InteractORF, predictions of human sORF functions from an interactome study

Mathilde Slivak, Sebastien Choteau, Philippe Pierre, Lionel Spinelli, Andreas Zanzoni, C. Brun

InteractORF, predictions of human sORF functions from an interactome study

Mathilde SLIVAK[1,&], Sébastien A. CHOTEAU[1,2,&,+], Philippe PIERRE[2,3,4], Lionel SPINELLI[1,2], Andreas ZANZONI[1], and Christine BRUN[1,5,*]

1 Aix-Marseille University, INSERM, TAGC, Turing Centre for Living Systems, 163 Avenue de Luminy, Marseille 13009, France

2 Aix-Marseille University, INSERM, CNRS, CIML, Turing Centre for Living Systems, 163 Avenue de Luminy, Marseille 13009, France,

3 Department of Medical Sciences, Institute for Research in Biomedicine (iBiMED) and Ilidio Pinho Foundation, University of Aveiro, Aveiro 3810-193, Portugal,

4 Shanghai Institute of Immunology, School of Medicine, Shanghai Jiao Tong University, Shanghai, China

5 CNRS, 31 Chemin Joseph Aiguier, Marseille 13009, France

+ Present address: IRSEA, 216 D201 Quartier Salignan, Apt 84400, France

& These authors contributed equally to the work.

*Corresponding author: christine-g.brun@inserm.fr

Keywords : sORF, microprotein, protein-protein interaction network, function prediction

Abstract

Short Open Reading Frames (sORFs) are ubiquitous genomic elements that have been overlooked for years, essentially due to their short length (< 100 residues) and the use of alternative start codons (other than AUG). However, some may encode functional peptides, so-called sORF-encoded peptides (sPEPs), the functions of which remain mainly 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 and analyzed the interfaces of interactions as well as the set of canonical proteins interacting with sPEPs. Second, by joining these sPEP-canonical proteins interactions with the human interactome, we predicted the first sPEP interactome network to date. Based on its topology, 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, metabolism, and signaling. Overall, the diversity in the predicted functions of the sPEPs underline the prevalence of their role in different biological mechanisms, suggesting that they are major regulatory actors.

Introduction

Open reading frames shorter than 100 codons were initially thought to be nonfunctional and discarded in most gene annotation programs with the notion they had no coding potential (1–5). More recent studies demonstrated that these sequences, called short open reading frames (sORFs), may actually encode functional peptides (5–9). 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) (5,6,10). 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 (11).

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 (12). Recent studies have demonstrated sPEPs to be involved in various cellular processes and diseases, notably cell proliferation, signaling, cell growth, death, metabolism or development (5). It has even been suggested that sPEPs may constitute a new pool of cancer-related peptides that could be targeted by immunotherapy (8). As an example, 168 novel major histocompatibility complex class I (MHC-I)-associated peptides derived from sORFs have been identified (13), demonstrating that sPEPs can also be involved in specialized functions such as antigen presentation.

Human monocytes are a heterogenous 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 (10,13,14). 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 (10,13).

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 (15). 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 is translated (9). 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. Only recently, a mass-spectrometry-based interactome screen with motif resolution, allowed predicting the functions of 226 sORF-encoded small peptides (16). However, no systematic large-scale 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 for which the functions are known and functional annotations are available. Indeed, protein-protein interactions (PPIs) drive biological functions and it has been demonstrated that protein functions can be assigned on the basis of the annotation of their neighbors in the PPI network (17). Hence, we hypothesize that analyzing the interactions between sPEPs and canonical proteins will allow performing a systematic functional annotation of the sPEPs. As we recently developed mimicINT (18), a computational method that allows inferring PPIs based on the presence of Short Linear Motifs (SLiMs) and globular domains in amino acid sequences, we herein predicted interactions between sPEPs and canonical proteins using this method, integrated those predicted interactions with the human interactome and studied network modularity to predict sPEP functions. We then investigated whether sPEPs do participate to specific functions in monocytes. For this, *(i)* we identified the SLiMs and domains with the highest occurrences as sPEP interaction interfaces to assess the biological processes to which sPEPs are participating; *(ii)* we predict sPEP functions, by analyzing the functional annotations of their protein partners in network clusters.

Methods

1- Collection of sPEPs identified in monocytes

The sequences of sPEPs have been collected from MetamORF (15) (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 (15). 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.

2- Collection of canonical proteins expressed in monocytes

All reviewed sequences of proteins experimentally identified in monocytes according to the Human

Proteome Atlas (19) have been downloaded from UniