ml streptomycin (GIBCO), and 50  $\mu$ M  $\beta$ -mercaptoethanol (VWR) supplemented with Flt3-L, produced using B16-Flt3-L hybridoma cells for DC differentiation or M-CSF for macrophages, as described previously (Wang et al, 2013). For the migration assays, GM-CSF was used instead of Flt3-L and cells were cultured during 10–12 d with

mRNA expression. (F) Levels of BiP mRNA expression measured by qRT-PCR during Flt3-L BMDcs differentiation and in MEFs stimulated with thapsigargin for 30 min. Data are mean  $\pm$  SD (n = 3). Statistical analysis was performed using Dunnett's multiple comparison (\*P < 0.05 and \*\*P < 0.01).

## B. Article: Proteostasis in dendritic cells is controlled by the PERK signaling axis independently of ATF4



changes in the medium each 3 d. GM-CSF was obtained from transfected J558 cells (Pierre et al, 1997). To obtain splenocytes, spleens were collected and injected with Liberase TL (Roche) and incubated 25 min at 37°C to disrupt the tissues. DC purification was performed using a CD11c+ positive selection kit (Miltenyi), according to the manufacturer's instructions and CD8 $\alpha$ + T-cell isolation was performed with a Dynabeads untouched mouse CD8 T cells kit from Thermo Fisher Scientific. CD8 $\alpha$ + T where incubate overnight with anti-CD3 (10  $\mu$ g/ml) and anti-CD28 (5  $\mu$ g/ml) antibodies, to mimic activation by APCs. MEFs used in this work, ATF4 $^{-/-}$  and matched WT (129 SvEv) were a kind gift from Prof. David Ron (Cambridge Institute for Medical Research). PERK KO $^{-/-}$  and matched WT were a kind gift from Prof. Douglas Cavener (Penn State University). MEFs were cultured in DMEM medium (GIBCO) with 5% FBS (Sigma-Aldrich) and 50  $\mu$ M 2-mercaptoethanol. For the experimental assays, cells were plated from 16 to 24 h before stimulation in six well plates, at 150,000 cells/ml in 2 ml of the same medium. After stimuli, cells were treated with trypsin-EDTA for 2 min at 37°C before washing to detach cells from the wells.

### Reagents

LPS (*E. coli* O55:B5), cycloheximide, puromycin, MRT67307, GSK2656157, rocaglamide, and thapsigargin were purchased from Sigma-Aldrich. Harringtonine is from ABCAM, Latrunculin A, and Jaspilakinolide are from Merck-Millipore. Low molecular weight polyinosinic-polycytidylic acid (LMW poly(I:C)) was from InvivoGen. SAR1 was kindly provided by Sanofi and Integrated Stress Response Inhibitor (ISRIB) was a gift from Carmela Sidrausky and Peter Walter (UCSF). Subtilase cytotoxin (Shiga toxicogenic *E. coli* strains) was purified from recombinant *E. coli*, as previously described (Paton et al, 2004). 4EGI-1 was purchased by Bertin bioreagent. HEL and the peptide HEL 46-61 were purchased from Thermo Fisher Scientific.

### Flow cytometry analysis

Cell suspensions were washed and incubated with a cocktail of coupled specific antibodies for cell surface markers in flow activated cell sorting (FACS) buffer (PBS, 1% FCS, and 2 mM EDTA) for 30 min at 4°C. For Flt3-L BMDcs, the antibodies used were CD11c (N418), Siglech (551), CD86 (GL-1), F4/80 (BM8), CD64 (X54-5/7.1) from BioLegend; Sirp $\alpha$  (P84), CD24 (M1/69), and MHC II (M5/114.15.2) from eBioscience CD11b (M1/70) from BD Bioscience. For splenic cells, the antibodies used were Nkp46 (29A14), CD4 (RM4-5), CD3 (145-2C11), CD11c (N418), CD19 (eBio1D3), CD8 $\alpha$  (53-6.7) from eBioscience, BST2 (927), Ly6G, F4/80, Ly6C from BioLegend; CD11b (M1/70), B220 (RA3-6B2), and CD69 (H12F3) from BD Biosciences. These antibodies were used in combination with the LIVE/DEAD Fixable Aqua Dead Cell Stain (Thermo Fisher Scientific). For intracellular staining, cells were next fixed with BD Phosflow Fix Buffer I (BD Biosciences) during 10 min at room temperature and washed with 10% Perm/wash Buffer I 1 $\times$  (BD Biosciences). Permeabilized cells were blocked during 10' with 10% Perm/wash buffer 1 $\times$ , 10% FCS, before staining with primary antibodies. When the primary antibody was not coupled, cells were washed after and blocked during 10 min with Perm/wash buffer 1 $\times$ , 10% FCS, and 10% of serum from the species where the secondary antibody was produced. Then, the incubation with the secondary antibody was performed at 4°C during 30 min. p-eIF2 $\alpha$ (S51) was purchased from ABCAM and

p-eEF2(Thr56) from Cell Signaling and Deoxyribonuclease (DNase I) was purchased from Invitrogen. Data were acquired on an LSR-II/UV instrument using FACS Diva software. The acquired data were analyzed with FlowJo software (BD Biosciences).

### Translation intensity and speed measurement

SUnSET technique to measure the intensity of protein synthesis was used as previously described (Schmidt et al, 2009). Puromycin was added in the culture medium at 12.5  $\mu$ g/ml, and the cells were incubated for 10 min at 37°C and 5% CO $_2$  before harvesting. Cells were washed with PBS before cell lysis and immunoblotting with the anti-puromycin 12D10 antibody (Merck Millipore). For flow cytometry (flow) cells were processed, as described below for the intracellular staining, using the  $\alpha$ -puromycin 12D10 antibody directly conjugated with Alexa 488 or A647 from Merck Millipore. The SUnRISE technique was performed as described (Arguello et al, 2018). Samples were treated with 2  $\mu$ g/ml of harringtonine at different time points (90, 60, 45, 30, 15, and 0 s) and then treated for 10 min with 12.5  $\mu$ g/ml of puromycin. For the measurement of Cap-dependent translation, the cells were treated with 4EGI-1 (100  $\mu$ M) or rocaglamide (RocA-1) (100 nM) for 0.5, 1, 2, or 4 h. Cells were then incubated for 15 min at 37°C and stained with the 12D10 antibody (Merck-Millipore).

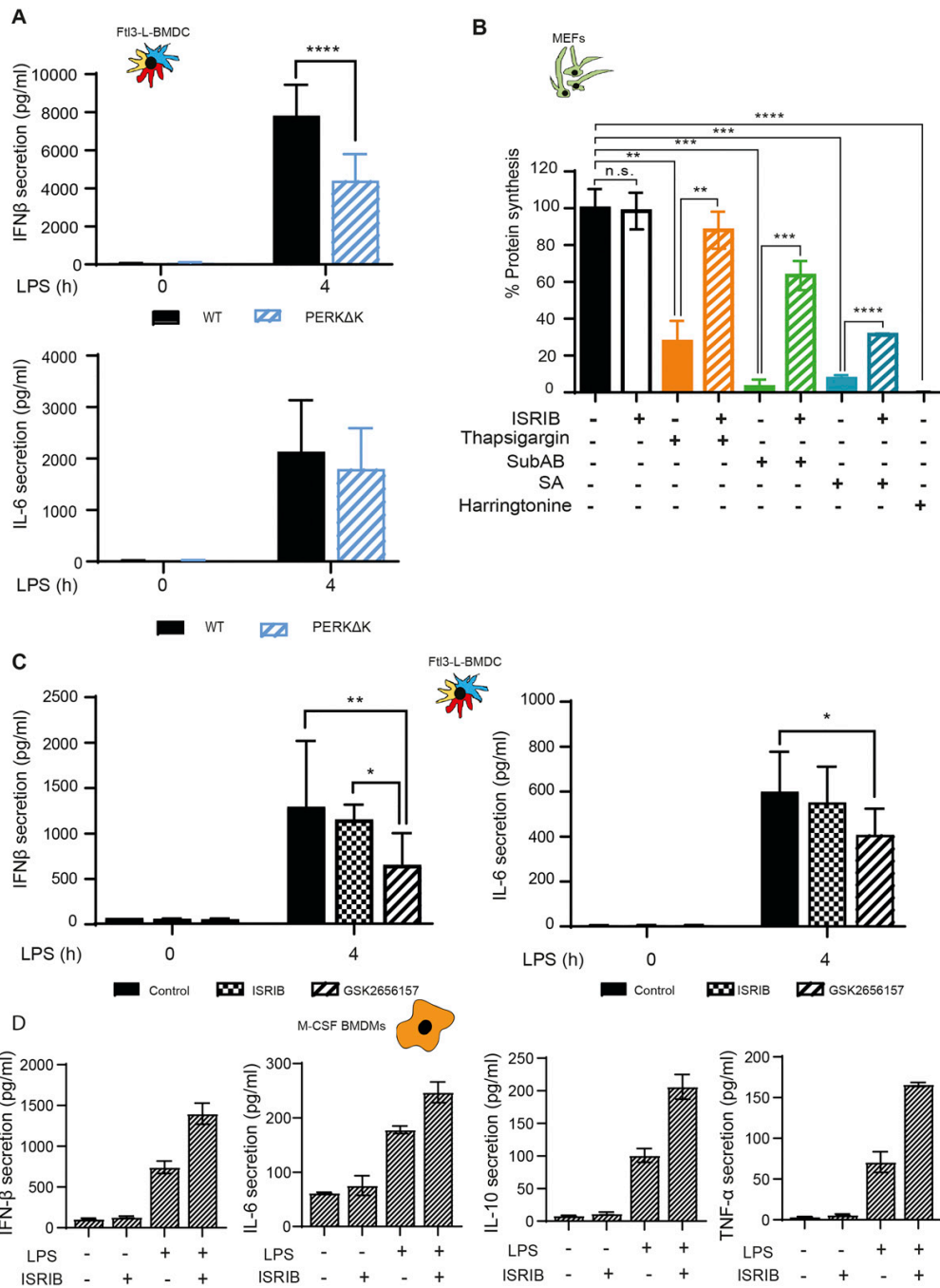
### Gene expression analysis

Total RNA was extracted from the DCs using the RNeasy Mini Kit (QIAGEN), including a DNA digestion step with RNase-free DNase (QIAGEN), and cDNA was synthesized using the Superscript II Reverse Transcriptase (Invitrogen). Quantitative PCR amplification was performed using SYBR Green PCR master mix (Takara) using 10 ng of cDNA and 200 nM of each specific primer on a 7500 Fast Real-PCR system (Applied Biosystems). cDNA concentration in each sample was normalized to GAPDH expression. The primers used for gene amplification were the following: GADD34 (S 5'-GACCCCTCC AACTCTCCTC-3', AS 5'-CTTCCTCAGCCTCAGCATT-3'); IL-6 (S 5'-CAT GTTCTCTGGGAAATCGTG-3', AS 5'-TCCAGTTTGGTAGCATCCATC-3'); IFN- $\beta$  (S 5'-CCCTATGGAGATGACGGAGA-3', AS 5'-ACCCAGTCTGGAGAAATTG-3'); IL-12 (S 5'-GGAATGCTGCGTGAAGCT-3', AS 5'-ACATGCCCACTTGTGCAT-3'); ATF4 (S 5'-AAGGAGGATGCCCTTTCCGGG-3', AS 5'-ATTGGGTTCACT GTCTGAGGG-3'); CHOP (S 5'-CACTCCGGAGAGACAGACAG-3', AS 5'-ATGA AGGAGAAGGAGCAGGAG-3'); PERK (S 5'-CGGATTCATTGAAAGCACCT-3', AS 5'-ACGCGATGGGAGTACAAAAC-3'); XBP1 (S 5'-CCGCAGCACTCAGACATATG-3', AS 5'-GGGTCCAACCTGTCCAGAAT-3'); spliced XBP1 (S 5'-CTGAGT CCGCAGCAGGT-3', AS 5'-AACATGACAGGGTCCAACCT-3'); GAPDH (S 5'-TGGAGAAACCTCCAAGTATG-3', AS 5'-GTTGAAGTCGAGGAGACAAC-3'); IL1- $\beta$  (S 5'-TGATGTGCTGCTGCGAGAGATT-3', AS 5'-TGCCATTTTGACAGTGA-3'); eIF2B $\epsilon$  (S 5'-GAGCCCTGGAGGAACACAGG-3' AS 5'-CACCACGTTGT CCTCATGGC-3'); BIP S (5'-ATTGGAGTGGGCAACCAA-3' AS 5'-TCGGTG GGCATATTGAAGT-3').

### GSEA

The GSEA was performed using published murine microarray datasets accessible through the Gene Expression Omnibus repository under the references GSE9810 (Robbins et al, 2008) and

*B. Article: Proteostasis in dendritic cells is controlled by the PERK signaling axis independently of ATF4*

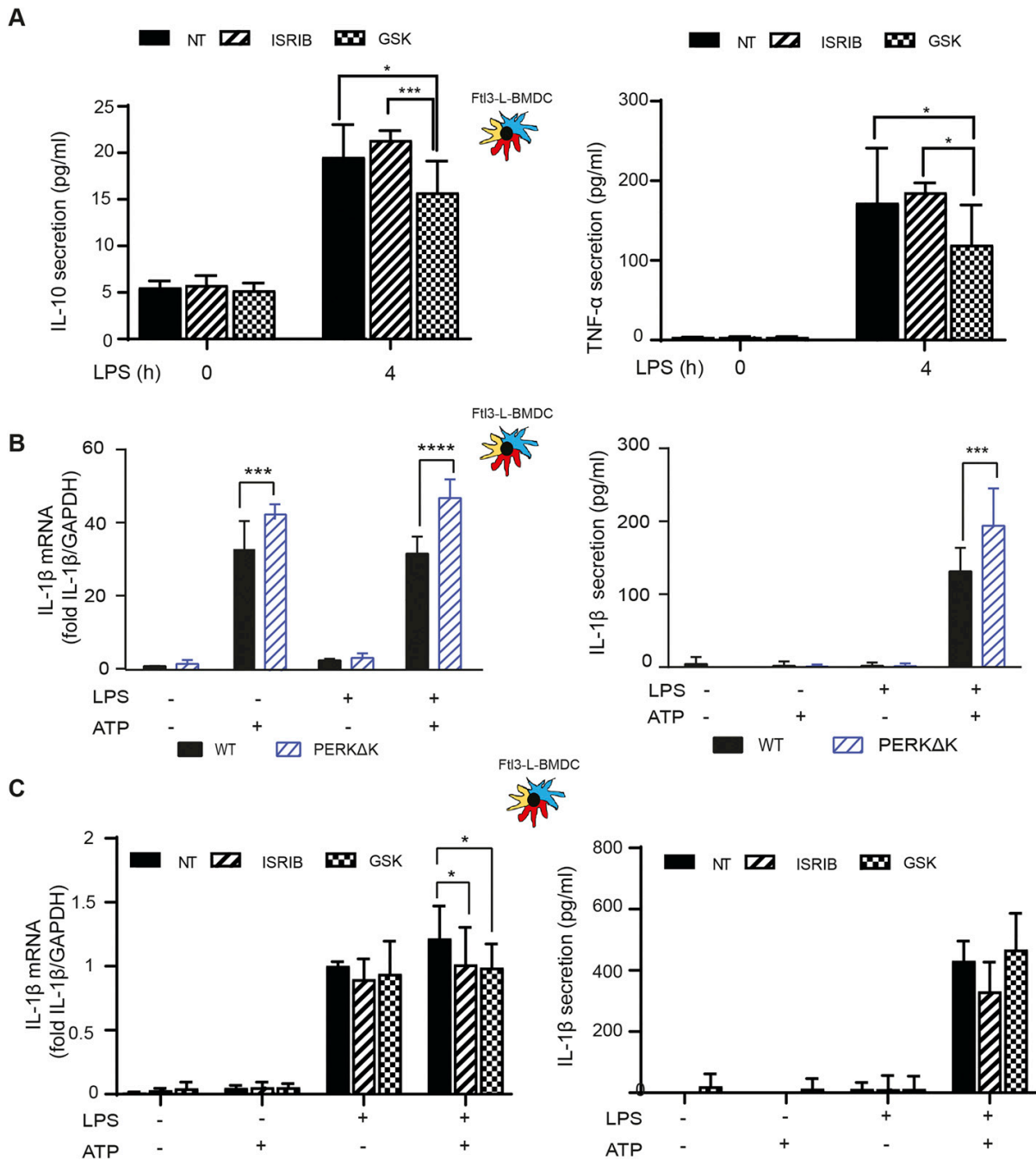


**Figure 7. ISIRIB does not inhibit cytokines expression in LPS activated DCs and Macrophages.**

(A) IFN- $\beta$  and IL-6 secretion was measured by Legendplex in WT and PERK $\Delta$ K Flt3-L BMDCs stimulated with LPS during 4 h. (B) Protein synthesis was measured by puromycylation and flow in MEFs treated with ISIRIB and different the integrated stress response inducing drugs, thapsigargin (Tg), subtilase cytotoxin (SubAB), and sodium arsenite for indicated times. (C, D) IFN- $\beta$ , IL-6, IL-10, and TNF secretion was measured by Legendplex in (C) Flt3-L BMDCs and (D) M-CSF BMDM. Data are mean  $\pm$  SD (n = 3). Statistical analysis was performed using Wilcoxon test (\* $P$  < 0.05 and \*\* $P$  < 0.01).



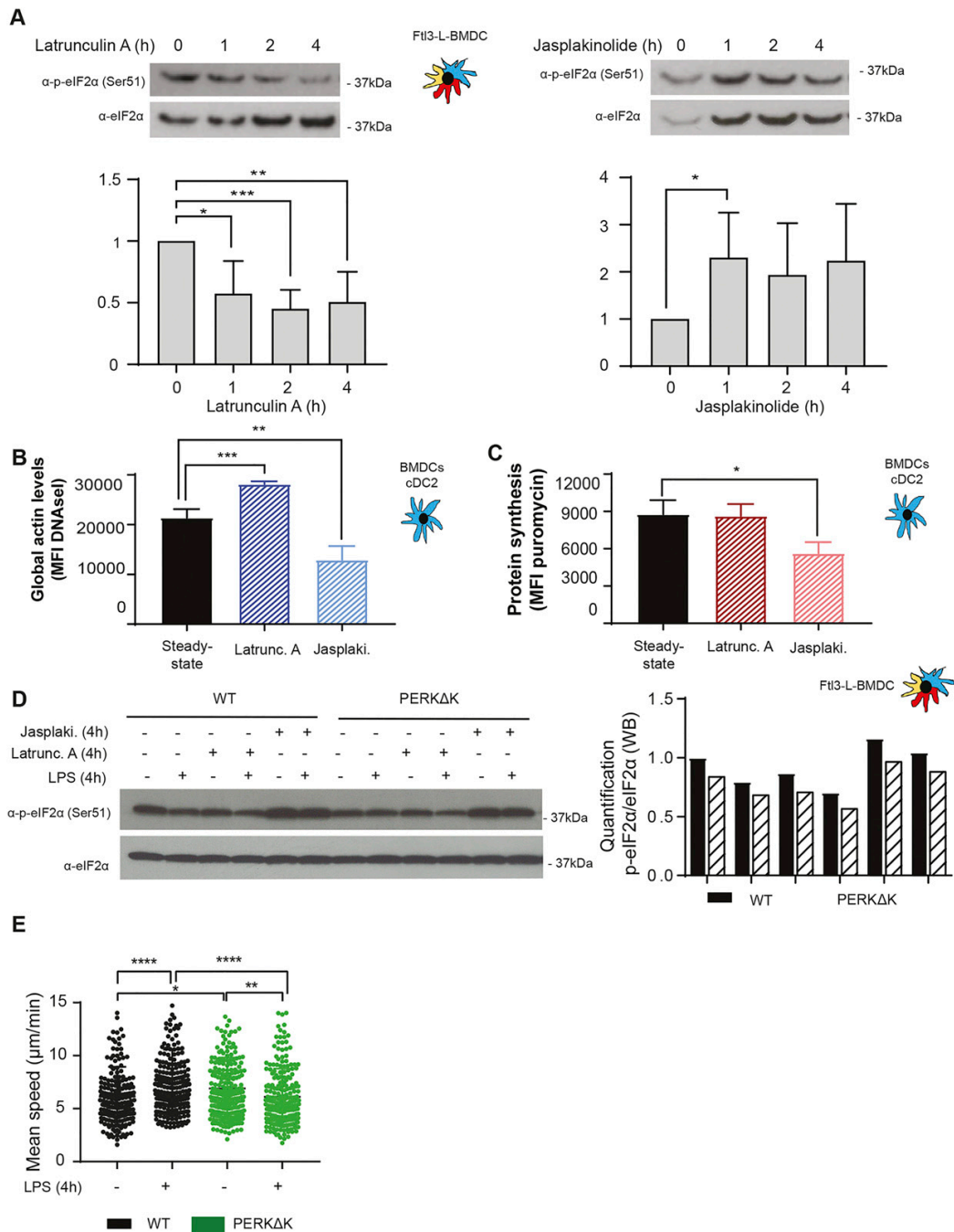
*B. Article: Proteostasis in dendritic cells is controlled by the PERK signaling axis independently of ATF4*



**Figure 8. PERK inactivation and cytokines expression.**

(A) IL-10 and TNF secretion was measured by Legendplex in Flt3-L BMDCs activated with LPS and treated or not with ISRIB or GSK2656157 for 4 h. (B) IL-1β mRNA expression measured by qRT-PCR (left) and secretion by ELISA (right) in WT and PERKΔK Flt3-L BMDCs stimulated with LPS during 4 h and with ATP for the last 30 min of treatment. (C) IL-1β mRNA expression measured by qRT-PCR (left) and secretion by ELISA (right) in Flt3-L BMDCs activated with LPS and treated with GSK2656157 and/or ISRIB for 4 h and with ATP for the last 30 min of treatment. (A, B, C) Statistical analysis was performed using Dunnett's multiple comparison (A, B) and Wilcoxon test (C). Data are mean ± SD (n = 3). independent experiments (\*P < 0.05, \*\*P < 0.01, and \*\*\*P < 0.001).

*B. Article: Proteostasis in dendritic cells is controlled by the PERK signaling axis independently of ATF4*



**Figure 9. eIF2 $\alpha$  phosphorylation levels are regulated by G-actin availability.**

WT Flt3-L BMDCs were treated with Latrunculin A (50 nM) (Latrunc A) and Jasplakinolide (1  $\mu$ M) (Jasp) for the indicated times. **(A)** Levels of p-eIF2 $\alpha$  and total eIF2 $\alpha$  detected by immunoblot in Flt3-L BMDCs. **(B)** Globular actin levels were measured by flow cytometry intracellular staining using a fluorochrome coupled DNase I protein in cDC2 population from Flt3-L BMDCs. The graph shows total mean fluorescence intensity levels. **(C)** Levels of protein synthesis in cDC2 were measured by puromycylation and flow intracellular staining. Cells were incubated with puromycin 10 min before harvesting. The graph shows total puromycin mean fluorescence intensity levels. WT and PERK $\Delta$ K Flt3-L BMDCs were treated with LPS (100 ng/ml), Latrunc A, and Jasp for 4 h. **(D)** Levels of p-eIF2 $\alpha$  and total eIF2 $\alpha$  monitored by immunoblot in Flt3-L BMDCs.

## B. Article: Proteostasis in dendritic cells is controlled by the PERK signaling axis independently of ATF4



GSE2389 (Fontenot et al, 2005). Raw microarray data describing CD8 $\alpha$ <sup>+</sup> cDCs (DC1), CD11b<sup>+</sup> cDC (DC2), pDC (plasmacytoid DCs), B cells, NK cells, CD8<sup>+</sup> T cells (Robbins et al, 2008), and CD4<sup>+</sup> T cells (Fontenot et al, 2005) in mice spleen were downloaded. For each of these cell types, the hybridization was performed using Affymetrix mouse 4302.0 gene chips. Microarray data were normalized by Robust Multi-array Average algorithm (Irizarry et al, 2003) using the oligo BioconductorR package (Carvalho & Irizarry, 2010). Normalization consists of a background correction of raw intensities, a log<sub>2</sub> transformation followed by quantile normalization to allow the comparison of each probe for each array. Before data usage, the absence of batch effect was assessed by Principal Component Analysis using ade4R package (Bougeard & Dray, 2018).

GSEA was performed using publicly available gene signatures reflecting an ISR state (Tables S1–S3). Lists of ATF4 and CHOP target genes identified by ChIP-seq experiments (Han et al, 2013) have been used to search for ATF4-dependent and CHOP-dependent signatures in DCs (Tables S2 and S3). Lists of genes for which a translational up-regulation after 1 h of thapsigargin (Tg) treatment and congruent (both transcriptional and translational) up-regulations after 16 h of Tg treatment (Guan et al, 2017) were used to search gene expression signatures, respectively, of acute and/or cISRs in DCs. GSEA was generated using BubbleGUM (Spinelli et al, 2015). Briefly, GSEA pairwise comparisons are performed for each probe and the multiple testing effects are corrected using a Benjamini–Yekutieli procedure. The corrected *P*-values are hence calculated based on a null hypothesis distribution built from the permutations of the gene sets across all the pairwise comparisons. In our analyses, 10,000 permutations of the gene sets have been performed to compute the *P*-values. All results with a FDR below the threshold of 0.25 have been considered as significant.

### Immunoblotting

Cells were lysed in RIPA buffer (25 mM Tris–HCl, pH 7.6, 150 mM NaCl, 1% NP-40; 1% sodium deoxycholate, 0.1% SDS) supplemented with Complete Mini Protease Inhibitor Mixture Tablets (Roche), NaF (Ser/Thr and acidic phosphatase inhibitor), Na<sub>3</sub>VO<sub>4</sub> (Tyr and alkaline phosphatase inhibitor) and MG132 (proteasome inhibitor). The nuclear extraction was performed using the Nuclear Extract kit (Active Motif) according with manufacturer's instructions. Protein quantification was performed using the BCA Protein Assay (Pierce). Around 20 μg of soluble proteins were run in 4–20% acrylamide gradient gels and for the immunoblot the concentration and time of incubation had to be optimized for each individual antibody. Rabbit antibodies against eIF2 $\alpha$ , p-eIF2(Thr56), eEF2, eIF2B, p-IRF3 (ser396), IRF3, p-S6, and PERK were purchased from Cell Signaling (ref 5324, 2331, 2332, 3592, 4947, 4302, 2211, and 3192, respectively). Rabbit antibody against p-eIF2 $\alpha$ (S51) was purchased from ABCAM (Ref 32157). Rabbit antibody against ATF4 was purchased from Santa Cruz

Biotechnology (sc-200). Mouse antibody against  $\beta$ -actin was purchased from Sigma-Aldrich (A2228). Mouse antibodies against HDAC1 and S6 were purchased from Cell Signaling (ref 5356, 2317, respectively). Mouse antibody against puromycin was purchased from Merck Millipore (MABE343). Mouse antibody against p-eIF2 $\beta$  was a kind gift from David Litchfield (University of Western Ontario). Mouse antibody against eIF2 $\beta$  was purchased from Santa Cruz Biotechnology (sc-9978). HRP secondary antibodies were from Jackson ImmunoResearch Laboratories.

### Cytokine measurement

The IL-6, IFN- $\beta$ , IL1- $\beta$ , and IL-2 quantifications from the cell culture supernatant were performed using the mouse Interleukin-6 ELISA Kit (eBioscience), the mouse IFN- $\beta$  ELISA Kits (PBL-Interferon Source or Thermo Fisher Scientific), the mouse IL1- $\beta$ , and mouse IL2 uncoated ELISA kits (Invitrogen) according to the manufacturer's instructions. Cytokine monitoring was also performed using Legendplex 740150 (BioLegend).

### Antigen presentation assays

FIt3-L-differentiated bmDC obtained from C3H/HeN were treated with indicated concentration of HEL or with 5 μM of 46-61 HEL peptide and incubated for 8 h with 100 nM of LPS in presence or not of GSK2656157. DCs were fixed mildly with 0.25% PFA, 2 min, RT, and prior quenching with 10 mM glycine. DCs were co-cultivated with 3A9 (HEL 46-61 on I-Ak) specific T hybridoma at 5:1 (T:DC) ratio for 18 h. CD69 up-regulation and IL-2 production was determined, respectively, by cytometry and ELISA.

### Immunohistochemistry

Spleens were snap frozen in Tissue Tek (Sakura Finetek). Frozen sections (8 μm) were fixed with acetone permeabilized with 0.05% saponin. The following antibodies were used for the staining: CD11c (N418) from BioLegend (Ref 117301), p-eIF2 $\alpha$  (Ser 52) from Invitrogen (Ref 44-728G), CD11b (M1/70) from BD Biosciences, CD8 $\alpha$ -biotin (53-6.7) BioLegend (Ref 100703), and B220 (RA3-6B2) from Invitrogen (Ref 14-0452-81). Images were collected using a Zeiss LSM 510 confocal microscope. Image processing was performed with Zeiss LSM software.

### Mice

Wild-type (WT) female C57BL/6 and C3H/HeN mice were purchased from Janvier. PKR<sup>-/-</sup> C57BL/6 were a kind gift from Dr Bryan Williams (Hudson Institute of Medical Research) (Kumar et al, 1997). PERK<sup>loxP/loxP</sup> mice were the kind gift of Dr Doug Cavener (Zhang et al, 2006) and purchased from Jackson Laboratories. GADD34 $\Delta$ C<sup>loxP/loxP</sup> mice were developed at the Centre d'Immunophénomique (CIPHE).

Quantification is shown on the right. (E) WT and PERK $\Delta$  GM-CSF BMDCs were treated with LPS (100 ng/ml) during 30 min previous to 16 h of migration. The graph represents instantaneous mean velocities of migration in 4 × 5-μm fibronectin-coated microchannels of at least 100 cells per condition. All data are representative of n = 3 independent experiments. (A, B, C) Statistical analysis was performed using Dunnett's multiple comparison (A) and Mann–Whitney test (B, C). Data are mean ± SD (n = 3). \**P* < 0.05, \*\**P* < 0.01, \*\*\**P* < 0.001, and \*\*\*\**P* < 0.0001.



## B. Article: Proteostasis in dendritic cells is controlled by the PERK signaling axis independently of ATF4



PERK<sup>loxp/loxp</sup> and GADD34<sup>loxp/loxp</sup> were crossed with Itgax-Cre+ mice (Caton et al, 2007) and backcrossed, to obtain stable homozygotic lines for the loxp sites expressing Cre. For all studies, age-matched WT and transgenic 6–9 wk females were used. All animals were maintained in the animal facility of CIML or CIPHE under specific pathogen-free conditions accredited by the French Ministry of Agriculture to perform experiments on live mice. These studies were carried out in strict accordance with Guide for the Care and Use of Laboratory Animals of the European Union. All experiments were approved by the Comité d’Ethique PACA and MESRI (approval number APAFIS#10010-201902071610358). All efforts were made to minimize animal suffering.

### Preparation of microchannels and speed of migration measurement

Microchannels were prepared as previously described (Vargas et al, 2016). For velocity measurements (carried out in 4-by-5  $\mu$ m microchannels), phase-contrast images of migrating cells were acquired during 16 h (frame rate of 2 min) on an epifluorescence Nikon Ti-E video microscope equipped with a cooled charge-coupled device camera (HQ2; Photometrics) and a 10 $\times$  objective. Kymographs of migrating cells were generated and analyzed using a custom program.

### Statistical analysis

Statistical analysis was performed using GraphPad Prism Software. The most appropriate statistical test was chosen according to each data set. Mainly, we used Wilcoxon test, Mann–Whitney test, t test, and multiple comparison with Dunnett’s correction. \* $P < 0.05$ , \*\*\* $P < 0.001$ , \*\*\*\* $P < 0.0001$ .

## Supplementary Information

Supplementary Information is available at <https://doi.org/10.26508/lsa.202000865>.

## Acknowledgements

We thank all the Centre d’Immunologie de Marseille-Luminy (CIML) cytometry and Imaging core facilities for expert assistance. The laboratory is supported by grants from La Fondation de l’Association pour la Recherche sur le Cancer (ARC). The laboratory is “Equipe de la Fondation de la Recherche Médicale” (FRM) sponsored by the grant DEQ20140329536. The project was also supported by grants from l’Agence Nationale de la Recherche (ANR), « ANR-FCT 12-ISV3-0002-01 » and « INFORM Labex ANR-11-LABEX-0054 », « DCBIOL Labex ANR-11-LABEX-0043 » and ANR-10-IDEX-0001-02 PSL\* and A\*MIDEX project ANR-11-IDEX-0001-02 funded by the “Investissements d’Avenir” French government program. Grant from French Agency for Research on AIDS and Viral Hepatitis (ANRS) ECTZ88500 “SMARTHCV” also supported this project. The research is supported by the Ilídio Pinho foundation, Maratona da Saúde and FCT—Fundação para a Ciência e a Tecnologia—and Programa Operacional Competitividade e Internacionalização—Compete2020 (FEDER)—references PTDC/BIA-CEL/28791/2017 and POCI-01-0145-FEDER-028791, POCI-01-0145-FEDER-030882 and PTDC/BIA-MOL/30882/2017 and UIDB/04501/2020. We thank Lionel Spinelli

and Thien-Phong Vu-Manh at CIML for bioinformatics and statistics support. We acknowledge financial support from no ANR-10-INBS-04-01 France Bio Imaging and the ImagImm CIML imaging core facility. The authors declare to have no competing interest.

### Author Contributions

A Mendes: conceptualization, formal analysis, validation, investigation, methodology, and writing—original draft, review, and editing.

JP Gigan: conceptualization, formal analysis, investigation, methodology, and writing—original draft, review, and editing.

C Rodriguez Rodrigues: conceptualization, formal analysis, investigation, and methodology.

SA Choteau: software, formal analysis, validation, investigation, and methodology.

D Sanseau: conceptualization, formal analysis, investigation, and methodology.

D Barros: conceptualization, formal analysis, investigation, and methodology.

C Almeida: conceptualization and investigation.

V Camosseto: conceptualization, formal analysis, investigation, and methodology.

L Chasson: formal analysis, investigation, and methodology.

AW Paton: resources.

JC Paton: resources.

RJ Argüello: conceptualization, investigation, and methodology.

A-M Lennon-Duménil: conceptualization and methodology.

E Gatti: formal analysis.

P Pierre: conceptualization, formal analysis, supervision, funding acquisition, validation, investigation, methodology, writing—original draft, and project administration.

### Conflict of Interest Statement

The authors declare that they have no conflict of interest.

## References

- Abdel-Nour M, Carneiro LAM, Downey J, Tsalikis J, Outlioua A, Prescott D, Da Costa LS, Hovingh ES, Farahvash A, Gaudet RG, et al (2019) The heme-regulated inhibitor is a cytosolic sensor of protein misfolding that controls innate immune signaling. *Science* 365: eaaw4144. doi:10.1126/science.aaw4144
- Adomavicius T, Guaita M, Zhou Y, Jennings MD, Latif Z, Roseman AM, Pavitt GD (2019) The structural basis of translational control by eIF2 phosphorylation. *Nat Commun* 10: 2136–2144. doi:10.1038/s41467-019-10167-3
- Akira S, Uematsu S, Takeuchi O (2006) Pathogen recognition and innate immunity. *Cell* 124: 783–801. doi:10.1016/j.cell.2006.02.015
- Andreev DE, O’Connor PB, Fahey C, Kenny EM, Terenin IM, Dmitriev SE, Cormican P, Morris DW, Shatsky IN, Baranov PV (2015) Translation of 5’ leaders is pervasive in genes resistant to eIF2 repression. *Elife* 4: e03971. doi:10.7554/elife.03971
- Arguello RJ, Reverendo M, Gatti E, Pierre P (2016) Regulation of protein synthesis and autophagy in activated dendritic cells: Implications for antigen processing and presentation. *Immunol Rev* 272: 28–38. doi:10.1111/immr.12427

## B. Article: Proteostasis in dendritic cells is controlled by the PERK signaling axis independently of ATF4



- Arguello RJ, Reverendo M, Mendes A, Camosseto V, Torres AG, Ribas de Pouplana L, van de Pavert SA, Gatti E, Pierre P (2018) SunRISE: Measuring translation elongation at single-cell resolution by means of flow cytometry. *J Cell Sci* 131: 132–142. doi:10.1242/jcs.214346
- Axten JM, Romeril SP, Shu A, Ralph J, Medina JR, Feng Y, Li WH, Grant SW, Heerding DA, Minthorn E, et al (2013) Discovery of GSK2656157: An optimized PERK inhibitor selected for preclinical development. *ACS Med Chem Lett* 4: 964–968. doi:10.1021/ml400228e
- Bougeard S, Dray S (2018) Supervised multiblock analysis in R with the ade4 package. *J Stat Softw* 86: 1–17. doi:10.18637/jss.v086.i01
- Boyle ST, Poltavets V, Kular J, Pyne NT, Sandow JJ, Lewis AC, Murphy KJ, Kolesnikoff N, Moretti PAB, Tea MN, et al (2020) ROCK-mediated selective activation of PERK signalling causes fibroblast reprogramming and tumour progression through a CRELD2-dependent mechanism. *Nat Cell Biol* 22: 908–918. doi:10.1038/s41556-020-0539-3
- Brasel K, De Smedt T, Smith JL, Maliszewski CR (2000) Generation of murine dendritic cells from fIt3-ligand-supplemented bone marrow cultures. *Blood* 96: 3029–3039. doi:10.1182/blood.v96.9.3029.h8003029\_3029\_3039
- Bretou M, Saez PJ, Sanseau D, Maurin M, Lankar D, Chabaud M, Spanpanato C, Malbec O, Barbier L, Muallem S, et al (2017) Lysosome signaling controls the migration of dendritic cells. *Sci Immunol* 16: eaak9573. doi:10.1126/sciimmunol.aak9573
- Carvalho BS, Irizarry RA (2010) A framework for oligonucleotide microarray preprocessing. *Bioinformatics* 26: 2363–2367. doi:10.1093/bioinformatics/btq431
- Caton ML, Smith-Raska MR, Reizis B (2007) Notch-RBP-J signaling controls the homeostasis of CD8<sup>+</sup> dendritic cells in the spleen. *J Exp Med* 204: 1653–1664. doi:10.1084/jem.20062648
- Ceppi M, Pereira PM, Dunand-Sauthier I, Barras E, Reith W, Santos MA, Pierre P (2009) MicroRNA-155 modulates the interleukin-1 signaling pathway in activated human monocyte-derived dendritic cells. *Proc Natl Acad Sci U S A* 106: 2735–2740. doi:10.1073/pnas.0811073106
- Chabaud M, Heuze ML, Bretou M, Vargas P, Maiuri P, Solanes P, Maurin M, Terriac E, Le Berre M, Lankar D, et al (2015) Cell migration and antigen capture are antagonistic processes coupled by myosin II in dendritic cells. *Nat Commun* 6: 7526–7542. doi:10.1038/ncomms8526
- Chambers JE, Dalton LE, Clarke HJ, Malzer E, Dominicus CS, Patel V, Moorhead G, Ron D, Marciniak SJ (2015) Actin dynamics tune the integrated stress response by regulating eukaryotic initiation factor 2alpha dephosphorylation. *Elife* 4: e04872. doi:10.7554/elife.04872
- Chen R, Rato C, Yan Y, Crespillo-Casado A, Clarke HJ, Harding HP, Marciniak SJ, Read RJ, Ron D (2015) G-actin provides substrate-specificity to eukaryotic initiation factor 2alpha holophosphatases. *Elife* 4: e04871. doi:10.7554/elife.04871
- Chiritoiu M, Brouwers N, Turacchio G, Pirozzi M, Malhotra V (2019) GRASP55 and UPR control interleukin-1beta aggregation and secretion. *Dev Cell* 49: 145–155. doi:10.1016/j.devcel.2019.02.011
- Clark K, Peggie M, Plater L, Sorcek RJ, Young ER, Madwed JB, Hough J, McIver EG, Cohen P (2011) Novel cross-talk within the IKK family controls innate immunity. *Biochem J* 434: 93–104. doi:10.1042/bj20101701
- Claudio N, Dalet A, Gatti E, Pierre P (2013) Mapping the crossroads of immune activation and cellular stress response pathways. *EMBO J* 32: 1214–1224. doi:10.1038/emboj.2013.80
- Clavarino G, Adriouach S, Quesada JL, Clay M, Chevreau M, Trocme C, Grange L, Gaudin P, Gatti E, Pierre P, et al (2016) Unfolded protein response gene GADD34 is overexpressed in rheumatoid arthritis and related to the presence of circulating anti-citrullinated protein antibodies. *Autoimmunity* 49: 172–178. doi:10.3109/08916934.2016.1138220
- Clavarino G, Claudio N, Couderc T, Dalet A, Judith D, Camosseto V, Schmidt EK, Wenger T, Lecuit M, Gatti E, et al (2012a) Induction of GADD34 is necessary for dsRNA-dependent interferon-beta production and participates in the control of Chikungunya virus infection. *PLoS Pathog* 8: e1002708. doi:10.1371/journal.ppat.1002708
- Clavarino G, Claudio N, Dalet A, Terawaki S, Couderc T, Chasson L, Ceppi M, Schmidt EK, Wenger T, Lecuit M, et al (2012b) Protein phosphatase 1 subunit Ppp1r15a/GADD34 regulates cytokine production in polyinosinic:polycytidylic acid-stimulated dendritic cells. *Proc Natl Acad Sci U S A* 109: 3006–3011. doi:10.1073/pnas.1104491109
- Costa-Mattioli M, Walter P (2020) The integrated stress response: From mechanism to disease. *Science* 368: eaat5314. doi:10.1126/science.aat5314
- Crespillo-Casado A, Chambers JE, Fischer PM, Marciniak SJ, Ron D (2017) PPP1R15A-mediated dephosphorylation of eIF2alpha is unaffected by Sephin1 or Guanabenz. *Elife* 6: e26109. doi:10.7554/elife.26109
- Crespillo-Casado A, Claes Z, Choy MS, Peti W, Bollen M, Ron D (2018) A Sephin1-insensitive tripartite holophosphatase dephosphorylates translation initiation factor 2alpha. *J Biol Chem* 293: 7766–7776. doi:10.1074/jbc.ra118.002325
- Dalet A, Arguello RJ, Combes A, Spinelli L, Jaeger S, Fallet M, Vu Manh TP, Mendes A, Perego J, Reverendo M, et al (2017) Protein synthesis inhibition and GADD34 control IFN-beta heterogeneous expression in response to dsRNA. *EMBO J* 15: 761–782. doi:10.15252/emboj.201695000
- Dalod M, Chelbi R, Malissen B, Lawrence T (2014) Dendritic cell maturation: Functional specialization through signaling specificity and transcriptional programming. *EMBO J* 33: 1104–1116. doi:10.1002/emboj.201488027
- Fitzgerald KA, Kagan JC (2020) Toll-like receptors and the control of immunity. *Cell* 180: 1044–1066. doi:10.1016/j.cell.2020.02.041
- Fontenot JD, Rasmussen JP, Williams LM, Dooley JL, Farr AG, Rudensky AY (2005) Regulatory T cell lineage specification by the forkhead transcription factor foxp3. *Immunity* 22: 329–341. doi:10.1016/j.immuni.2005.01.016
- Freaney JE, Kim R, Mandhana R, Horvath CM (2013) Extensive cooperation of immune master regulators IRF3 and NFkappaB in RNA Pol II recruitment and pause release in human innate antiviral transcription. *Cell Rep* 4: 959–973. doi:10.1016/j.celrep.2013.07.043
- Fusakio ME, Willy JA, Wang Y, Mirek ET, Al Baghdadi RJ, Adams CM, Anthony TG, Wek RC (2016) Transcription factor ATF4 directs basal and stress-induced gene expression in the unfolded protein response and cholesterol metabolism in the liver. *Mol Biol Cell* 27: 1536–1551. doi:10.1091/mbc.e16-01-0039
- Gandin V, Masvidal L, Cargnello M, Gyenis L, McLaughlan S, Cai Y, Tenkerian C, Morita M, Balanathan P, Jean-Jean O, et al (2016) mTORC1 and CK2 coordinate ternary and eIF4F complex assembly. *Nat Commun* 7: 11127. doi:10.1038/ncomms11127
- Guan BJ, van Hoef V, Jobava R, Elroy-Stein O, Valasek LS, Cargnello M, Gao XH, Krokowski D, Merrick WC, Kimball SR, et al (2017) A unique ISR program determines cellular responses to chronic stress. *Mol Cell* 68: 885–900. doi:10.1016/j.molcel.2017.11.007
- Han J, Back SH, Hur J, Lin YH, Gildersleeve R, Shan J, Yuan CL, Krokowski D, Wang S, Hatzoglou M, et al (2013) ER-stress-induced transcriptional regulation increases protein synthesis leading to cell death. *Nat Cell Biol* 15: 481–490. doi:10.1038/ncb2738
- Harding HP, Zhang Y, Scheuner D, Chen JJ, Kaufman RJ, Ron D (2009) Ppp1r15 gene knockout reveals an essential role for translation initiation factor 2 alpha (eIF2alpha) dephosphorylation in mammalian development. *Proc Natl Acad Sci U S A* 106: 1832–1837. doi:10.1073/pnas.0809632106
- Heuze ML, Vargas P, Chabaud M, Le Berre M, Liu YJ, Collin O, Solanes P, Voituriez R, Piel M, Lennon-Dumenil AM (2013) Migration of dendritic cells: Physical principles, molecular mechanisms, and functional implications. *Immunol Rev* 256: 240–254. doi:10.1111/imr.12108
- Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003) Exploration, normalization, and summaries of high

## B. Article: Proteostasis in dendritic cells is controlled by the PERK signaling axis independently of ATF4



- density oligonucleotide array probe level data. *Biostatistics* 4: 249–264. doi:10.1093/biostatistics/4.2.249
- Ito S, Tanaka Y, Oshino R, Aiba K, Thanasegaran S, Nishio N, Isobe K (2015) GADD34 inhibits activation-induced apoptosis of macrophages through enhancement of autophagy. *Sci Rep* 5: 8327. doi:10.1038/srep08327
- Iwasaki S, Iwasaki W, Takahashi M, Sakamoto A, Watanabe C, Shichino Y, Floor SN, Fujiwara K, Mito M, Dodo K, et al (2019) The translation inhibitor Rocaglamide targets a bimolecular cavity between eIF4A and polypurine RNA. *Mol Cell* 73: 738–748. doi:10.1016/j.molcel.2018.11.026
- Kim JH, Park SM, Park JH, Keum SJ, Jang SK (2011) eIF2A mediates translation of hepatitis C viral mRNA under stress conditions. *EMBO J* 30: 2454–2464. doi:10.1038/emboj.2011.146
- Krishna KH, Kumar MS (2018) Molecular evolution and functional divergence of eukaryotic translation initiation factor 2- $\alpha$  kinases. *PLoS One* 13: e0194335. doi:10.1371/journal.pone.0194335
- Kumar A, Yang YL, Flati V, Der S, Kadereit S, Deb A, Haque J, Reis L, Weissmann C, Williams BR (1997) Deficient cytokine signaling in mouse embryo fibroblasts with a targeted deletion in the PKR gene: Role of IRF-1 and NF- $\kappa$ B. *EMBO J* 16: 406–416. doi:10.1093/emboj/16.2.406
- Lazarus MB, Levin RS, Shokat KM (2017) Discovery of new substrates of the elongation factor-2 kinase suggests a broader role in the cellular nutrient response. *Cell Signal* 29: 78–83. doi:10.1016/j.cellsig.2016.10.006
- Lazear HM, Lancaster A, Wilkins C, Suthar MS, Huang A, Vick SC, Clepper L, Thackray L, Brassil MM, Virgin HW, et al (2013) IRF-3, IRF-5, and IRF-7 coordinately regulate the type I IFN response in myeloid dendritic cells downstream of MAVS signaling. *PLoS Pathog* 9: e1003118. doi:10.1371/journal.ppat.1003118
- Lelouard H, Schmidt EK, Camosseto V, Clavarino G, Ceppi M, Hsu HT, Pierre P (2007) Regulation of translation is required for dendritic cell function and survival during activation. *J Cell Biol* 179: 1427–1439. doi:10.1083/jcb.200707166
- Liang Q, Deng H, Sun CW, Townes TM, Zhu F (2011) Negative regulation of IRF7 activation by activating transcription factor 4 suggests a cross-regulation between the IFN responses and the cellular integrated stress responses. *J Immunol* 186: 1001–1010. doi:10.4049/jimmunol.1002240
- Marciniak SJ, Yun CY, Ouyadomari S, Novoa I, Zhang Y, Jungreis R, Nagata K, Harding HP, Ron D (2004) CHOP induces death by promoting protein synthesis and oxidation in the stressed endoplasmic reticulum. *Genes Dev* 18: 3066–3077. doi:10.1101/gad.1250704
- Mellman I (2013) Dendritic cells: Master regulators of the immune response. *Cancer Immunol Res* 1: 145–149. doi:10.1158/2326-6066.cir-13-0102
- Moerke NJ, Aktas H, Chen H, Cantel S, Reibarkh MY, Fahmy A, Gross JD, Degtarev A, Yuan J, Chorev M, et al (2007) Small-molecule inhibition of the interaction between the translation initiation factors eIF4E and eIF4G. *Cell* 128: 257–267. doi:10.1016/j.cell.2006.11.046
- Munn DH, Sharma MD, Hou D, Baban B, Lee JR, Antonia SJ, Messina JL, Chandler P, Koni PA, Mellor AL (2004) Expression of indoleamine 2,3-dioxygenase by plasmacytoid dendritic cells in tumor-draining lymph nodes. *J Clin Invest* 114: 280–290. doi:10.1172/jci21583
- Novoa I, Zeng H, Harding HP, Ron D (2001) Feedback inhibition of the unfolded protein response by GADD34-mediated dephosphorylation of eIF2 $\alpha$ . *J Cell Biol* 153: 1011–1022. doi:10.1083/jcb.153.5.1011
- Novoa I, Zhang Y, Zeng H, Jungreis R, Harding HP, Ron D (2003) Stress-induced gene expression requires programmed recovery from translational repression. *EMBO J* 22: 1180–1187. doi:10.1093/emboj/cdg112
- Palam LR, Baird TD, Wek RC (2011) Phosphorylation of eIF2 facilitates ribosomal bypass of an inhibitory upstream ORF to enhance CHOP translation. *J Biol Chem* 286: 10939–10949. doi:10.1074/jbc.M110.216093
- Pasini S, Liu J, Corona C, Peze-Heidsieck E, Shelanski M, Greene LA (2016) Activating transcription factor 4 (ATF4) modulates Rho GTPase levels and function via regulation of RhoGD $\alpha$ . *Sci Rep* 6: 36952. doi:10.1038/srep36952
- Paton AW, Beddoe T, Thorpe CM, Whisstock JC, Wilce MC, Rossjohn J, Talbot UM, Paton JC (2006) AB5 subtilase cytotoxin inactivates the endoplasmic reticulum chaperone BiP. *Nature* 443: 548–552. doi:10.1038/nature05124
- Paton AW, Srimanote P, Talbot UM, Wang H, Paton JC (2004) A new family of potent AB(5) cytotoxins produced by Shiga toxin-producing *Escherichia coli*. *J Exp Med* 200: 35–46. doi:10.1084/jem.20040392
- Perego J, Mendes A, Bourbon C, Camosseto V, Combes A, Liu H, Manh TV, Dalet A, Chasson L, Spinelli L, et al (2018) Guanabenz inhibits TLR9 signaling through a pathway that is independent of eIF2 $\alpha$  dephosphorylation by the GADD34/PP1c complex. *Sci Signal* 11: eaam8104. doi:10.1126/scisignal.aam8104
- Pierre P (2019) Integrating stress responses and immunity. *Science* 365: 28–29. doi:10.1126/science.aay0987
- Pierre P, Turley SJ, Gatti E, Hull M, Meltzer J, Mirza A, Inaba K, Steinman RM, Mellman I (1997) Developmental regulation of MHC class II transport in mouse dendritic cells. *Nature* 388: 787–792. doi:10.1038/42039
- Prete F, Catucci M, Labrada M, Gobessi S, Castiello MC, Bonomi E, Aiuti A, Vermi W, Cancrini C, Metin A, et al (2013) Wiskott-Aldrich syndrome protein-mediated actin dynamics control type-I interferon production in plasmacytoid dendritic cells. *J Exp Med* 210: 355–374. doi:10.1084/jem.20120363
- Pulecio J, Petrovic J, Prete F, Chiaruttini G, Lennon-Dumenil AM, Desdouets C, Gasman S, Burrone OR, Benvenuti F (2010) Cdc42-mediated MTOC polarization in dendritic cells controls targeted delivery of cytokines at the immune synapse. *J Exp Med* 207: 2719–2732. doi:10.1084/jem.20100007
- Reid DW, Tay AS, Sundaram JR, Lee IC, Chen Q, George SE, Nicchitta CV, Shenolikar S (2016) Complementary roles of GADD34- and CREP-containing eukaryotic initiation factor 2 $\alpha$  phosphatases during the unfolded protein response. *Mol Cell Biol* 36: 1868–1880. doi:10.1128/mcb.00190-16
- Reverendo M, Arguello RJ, Polte C, Valecka J, Camosseto V, Auphan-Anezin N, Ignatova Z, Gatti E, Pierre P (2019) Polymerase III transcription is necessary for T cell priming by dendritic cells. *Proc Natl Acad Sci U S A* 116: 22721–22729. doi:10.1073/pnas.1904396116
- Reverendo M, Mendes A, Arguello RJ, Gatti E, Pierre P (2018) At the crossway of ER-stress and proinflammatory responses. *FEBS J* 286: 297–310. doi:10.1111/febs.14391
- Robbins SH, Walzer T, Dembele D, Thibault C, Defays A, Bessou G, Xu H, Vivier E, Sellars M, Pierre P, et al (2008) Novel insights into the relationships between dendritic cell subsets in human and mouse revealed by genome-wide expression profiling. *Genome Biol* 9: R17. doi:10.1186/gb-2008-9-1-r17
- Ryazanov AG (2002) Elongation factor-2 kinase and its newly discovered relatives. *FEBS Lett* 514: 26–29. doi:10.1016/s0014-5793(02)02299-8
- Sanchez CL, Sims SG, Nowery JD, Meares GP (2019) Endoplasmic reticulum stress differentially modulates the IL-6 family of cytokines in murine astrocytes and macrophages. *Sci Rep* 9: 14931. doi:10.1038/s41598-019-51481-6
- Schmidt EK, Clavarino G, Ceppi M, Pierre P (2009) SUnSET, a nonradioactive method to monitor protein synthesis. *Nat Methods* 6: 275–287. doi:10.1038/nmeth.1314
- Sidrauski C, McGeachy AM, Ingolia NT, Walter P (2015) The small molecule ISRIB reverses the effects of eIF2 $\alpha$  phosphorylation on translation and stress granule assembly. *Elife* 4: 05033. doi:10.7554/elife.05033
- Spinelli L, Carpentier S, Montanana Sanchis F, Dalod M, Vu Manh TP (2015) BubbleGUM: Automatic extraction of phenotype molecular signatures

## B. Article: Proteostasis in dendritic cells is controlled by the PERK signaling axis independently of ATF4



- and comprehensive visualization of multiple gene set enrichment analyses. *BMC Genomics* 16: 814. doi:[10.1186/s12864-015-2012-4](https://doi.org/10.1186/s12864-015-2012-4)
- Starck SR, Tsai JC, Chen K, Shodiya M, Wang L, Yahiro K, Martins-Green M, Shastri N, Walter P (2016) Translation from the 5' untranslated region shapes the integrated stress response. *Science* 351: aad3867. doi:[10.1126/science.aad3867](https://doi.org/10.1126/science.aad3867)
- Steinman RM (2007) Dendritic cells: Understanding immunogenicity. *Eur J Immunol* 37: S53–S60. doi:[10.1002/eji.200737400](https://doi.org/10.1002/eji.200737400)
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102: 15545–15550. doi:[10.1073/pnas.0506580102](https://doi.org/10.1073/pnas.0506580102)
- Tavernier SJ, Osorio F, Vandersarren L, Vettters J, Vanlangenakker N, Van Isterdael G, Vergote K, De Rycke R, Parthoens E, van de Laar L, et al (2017) Regulated IRE1-dependent mRNA decay sets the threshold for dendritic cell survival. *Nat Cell Biol* 19: 698–710. doi:[10.1038/ncb3518](https://doi.org/10.1038/ncb3518)
- Terawaki S, Camosseto V, Prete F, Wenger T, Papadopoulos A, Rondeau C, Combes A, Rodriguez Rodrigues C, Vu Manh TP, et al (2015) RUN and FYVE domain-containing protein 4 enhances autophagy and lysosome tethering in response to Interleukin-4. *J Cell Biol* 210: 1133–1152. doi:[10.1083/jcb.201501059](https://doi.org/10.1083/jcb.201501059)
- van Vliet AR, Giordano F, Gerlo S, Segura I, Van Eygen S, Molenberghs G, Rocha S, Houcine A, Derua R, Verfaillie T, et al (2017) The ER stress sensor PERK coordinates ER-plasma membrane contact site formation through interaction with filamin-A and F-actin remodeling. *Mol Cell* 65: 885–899.e6. doi:[10.1016/j.molcel.2017.01.020](https://doi.org/10.1016/j.molcel.2017.01.020)
- Vargas P, Chabaud M, Thiam HR, Lankar D, Piel M, Lennon-Dumenil AM (2016) Study of dendritic cell migration using micro-fabrication. *J Immunol Methods* 432: 30–44. doi:[10.1016/j.jim.2015.12.005](https://doi.org/10.1016/j.jim.2015.12.005)
- Walter P, Ron D (2011) The unfolded protein response: From stress pathway to homeostatic regulation. *Science* 334: 1081–1086. doi:[10.1126/science.1209038](https://doi.org/10.1126/science.1209038)
- Wang C, Yu X, Cao Q, Wang Y, Zheng G, Tan TK, Zhao H, Zhao Y, Wang Y, Harris D (2013) Characterization of murine macrophages from bone marrow, spleen and peritoneum. *BMC Immunol* 14: 6. doi:[10.1186/1471-2172-14-6](https://doi.org/10.1186/1471-2172-14-6)
- West MA, Wallin RP, Matthews SP, Svensson HG, Zaru R, Ljunggren HG, Prescott AR, Watts C (2004) Enhanced dendritic cell antigen capture via toll-like receptor-induced actin remodeling. *Science* 305: 1153–1157. doi:[10.1126/science.1099153](https://doi.org/10.1126/science.1099153)
- Woo CW, Cui D, Arellano J, Dorweiler B, Harding H, Fitzgerald KA, Ron D, Tabas I (2009) Adaptive suppression of the ATF4-CHOP branch of the unfolded protein response by toll-like receptor signalling. *Nat Cell Biol* 11: 1473–1480. doi:[10.1038/ncb1996](https://doi.org/10.1038/ncb1996)
- Woo CW, Kutzler L, Kimball SR, Tabas I (2012) Toll-like receptor activation suppresses ER stress factor CHOP and translation inhibition through activation of eIF2B. *Nat Cell Biol* 14: 192–200. doi:[10.1038/ncb2408](https://doi.org/10.1038/ncb2408)
- Yamasaki S, Anderson P (2008) Reprogramming mRNA translation during stress. *Curr Opin Cell Biol* 20: 222–236. doi:[10.1016/j.ceb.2008.01.013](https://doi.org/10.1016/j.ceb.2008.01.013)
- Zhang W, Feng D, Li Y, Iida K, McGrath B, Cavener DR (2006) PERK EIF2AK3 control of pancreatic beta cell differentiation and proliferation is required for postnatal glucose homeostasis. *Cell Metab* 4: 491–507. doi:[10.1016/j.cmet.2006.11.002](https://doi.org/10.1016/j.cmet.2006.11.002)



**License:** This article is available under a Creative Commons License (Attribution 4.0 International, as described at <https://creativecommons.org/licenses/by/4.0/>).

# Bibliography

1. Abeyruwan, S., Vempati, U. D., Küçük-McGinty, H., *et al.* Evolving BioAssay Ontology (BAO): modularization, integration and applications. *Journal of Biomedical Semantics* **5**, S5. ISSN: 2041-1480 (Suppl 1 Proceedings of the Bio-Ontologies Spec Interest G 2014).
2. Adai, A. T., Date, S. V., Wieland, S., *et al.* LGL: Creating a Map of Protein Function with an Algorithm for Visualizing Very Large Biological Networks. *Journal of Molecular Biology* **340**, 179–190. ISSN: 00222836. <https://linkinghub.elsevier.com/retrieve/pii/S0022283604004851> (2021) (June 2004).
3. An, G., Mi, Q., Dutta-Moscato, J., *et al.* Agent-based models in translational systems biology. *WIREs Systems Biology and Medicine* **1**, 159–171. ISSN: 1939-5094, 1939-005X. <https://onlinelibrary.wiley.com/doi/10.1002/wsbm.45> (2022) (Sept. 2009).
4. Anderson, D. M., Makarewich, C. A., Anderson, K. M., *et al.* Widespread control of calcium signaling by a family of SERCA-inhibiting micropeptides. *Science Signaling* **9**, ra119–ra119. ISSN: 1945-0877, 1937-9145. <https://stke.sciencemag.org/lookup/doi/10.1126/scisignal.aaj1460> (2021) (Dec. 6, 2016).
5. Andreev, D. E., Baranov, P. V., Milogrodskii, A., *et al.* A deterministic model for non-monotone relationship between translation of upstream and downstream open reading frames. *arXiv:2012.15269 [math]*. arXiv: 2012.15269. <http://arxiv.org/abs/2012.15269> (2021) (Dec. 30, 2020).
6. Andreev, D. E., Arnold, M., Kiniry, S. J., *et al.* TASEP modelling provides a parsimonious explanation for the ability of a single uORF to derepress translation during the integrated stress response. *eLife* **7**. ISSN: 2050-084X. <https://elifesciences.org/articles/32563> (June 2018).
7. Andreev, D. E., O'Connor, P. B., Fahey, C., *et al.* Translation of 5' leaders is pervasive in genes resistant to eIF2 repression. *eLife* **4**. ISSN: 2050-084X. <https://elifesciences.org/articles/03971> (Jan. 2015).
8. Andrews, S. J. & Rothnagel, J. A. Emerging evidence for functional peptides encoded by short open reading frames. *Nature Reviews Genetics* **15**, 193–204. ISSN: 1471-0056, 1471-0064. <http://www.nature.com/articles/nrg3520> (Mar. 2014).
9. Ashburner, M., Ball, C. A., Blake, J. A., *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25–29. ISSN: 1061-4036, 1546-1718. [http://www.nature.com/articles/ng0500\\_25](http://www.nature.com/articles/ng0500_25) (2022) (May 2000).

10. Aspden, J. L., Eyre-Walker, Y. C., Phillips, R. J., *et al.* Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. *eLife* **3**. ISSN: 2050-084X. <https://elifesciences.org/articles/03528> (Aug. 2014).
11. Auparakkitanon, S. & Wilairat, P. Universal scanning-free initiation of eukaryote protein translation—a new normal. *Biomolecular Concepts* **12**, 129–131. ISSN: 1868-503X, 1868-5021. <https://www.degruyter.com/document/doi/10.1515/bmc-2021-0014/html> (2021) (Jan. 1, 2021).
12. Badger, J. H. & Olsen, G. J. CRITICA: coding region identification tool invoking comparative analysis. *Molecular Biology and Evolution* **16**, 512–524. ISSN: 0737-4038 (Apr. 1999).
13. Bandrowski, A., Brinkman, R., Brochhausen, M., *et al.* The Ontology for Biomedical Investigations. *PloS One* **11**, e0154556. ISSN: 1932-6203 (2016).
14. Bartholomäus, A., Del Campo, C. & Ignatova, Z. Mapping the non-standardized biases of ribosome profiling. *Biological Chemistry* **397**, 23–35. ISSN: 1437-4315 (Jan. 2016).
15. Bauer, A. L., Beauchemin, C. A. & Perelson, A. S. Agent-based modeling of host–pathogen systems: The successes and challenges. *Information Sciences* **179**, 1379–1389. ISSN: 00200255. <https://linkinghub.elsevier.com/retrieve/pii/S0020025508004726> (2021) (Apr. 29, 2009).
16. Bazzini, A. A., Johnstone, T. G., Christiano, R., *et al.* Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *The EMBO Journal* **33**, 981–993. ISSN: 0261-4189, 1460-2075. <http://emboj.embopress.org/cgi/doi/10.1002/embj.201488411> (May 2014).
17. Beaudoin, C. A., Bartas, M., Volná, A., *et al.* Are There Hidden Genes in DNA/RNA Vaccines? *Frontiers in Immunology* **13**, 801915. ISSN: 1664-3224. <https://www.frontiersin.org/articles/10.3389/fimmu.2022.801915/full> (2022) (Feb. 8, 2022).
18. Becker, E., Robisson, B., Chapple, C. E., *et al.* Multifunctional proteins revealed by overlapping clustering in protein interaction network. *Bioinformatics* **28**, 84–90. ISSN: 1460-2059, 1367-4803. <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr621> (2022) (Jan. 1, 2012).
19. Berger, C. & Mahdavi, A. Review of current trends in agent-based modeling of building occupants for energy and indoor-environmental performance analysis. *Building and Environment* **173**, 106726. ISSN: 03601323. <https://linkinghub.elsevier.com/retrieve/pii/S0360132320300846> (2022) (Apr. 2020).
20. Blum, M., Chang, H.-Y., Chuguransky, S., *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research* **49**, D344–D354. ISSN: 1362-4962 (D1 Jan. 8, 2021).

21. Brun, C., Chevenet, F., Martin, D., *et al.* Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biology* **5**, R6. ISSN: 1474-760X (2003).
22. Brun, C., Herrmann, C. & Guénoche, A. Clustering proteins from interaction networks for the prediction of cellular functions. *BMC bioinformatics* **5**, 95. ISSN: 1471-2105 (July 13, 2004).
23. Brunet, M. A., Brunelle, M., Lucier, J.-F., *et al.* OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucleic Acids Research*. ISSN: 0305-1048, 1362-4962. <https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gky936/5123790> (2022) (Oct. 9, 2018).
24. Brunet, M. A., Lucier, J.-F., Levesque, M., *et al.* OpenProt 2021: deeper functional annotation of the coding potential of eukaryotic genomes. *Nucleic Acids Research* **49**, D380–D388. ISSN: 0305-1048, 1362-4962. <https://academic.oup.com/nar/article/49/D1/D380/5976898> (2022) (D1 Jan. 8, 2021).
25. Cabrera-Quio, L. E., Herberg, S. & Pauli, A. Decoding sORF translation – from small proteins to gene regulation. *RNA Biology* **13**, 1051–1059. ISSN: 1547-6286, 1555-8584. <https://www.tandfonline.com/doi/full/10.1080/15476286.2016.1218589> (Nov. 2016).
26. Caillou, P., Gaudou, B., Grignard, A., *et al.* in *Advances in Social Simulation 2015* (eds Jager, W., Verbrugge, R., Flache, A., *et al.*) Series Title: Advances in Intelligent Systems and Computing, 15–28 (Springer International Publishing, Cham, 2017). ISBN: 978-3-319-47252-2. [http://link.springer.com/10.1007/978-3-319-47253-9\\_2](http://link.springer.com/10.1007/978-3-319-47253-9_2) (2022).
27. Calviello, L., Mukherjee, N., Wyler, E., *et al.* Detecting actively translated open reading frames in ribosome profiling data. *Nature Methods* **13**, 165–170. ISSN: 1548-7091, 1548-7105. <http://www.nature.com/articles/nmeth.3688> (2020) (Feb. 2016).
28. Calvo, S. E., Pagliarini, D. J. & Mootha, V. K. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proceedings of the National Academy of Sciences* **106**, 7507–7512. ISSN: 0027-8424, 1091-6490. <https://pnas.org/doi/full/10.1073/pnas.0810916106> (2022) (May 5, 2009).
29. Chang, A., Schomburg, I., Placzek, S., *et al.* BRENDA in 2015: exciting developments in its 25th year of existence. *Nucleic Acids Research* **43**, D439–446. ISSN: 1362-4962 (Database issue Jan. 2015).
30. Chapple, C. E. & Brun, C. Redefining protein moonlighting. *Oncotarget* **6**, 16812–16813. ISSN: 1949-2553. <https://www.oncotarget.com/lookup/doi/10.18632/oncotarget.4793> (2021) (July 10, 2015).



31. Chassé, H., Boulben, S., Costache, V., *et al.* Analysis of translation using polysome profiling. *Nucleic Acids Research*, gkw907. ISSN: 0305-1048, 1362-4962. <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw907> (2021) (Oct. 7, 2016).
32. Chew, G.-L., Pauli, A. & Schier, A. F. Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish. *Nature Communications* **7**, 11663. ISSN: 2041-1723. <http://www.nature.com/doi/10.1038/ncomms11663> (May 2016).
33. Choteau, S. A., Wagner, A., Pierre, P., *et al.* MetamORF: a repository of unique short open reading frames identified by both experimental and computational approaches for gene and metagene analyses. *Database* **2021**, baab032. ISSN: 1758-0463. <https://academic.oup.com/database/article/doi/10.1093/database/baab032/6307706> (2022) (June 22, 2021).
34. Chun, S. Y., Rodriguez, C. M., Todd, P. K., *et al.* SPECTre: a spectral coherence--based classifier of actively translated transcripts from ribosome profiling sequence data. *BMC Bioinformatics* **17**, 482. ISSN: 1471-2105. <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1355-4> (2022) (Dec. 2016).
35. Cláudio, N., Dalet, A., Gatti, E., *et al.* Mapping the crossroads of immune activation and cellular stress response pathways. *The EMBO Journal* **32**, 1214–1224. ISSN: 0261-4189, 1460-2075. <http://emboj.embopress.org/cgi/doi/10.1038/emboj.2013.80> (2021) (Apr. 12, 2013).
36. Couso, J.-P. & Patraquim, P. Classification and function of small open reading frames. *Nature Reviews Molecular Cell Biology* **18**, 575–589. ISSN: 1471-0072, 1471-0080. <http://www.nature.com/doi/10.1038/nrm.2017.58> (July 2017).
37. Crappé, J., Ndah, E., Koch, A., *et al.* PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Research* **43**, e29–e29. ISSN: 1362-4962, 0305-1048. <http://academic.oup.com/nar/article/43/5/e29/2453155/PROTEOFORMER-deep-proteome-coverage-through> (2022) (Mar. 11, 2015).
38. Crowe, M. L., Wang, X.-Q. & Rothnagel, J. A. Evidence for conservation and selection of upstream open reading frames suggests probable encoding of bioactive peptides. *BMC Genomics* **7**, 16. ISSN: 1471-2164. <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2164-7-16> (2022) (Dec. 2006).
39. Davey, N. E., Van Roey, K., Weatheritt, R. J., *et al.* Attributes of short linear motifs. *Mol. BioSyst.* **8**, 268–281. ISSN: 1742-206X, 1742-2051. <http://xlink.rsc.org/?DOI=C1MB05231D> (2021) (2012).



40. Diehl, A. D., Meehan, T. F., Bradford, Y. M., *et al.* The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *Journal of Biomedical Semantics* **7**, 44. ISSN: 2041-1480 (July 4, 2016).
41. Dosztányi, Z., Csizmok, V., Tompa, P., *et al.* IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics (Oxford, England)* **21**, 3433–3434. ISSN: 1367-4803 (Aug. 15, 2005).
42. Doudna, J. A. & Charpentier, E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science (New York, N.Y.)* **346**, 1258096. ISSN: 1095-9203 (Nov. 28, 2014).
43. Edwards, R. J., Paulsen, K., Aguilar Gomez, C. M., *et al.* in *Intrinsically Disordered Proteins* (eds Kragelund, B. B. & Skriver, K.) 37–72 (Springer US, New York, NY, 2020). ISBN: 978-1-07-160523-3. [https://link.springer.com/10.1007/978-1-0716-0524-0\\_3](https://link.springer.com/10.1007/978-1-0716-0524-0_3) (2022).
44. Erhard, F., Halenius, A., Zimmermann, C., *et al.* Improved Ribo-seq enables identification of cryptic translation events. *Nature Methods* **15**, 363–366. ISSN: 1548-7091, 1548-7105. <http://www.nature.com/doifinder/10.1038/nmeth.4631> (Mar. 2018).
45. Evans, V. C., Barker, G., Heesom, K. J., *et al.* De novo derivation of proteomes from transcriptomes for transcript and protein identification. *Nature Methods* **9**, 1207–1211. ISSN: 1548-7105 (Dec. 2012).
46. Fields, A. P., Rodriguez, E. H., Jovanovic, M., *et al.* A regression-based analysis of ribosome-profiling data reveals a conserved complexity to mammalian translation. *Molecular Cell* **60**, 816–827. ISSN: 10972765. <https://linkinghub.elsevier.com/retrieve/pii/S1097276515009053> (Dec. 2015).
47. Golbreich, C., Grosjean, J. & Darmoni, S. J. The Foundational Model of Anatomy in OWL 2 and its use. *Artificial Intelligence in Medicine* **57**, 119–132. ISSN: 1873-2860 (Feb. 2013).
48. Gray, T., Storz, G. & Papenfort, K. Small Proteins; Big Questions. *Journal of Bacteriology*. ISSN: 0021-9193, 1098-5530. <https://journals.asm.org/doi/10.1128/JB.00341-21> (2021) (July 26, 2021).
49. Hagai, T., Azia, A., Babu, M. M., *et al.* Use of Host-like Peptide Motifs in Viral Proteins Is a Prevalent Strategy in Host-Virus Interactions. *Cell Reports* **7**, 1729–1739. ISSN: 22111247. <https://linkinghub.elsevier.com/retrieve/pii/S2211124714003702> (2022) (June 2014).
50. Hanada, K., Akiyama, K., Sakurai, T., *et al.* sORF finder: a program package to identify small open reading frames with high coding potential. *Bioinformatics* **26**, 399–400. ISSN: 1460-2059, 1367-4803. <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp688> (2022) (Feb. 1, 2010).

51. Hao, Y., Zhang, L., Niu, Y., *et al.* SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. *Briefings in Bioinformatics*, bbx005. ISSN: 1467-5463, 1477-4054. <https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbx005> (Jan. 2017).
52. Hazarika, R. R., Sostaric, N., Sun, Y., *et al.* Large-scale docking predicts that sORF-encoded peptides may function through protein-peptide interactions in *Arabidopsis thaliana*. *PLoS ONE* **13** (ed Helmer-Citterich, M.) e0205179. ISSN: 1932-6203. <http://dx.plos.org/10.1371/journal.pone.0205179> (Oct. 2018).
53. Hernández, G., Osnaya, V. G. & Pérez-Martínez, X. Conservation and Variability of the AUG Initiation Codon Context in Eukaryotes. *Trends in Biochemical Sciences* **44**, 1009–1021. ISSN: 09680004. <https://linkinghub.elsevier.com/retrieve/pii/S096800041930146X> (Dec. 2019).
54. Hinnebusch, A. G., Ivanov, I. P. & Sonenberg, N. Translational control by 5'-untranslated regions of eukaryotic mRNAs. *Science* **352**, 1413–1416. ISSN: 0036-8075, 1095-9203. <https://www.science.org/doi/10.1126/science.aad9868> (2021) (June 17, 2016).
55. Huang, J., Gutierrez, F., Strachan, H. J., *et al.* OmniSearch: a semantic search system based on the Ontology for MicroRNA Target (OMIT) for microRNA-target gene interaction data. *Journal of Biomedical Semantics* **7**, 25. ISSN: 2041-1480 (2016).
56. Iacono, M., Mignone, F. & Pesole, G. uAUG and uORFs in human and rodent 5'-untranslated mRNAs. *Gene* **349**, 97–105. ISSN: 03781119. <https://linkinghub.elsevier.com/retrieve/pii/S0378111904007188> (2022) (Apr. 2005).
57. Ingolia, N. T. Ribosome footprint profiling of translation throughout the genome. *Cell* **165**, 22–33. ISSN: 00928674. <https://linkinghub.elsevier.com/retrieve/pii/S0092867416302161> (Mar. 2016).
58. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., *et al.* Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* **324**, 218–223. ISSN: 0036-8075, 1095-9203. <https://www.science.org/doi/10.1126/science.1168978> (2022) (Apr. 10, 2009).
59. Janssen, S., Sharpanskykh, A., Curran, R., *et al.* Using causal discovery to analyze emergence in agent-based models. *Simulation Modelling Practice and Theory* **96**, 101940. ISSN: 1569190X. <https://linkinghub.elsevier.com/retrieve/pii/S1569190X19300735> (2022) (Nov. 2019).
60. Ji, Z., Song, R., Regev, A., *et al.* Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife* **4**, e08890. ISSN: 2050-084X (Dec. 2015).

61. Johnstone, T. G., Bazzini, A. A. & Giraldez, A. J. Upstream ORFs are prevalent translational repressors in vertebrates. *The EMBO Journal* **35**, 706–723. ISSN: 0261-4189, 1460-2075. <http://emboj.embopress.org/lookup/doi/10.15252/emboj.201592759> (Apr. 2016).
62. Jones, P., Binns, D., Chang, H.-Y., *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240. ISSN: 1367-4803, 1460-2059. <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu031> (2022) (May 1, 2014).
63. Jørgensen, A. C. S., Ghosh, A., Sturrock, M., *et al.* *Efficient inference for agent-based models of real-world phenomena* preprint (Bioinformatics, Oct. 5, 2021). <http://biorxiv.org/lookup/doi/10.1101/2021.10.04.462980> (2021).
64. Juhas, U., Ryba-Stanisławowska, M., Szargiej, P., *et al.* Different pathways of macrophage activation and polarization. *Postępy Higieny i Medycyny Doświadczalnej* **69**, 496–502. ISSN: 1732-2693. <https://publisherspanel.com/ucid/1150133> (2022) (Apr. 22, 2015).
65. Kent, W. J., Sugnet, C. W., Furey, T. S., *et al.* The Human Genome Browser at UCSC. *Genome Research* **12**, 996–1006. ISSN: 1088-9051. <http://www.genome.org/cgi/doi/10.1101/gr.229102> (May 2002).
66. King, H. A. & Gerber, A. P. Translatome profiling: methods for genome-scale analysis of mRNA translation. *Briefings in Functional Genomics*, elu045. ISSN: 2041-2649, 2041-2657. <https://academic.oup.com/bfg/article-lookup/doi/10.1093/bfgp/elu045> (2022) (Nov. 6, 2014).
67. Kozak, M. Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes. *Proceedings of the National Academy of Sciences* **87**, 8301–8305. ISSN: 0027-8424, 1091-6490. <https://pnas.org/doi/full/10.1073/pnas.87.21.8301> (2022) (Nov. 1990).
68. Kozak, M. Effects of intercistronic length on the efficiency of reinitiation by eucaryotic ribosomes. *Molecular and Cellular Biology* **7**, 3438–3445. ISSN: 0270-7306 (Oct. 1987).
69. Kozak, M. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* **44**, 283–292. ISSN: 00928674. <https://linkinghub.elsevier.com/retrieve/pii/0092867486907622> (2022) (Jan. 1986).
70. Kumar, M., Gouw, M., Michael, S., *et al.* ELM—the eukaryotic linear motif resource in 2020. *Nucleic Acids Research*, gkz1030. ISSN: 0305-1048, 1362-4962. <https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkz1030/5611669> (2021) (Nov. 4, 2019).
71. Kumar, M., Michael, S., Alvarado-Valverde, J., *et al.* The Eukaryotic Linear Motif resource: 2022 release. *Nucleic Acids Research* **50**, D497–D508. ISSN: 0305-1048, 1362-4962. <https://academic.oup.com/nar/article/50/D1/D497/6414054> (2022) (D1 Jan. 7, 2022).

72. Laumont, C. M., Daouda, T., Laverdure, J.-P., *et al.* Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nature Communications* **7**, 10238. ISSN: 2041-1723. <http://www.nature.com/doi/10.1038/ncomms10238> (Jan. 2016).
73. Le, Q. G. & Kimata, Y. Multiple ways for stress sensing and regulation of the endoplasmic reticulum-stress sensors. *Cell Structure and Function*. ISSN: 0386-7196, 1347-3700. [https://www.jstage.jst.go.jp/article/csf/advpub/0/advpub\\_21015/\\_article](https://www.jstage.jst.go.jp/article/csf/advpub/0/advpub_21015/_article) (2021) (2021).
74. Lee, D. S. M., Park, J., Kromer, A., *et al.* Disrupting upstream translation in mRNAs is associated with human disease. *Nature Communications* **12**, 1515. ISSN: 2041-1723. <http://www.nature.com/articles/s41467-021-21812-1> (2021) (Dec. 2021).
75. Lee, S., Liu, B., Lee, S., *et al.* Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proceedings of the National Academy of Sciences of the United States of America* **109**, E2424–2432. ISSN: 1091-6490 (Sept. 2012).
76. Lee, W.-W. & Jeon, B. S. Clinical spectrum of dopa-responsive dystonia and related disorders. *Current Neurology and Neuroscience Reports* **14**, 461. ISSN: 1534-6293 (July 2014).
77. Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275–i282. ISSN: 1367-4803, 1460-2059. <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr209> (July 2011).
78. Liu, W., Xiang, L., Zheng, T., *et al.* TranslatomeDB: a comprehensive database and cloud-based analysis platform for translatome sequencing data. *Nucleic Acids Research* **46**, D206–D212. ISSN: 0305-1048, 1362-4962 (D1 Jan. 2018).
79. Lv, D., Chang, Z., Cai, Y., *et al.* TransLnc: a comprehensive resource for translatable lncRNAs extends immunopeptidome. *Nucleic Acids Research*, gkab847. ISSN: 0305-1048, 1362-4962. <https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkab847/6376020> (2021) (Sept. 27, 2021).
80. Ma, J., Diedrich, J. K., Jungreis, I., *et al.* Improved Identification and Analysis of Small Open Reading Frame Encoded Polypeptides. *Analytical Chemistry* **88**, 3967–3975. ISSN: 0003-2700, 1520-6882. <http://pubs.acs.org/doi/10.1021/acs.analchem.6b00191> (Apr. 2016).
81. Mackowiak, S. D., Zauber, H., Bielow, C., *et al.* Extensive identification and analysis of conserved small ORFs in animals. *Genome Biology* **16**. ISSN: 1474-760X. <http://genomebiology.com/2015/16/1/179> (Dec. 2015).
82. Makarewich, C. A. & Olson, E. N. Mining for Micropeptides. *Trends in Cell Biology* **27**, 685–696. ISSN: 09628924. <https://linkinghub.elsevier.com/retrieve/pii/S0962892417300648> (Sept. 2017).

83. Malone, J., Holloway, E., Adamusiak, T., *et al.* Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics (Oxford, England)* **26**, 1112–1118. ISSN: 1367-4811 (Apr. 15, 2010).
84. McGillivray, P., Ault, R., Pawashe, M., *et al.* A comprehensive catalog of predicted functional upstream open reading frames in humans. *Nucleic Acids Research* **46**, 3326–3338. ISSN: 0305-1048, 1362-4962. <https://academic.oup.com/nar/article/46/7/3326/4942470> (Apr. 2018).
85. Merkel & Dirk. Docker: lightweight linux containers for consistent development and deployment. *Linux journal* **2014**, 2 (2014).
86. Michel, A. M., Fox, G., M. Kiran, A., *et al.* GWIPS-viz: development of a ribo-seq genome browser. *Nucleic Acids Research* **42**, D859–D864. ISSN: 0305-1048, 1362-4962. <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkt1035> (D1 Jan. 2014).
87. Mölder, F., Jablonski, K. P., Letcher, B., *et al.* Sustainable data analysis with Snakemake. *F1000Research* **10**, 33. ISSN: 2046-1402. <https://f1000research.com/articles/10-33/v1> (2022) (Jan. 18, 2021).
88. Moro, S. G., Hermans, C., Ruiz-Orera, J., *et al.* Impact of uORFs in mediating regulation of translation in stress conditions. *BMC Molecular and Cell Biology* **22**, 29. ISSN: 2661-8850. <https://bmcmolcellbiol.biomedcentral.com/articles/10.1186/s12860-021-00363-9> (2021) (Dec. 2021).
89. Mosca, R., Céol, A., Stein, A., *et al.* 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Research* **42**, D374–D379. ISSN: 0305-1048, 1362-4962. <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkt887> (2022) (D1 Jan. 2014).
90. Mouilleron, H., Delcourt, V. & Roucou, X. Death of a dogma: eukaryotic mRNAs can code for more than one protein. *Nucleic Acids Research* **44**, 14–23. ISSN: 1362-4962 (Jan. 8, 2016).
91. Mudge, J. M., Ruiz-Orera, J., Prensner, J. R., *et al.* Standardized annotation of translated open reading frames. *Nature Biotechnology* **40**, 994–999. ISSN: 1087-0156, 1546-1696. <https://www.nature.com/articles/s41587-022-01369-0> (2022) (July 2022).
92. Mumtaz, M. A. S. & Couso, J. P. Ribosomal profiling adds new coding sequences to the proteome. *Biochemical Society Transactions* **43**, 1271–1276. ISSN: 0300-5127, 1470-8752. <http://biochemsoctrans.org/cgi/doi/10.1042/BST20150170> (Dec. 2015).
93. Neville, M. D. C., Kohze, R., Erady, C., *et al.* A platform for curated products from novel open reading frames prompts reinterpretation of disease variants. *Genome Research*. ISSN: 1549-5469 (Jan. 19, 2021).

94. Olexiouk, V., Crappé, J., Verbruggen, S., *et al.* sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Research* **44**, D324–D329. ISSN: 0305-1048, 1362-4962. <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1175> (D1 Jan. 2016).
95. Olexiouk, V., Van Criekinge, W. & Menschaert, G. An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Research* **46**, D497–D502. ISSN: 0305-1048, 1362-4962. <http://academic.oup.com/nar/article/46/D1/D497/4621340> (D1 Jan. 2018).
96. Olingy, C. E., Dinh, H. Q. & Hedrick, C. C. Monocyte heterogeneity and functions in cancer. *Journal of Leukocyte Biology* **106**, 309–322. ISSN: 1938-3673 (Aug. 2019).
97. Orr, M. W., Mao, Y., Storz, G., *et al.* Alternative ORFs and small ORFs: shedding light on the dark proteome. *Nucleic Acids Research*. ISSN: 1362-4962 (Aug. 2019).
98. Paiano, A., Margiotta, A., De Luca, M., *et al.* Yeast Two-Hybrid Assay to Identify Interacting Proteins. *Current Protocols in Protein Science* **95**, e70. ISSN: 19343655. <https://onlinelibrary.wiley.com/doi/10.1002/cpp.70> (2022) (Feb. 2019).
99. Pakos-Zebrucka, K., Koryga, I., Mnich, K., *et al.* The integrated stress response. *EMBO reports* **17**, 1374–1395. ISSN: 1469-221X, 1469-3178. <https://onlinelibrary.wiley.com/doi/10.15252/embr.201642195> (2021) (Oct. 2016).
100. Pavlov, M. Y., Ullman, G., Ignatova, Z., *et al.* Estimation of peptide elongation times from ribosome profiling spectra. *Nucleic Acids Research* **49**, 5124–5142. ISSN: 0305-1048, 1362-4962. <https://academic.oup.com/nar/article/49/9/5124/6246399> (2021) (May 21, 2021).
101. Petell, C. J., Randene, K., Pappas, M., *et al.* Mechanically transduced immunosorbent assay to measure protein-protein interactions. *eLife* **10**, e67525. ISSN: 2050-084X. <https://elifesciences.org/articles/67525> (2021) (Sept. 28, 2021).
102. Plaza, S., Menschaert, G. & Payre, F. In Search of Lost Small Peptides. *Annual Review of Cell and Developmental Biology* **33**, 391–416. ISSN: 1081-0706, 1530-8995. <http://www.annualreviews.org/doi/10.1146/annurev-cellbio-100616-060516> (Oct. 2017).
103. Pozzati, G., Kundrotas, P. & Elofsson, A. *Improved protein docking by predicted interface residues* preprint (Bioinformatics, Aug. 26, 2021). <http://biorxiv.org/lookup/doi/10.1101/2021.08.25.457642> (2021).
104. Prasad, V. & Greber, U. F. The endoplasmic reticulum unfolded protein response – homeostasis, cell death and evolution in virus infections. *FEMS Microbiology Reviews*, fuab016. ISSN: 0168-6445, 1574-6976. <https://academic.oup.com/femsre/advance-article/doi/10.1093/femsre/fuab016/6188392> (2021) (Mar. 25, 2021).

105. Prel, A., Dozier, C., Combier, J.-P., *et al.* Evidence That Regulation of Pri-miRNA/miRNA Expression Is Not a General Rule of miPEPs Function in Humans. *International Journal of Molecular Sciences* **22**, 3432. ISSN: 1422-0067. <https://www.mdpi.com/1422-0067/22/7/3432> (2021) (Mar. 26, 2021).
106. Pueyo, J. I., Magny, E. G. & Couso, J. P. New peptides under the s(ORF)ace of the genome. *Trends in Biochemical Sciences* **41**, 665–678. ISSN: 09680004. <https://linkinghub.elsevier.com/retrieve/pii/S0968000416300317> (Aug. 2016).
107. Raj, A., Wang, S. H., Shim, H., *et al.* Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *eLife* **5**, e13328. ISSN: 2050-084X. <https://elifesciences.org/articles/13328> (May 2016).
108. Ren, L.-H., Ding, Y.-S., Shen, Y.-Z., *et al.* Multi-agent-based bio-network for systems biology: protein–protein interaction network as an example. *Amino Acids* **35**, 565–572. ISSN: 0939-4451, 1438-2199. <http://link.springer.com/10.1007/s00726-008-0081-2> (2021) (Oct. 2008).
109. Renz, P. F., Valdivia-Francia, F. & Sandoel, A. Some like it translated: small ORFs in the 5'UTR. *Experimental Cell Research* **396**, 112229. ISSN: 00144827. <https://linkinghub.elsevier.com/retrieve/pii/S001448272030478X> (2021) (Nov. 2020).
110. Rodriguez, C. M., Chun, S. Y., Mills, R. E., *et al.* Translation of upstream open reading frames in a model of neuronal differentiation. *BMC genomics* **20**, 391. ISSN: 1471-2164 (May 2019).
111. Rossol, M., Heine, H., Meusch, U., *et al.* LPS-induced Cytokine Production in Human Monocytes and Macrophages. *Critical Reviews in Immunology* **31**, 379–446. ISSN: 2162-6472. <http://www.dl.begellhouse.com/journals/2ff21abf44b19838,47d4cd1f0e2c889b,5453e7eb10564e78.html> (2022) (2011).
112. Rundlet, E. J., Holm, M., Schacherl, M., *et al.* Structural basis of early translocation events on the ribosome. *Nature* **595**, 741–745. ISSN: 0028-0836, 1476-4687. <http://www.nature.com/articles/s41586-021-03713-x> (2021) (July 29, 2021).
113. Saghatelian, A. & Couso, J. P. Discovery and characterization of smORF-encoded bioactive polypeptides. *Nature Chemical Biology* **11**, 909–916. ISSN: 1552-4450, 1552-4469. <http://www.nature.com/articles/nchembio.1964> (Dec. 2015).
114. Saha, S., Chatzimichali, E. A., Matthews, D. A., *et al.* PITDB: a database of translated genomic elements. *Nucleic Acids Research* **46**, D1223–D1228. ISSN: 0305-1048, 1362-4962. <http://academic.oup.com/nar/article/46/D1/D1223/4372529> (2022) (D1 Jan. 4, 2018).

115. Sakharov, P. A., Smolin, E. A., Lyabin, D. N., *et al.* ATP-Independent Initiation during Cap-Independent Translation of m6A-Modified mRNA. *International Journal of Molecular Sciences* **22**, 3662. ISSN: 1422-0067. <https://www.mdpi.com/1422-0067/22/7/3662> (2021) (Apr. 1, 2021).
116. Sakiyama, K., Shimokawa-Chiba, N., Fujiwara, K., *et al.* Search for translation arrest peptides encoded upstream of genes for components of protein localization pathways. *Nucleic Acids Research* **49**, 1550–1566. ISSN: 0305-1048, 1362-4962. <https://academic.oup.com/nar/article/49/3/1550/6121473> (2021) (Feb. 22, 2021).
117. Samandi, S., Roy, A. V., Delcourt, V., *et al.* Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. *eLife* **6**. ISSN: 2050-084X. <https://elifesciences.org/articles/27860> (e27860 Oct. 2017).
118. Sarntivijai, S., Lin, Y., Xiang, Z., *et al.* CLO: The cell line ontology. *Journal of Biomedical Semantics* **5**, 37. ISSN: 2041-1480 (2014).
119. Sato, H. & Singer, R. H. *Cellular variability of nonsense-mediated mRNA decay* preprint (Cell Biology, Mar. 31, 2021). <http://biorxiv.org/lookup/doi/10.1101/2021.03.31.437867> (2021).
120. Scholz, A., Eggenhofer, F., Gelhausen, R., *et al.* uORF-Tools—Workflow for the determination of translation-regulatory upstream open reading frames. *PLOS ONE* **14** (ed Jan, E.) e0222459. ISSN: 1932-6203. <https://dx.plos.org/10.1371/journal.pone.0222459> (2021) (Sept. 12, 2019).
121. Schoof, M., Boone, M., Wang, L., *et al.* eIF2B conformation and assembly state regulate the integrated stress response. *eLife* **10**, e65703. ISSN: 2050-084X. <https://elifesciences.org/articles/65703> (2021) (Mar. 10, 2021).
122. Schott, J., Reitter, S., Lindner, D., *et al.* Nascent Ribo-Seq measures ribosomal loading time and reveals kinetic impact on ribosome density. *Nature Methods*. ISSN: 1548-7091, 1548-7105. <https://www.nature.com/articles/s41592-021-01250-z> (2021) (Sept. 3, 2021).
123. Sharipov, R. N., Yevshin, I. S., Kondrakhin, Y. V., *et al.* RiboSeqDB – a repository of selected human and mouse ribosome footprint and RNA-seq data. *Virtual Biology* **1**, 37. ISSN: 2306-8140 (Dec. 2014).
124. Siepel, A., Bejerano, G., Pedersen, J. S., *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* **15**, 1034–1050. ISSN: 1088-9051 (Aug. 2005).
125. Singh, J., Hanson, J., Paliwal, K., *et al.* RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nature Communications* **10**, 5407. ISSN: 2041-1723. <http://www.nature.com/articles/s41467-019-13395-9> (2021) (Dec. 2019).



126. Sioutos, N., de Coronado, S., Haber, M. W., *et al.* NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics* **40**, 30–43. ISSN: 1532-0480 (Feb. 2007).
127. Skarszewski, A., Stanton-Cook, M., Huber, T., *et al.* uPEPPERoni: An online tool for upstream open reading frame location and analysis of transcript conservation. *BMC Bioinformatics* **15**, 36. ISSN: 1471-2105. <http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-36> (2014).
128. Skinnider, M. A., Scott, N. E., Prudova, A., *et al.* An atlas of protein-protein interactions across mouse tissues. *Cell* **184**, 4073–4089.e17. ISSN: 00928674. <https://linkinghub.elsevier.com/retrieve/pii/S0092867421007042> (2021) (July 2021).
129. Somers, J., Pöyry, T. & Willis, A. E. A perspective on mammalian upstream open reading frame function. *The International Journal of Biochemistry & Cell Biology* **45**, 1690–1700. ISSN: 1878-5875 (Aug. 2013).
130. Song, B., Jiang, M. & Gao, L. RiboNT: A Noise-Tolerant Predictor of Open Reading Frames from Ribosome-Protected Footprints. *Life* **11**, 701. ISSN: 2075-1729. <https://www.mdpi.com/2075-1729/11/7/701> (2021) (July 16, 2021).
131. Spealman, P., Naik, A. W., May, G. E., *et al.* Conserved non-AUG uORFs revealed by a novel regression analysis of ribosome profiling data. *Genome Research* **28**, 214–222. ISSN: 1088-9051, 1549-5469. <http://genome.cshlp.org/lookup/doi/10.1101/gr.221507.117> (Feb. 2018).
132. Starck, S. R., Tsai, J. C., Chen, K., *et al.* Translation from the 5' untranslated region shapes the integrated stress response. *Science* **351**, aad3867. ISSN: 0036-8075, 1095-9203. <http://www.sciencemag.org/cgi/doi/10.1126/science.aad3867> (Jan. 2016).
133. Tharakan, R. & Sawa, A. Minireview: Novel Micropeptide Discovery by Proteomics and Deep Sequencing Methods. *Frontiers in Genetics* **12**, 651485. ISSN: 1664-8021. <https://www.frontiersin.org/articles/10.3389/fgene.2021.651485/full> (2021) (May 6, 2021).
134. The Gene Ontology Consortium, Carbon, S., Douglass, E., *et al.* The Gene Ontology resource: enriching a Gold mine. *Nucleic Acids Research* **49**, D325–D334. ISSN: 0305-1048, 1362-4962. <https://academic.oup.com/nar/article/49/D1/D325/6027811> (2022) (D1 Jan. 8, 2021).
135. Tian, T., Li, S., Lang, P., *et al.* Full-length ribosome density prediction by a multi-input and multi-output model. *PLOS Computational Biology* **17** (ed Roy, S.) e1008842. ISSN: 1553-7358. <https://dx.plos.org/10.1371/journal.pcbi.1008842> (2021) (Mar. 26, 2021).
136. Tjeldnes, H., Labun, K., Torres Cleuren, Y., *et al.* ORFik: a comprehensive R toolkit for the analysis of translation. *BMC Bioinformatics* **22**, 336. ISSN: 1471-2105. <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-021-04254-w> (2021) (Dec. 2021).

137. Um, T., Park, T., Shim, J. S., *et al.* Application of Upstream Open Reading Frames (uORFs) Editing for the Development of Stress-Tolerant Crops. *International Journal of Molecular Sciences* **22**, 3743. ISSN: 1422-0067. <https://www.mdpi.com/1422-0067/22/7/3743> (2021) (Apr. 3, 2021).
138. Van den Akker, G. G. H., Zacchini, F., Housmans, B. A. C., *et al.* Current Practice in Bicistronic IRES Reporter Use: A Systematic Review. *International Journal of Molecular Sciences* **22**, 5193. ISSN: 1422-0067. <https://www.mdpi.com/1422-0067/22/10/5193> (2021) (May 14, 2021).
139. Vanderperre, B., Lucier, J.-F., Bissonnette, C., *et al.* Direct detection of alternative Open Reading Frames translation products in Human significantly expands the proteome. *PLoS ONE* **8**, e70698. ISSN: 1932-6203. <http://dx.plos.org/10.1371/journal.pone.0070698> (Aug. 2013).
140. Vanderperre, B., Lucier, J.-F. & Roucou, X. HAltORF: a database of predicted out-of-frame alternative open reading frames in human. *Database* **2012**, bas025. ISSN: 1758-0463. <https://academic.oup.com/database/article-lookup/doi/10.1093/database/bas025> (May 2012).
141. VanInsberghe, M., van den Berg, J., Andersson-Rolf, A., *et al.* Single-cell Ribo-seq reveals cell cycle-dependent translational pausing. *Nature* **597**, 561–565. ISSN: 0028-0836, 1476-4687. <https://www.nature.com/articles/s41586-021-03887-4> (2022) (Sept. 23, 2021).
142. Vazquez-Laslop, N., Sharma, C. M., Mankin, A., *et al.* Identifying Small Open Reading Frames in Prokaryotes with Ribosome Profiling. *Journal of Bacteriology* **204** (ed Henkin, T. M.) e00294–21. ISSN: 0021-9193, 1098-5530. <https://journals.asm.org/doi/10.1128/JB.00294-21> (2022) (Jan. 18, 2022).
143. Velleman, J. D. & Bratman, M. E. Intention, Plans, and Practical Reason. *The Philosophical Review* **100**, 277. ISSN: 00318108. <https://www.jstor.org/stable/2185304?origin=crossref> (2022) (1987).
144. Verbruggen, S., Gessulat, S., Gabriels, R., *et al.* Spectral Prediction Features as a Solution for the Search Space Size Problem in Proteogenomics. *Molecular & Cellular Proteomics* **20**, 100076. ISSN: 15359476. <https://linkinghub.elsevier.com/retrieve/pii/S1535947621000499> (2021) (2021).
145. Verbruggen, S., Ndah, E., Van Criekinge, W., *et al.* PROTEOFORMER 2.0: Further Developments in the Ribosome Profiling-assisted Proteogenomic Hunt for New Proteoforms. *Molecular & cellular proteomics: MCP* **18**, S126–S140. ISSN: 1535-9484 (Aug. 9, 2019).
146. Via, A., Uyar, B., Brun, C., *et al.* How pathogens use linear motifs to perturb host cell networks. *Trends in Biochemical Sciences* **40**, 36–48. ISSN: 09680004. <https://linkinghub.elsevier.com/retrieve/pii/S0968000414002059> (2022) (Jan. 2015).

147. Vitorino, R., Guedes, S., Amado, F., *et al.* The role of micropeptides in biology. *Cellular and Molecular Life Sciences*. ISSN: 1420-682X, 1420-9071. <http://link.springer.com/10.1007/s00018-020-03740-3> (2021) (Jan. 28, 2021).
148. Von Bohlen, A. E., Böhm, J., Pop, R., *et al.* A mutation creating an upstream initiation codon in the SOX9 5' UTR causes acampomelic campomelic dysplasia. *Molecular Genetics & Genomic Medicine* **5**, 261–268. ISSN: 23249269. <https://onlinelibrary.wiley.com/doi/10.1002/mgg3.282> (2022) (May 2017).
149. Wadie, B., Kleshchevnikov, V., Sandaltzopoulou, E., *et al.* *Use of viral motif mimicry improves the proteome-wide discovery of human linear motifs* preprint (Systems Biology, June 26, 2021). <http://biorxiv.org/lookup/doi/10.1101/2021.06.25.449930> (2021).
150. Walter, P. & Ron, D. The Unfolded Protein Response: From Stress Pathway to Homeostatic Regulation. *Science* **334**, 1081–1086. ISSN: 0036-8075, 1095-9203. <https://www.sciencemag.org/lookup/doi/10.1126/science.1209038> (2021) (Nov. 25, 2011).
151. Wan, J. & Qian, S.-B. TISdb: a database for alternative translation initiation in mammalian cells. *Nucleic Acids Research* **42**, D845–D850. ISSN: 0305-1048, 1362-4962. <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkt1085> (2022) (D1 Jan. 2014).
152. Wang, H., Yang, L., Wang, Y., *et al.* RPFdb v2.0: an updated database for genome-wide information of translated mRNA generated from ribosome profiling. *Nucleic Acids Research* **47**, D230–D234. ISSN: 0305-1048 (D1 Oct. 2018).
153. Weatheritt, R. J., Luck, K., Petsalaki, E., *et al.* The identification of short linear motif-mediated interfaces within the human interactome. *Bioinformatics* **28**, 976–982. ISSN: 1460-2059, 1367-4803. <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts072> (2022) (Apr. 1, 2012).
154. Welter, D., Osumi-Sutherland, D. & Jupp, S. Human Cell Atlas Ontology. *CEUR-WS.org* **2285**, 2 (2018).
155. Wethmar, K., Barbosa-Silva, A., Andrade-Navarro, M. A., *et al.* uORFdb—a comprehensive literature database on eukaryotic uORF biology. *Nucleic Acids Research* **42**, D60–D67. ISSN: 0305-1048, 1362-4962. <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkt952> (D1 Jan. 2014).
156. Wu, W.-S., Tsao, Y.-H., Shiu, S.-C., *et al.* A tool for analyzing and visualizing ribo-seq data at the isoform level. *BMC Bioinformatics* **22**, 271. ISSN: 1471-2105. <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-021-04192-7> (2021) (S10 May 2021).
157. Yang, T.-H., Wang, C.-Y., Tsai, H.-C., *et al.* Human IRES Atlas: an integrative platform for studying IRES-driven translational regulation in humans. *Database* **2021**, baab025. ISSN: 1758-0463. <https://academic.oup.com/database/article/doi/10.1093/database/baab025/6263636> (2021) (May 18, 2021).

158. You, R., Yao, S., Mamitsuka, H., *et al.* DeepGraphGO: graph neural network for large-scale, multispecies protein function prediction. *Bioinformatics* **37**, i262–i271. ISSN: 1367-4803, 1460-2059. [https://academic.oup.com/bioinformatics/article/37/Supplement\\_1/i262/6319663](https://academic.oup.com/bioinformatics/article/37/Supplement_1/i262/6319663) (2021) (Supplement\_1 Aug. 4, 2021).
159. Zaheed, O., Kiniry, S. J., Baranov, P. V., *et al.* Exploring Evidence of Non-coding RNA Translation With Trips-Viz and GWIPS-Viz Browsers. *Frontiers in Cell and Developmental Biology* **9**, 703374. ISSN: 2296-634X. <https://www.frontiersin.org/articles/10.3389/fcell.2021.703374/full> (2021) (Aug. 12, 2021).
160. Zanet, J., Benrabah, E., Li, T., *et al.* Pri sORF peptides induce selective proteasome-mediated protein processing. *Science* **349**, 1356–1358. ISSN: 0036-8075, 1095-9203. <http://www.sciencemag.org/cgi/doi/10.1126/science.aac5677> (Sept. 2015).
161. Zanet, J., Chanut-Delalande, H., Plaza, S., *et al.* in *Current Topics in Developmental Biology* 199–219 (Elsevier, 2016). ISBN: 978-0-12-801382-3. <https://linkinghub.elsevier.com/retrieve/pii/S0070215315001234>.
162. Zanzoni, A., Spinelli, L., Braham, S., *et al.* Perturbed human sub-networks by *Fusobacterium nucleatum* candidate virulence proteins. *Microbiome* **5**, 89. ISSN: 2049-2618. <http://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-017-0307-1> (2021) (Dec. 2017).
163. Zhang, H., Wang, Y., Wu, X., *et al.* Determinants of genome-wide distribution and evolution of uORFs in eukaryotes. *Nature Communications* **12**, 1076. ISSN: 2041-1723. <http://www.nature.com/articles/s41467-021-21394-y> (2021) (Dec. 2021).
164. Zhang, P., He, D., Xu, Y., *et al.* Genome-wide identification and differential analysis of translational initiation. *Nature Communications* **8**, 1749. ISSN: 2041-1723. <http://www.nature.com/articles/s41467-017-01981-8> (Dec. 2017).
165. Zhou, X., Huang, X. & Du, Z. *A Computational and Biochemical Study of -1 Ribosomal Frameshifting in Human mRNAs* preprint (Biochemistry, Apr. 24, 2021). <http://biorxiv.org/lookup/doi/10.1101/2021.04.23.441185> (2021).
166. Zhu, M. & Gribskov, M. MiPepid: MicroPeptide identification tool using machine learning. *BMC bioinformatics* **20**, 559. ISSN: 1471-2105 (Nov. 8, 2019).
167. Zimmer, M. H., Niesen, M. J. & Miller, T. F. Force transduction creates long-ranged coupling in ribosomes stalled by arrest peptides. *Biophysical Journal* **120**, 2425–2435. ISSN: 00063495. <https://linkinghub.elsevier.com/retrieve/pii/S0006349521003350> (2021) (June 2021).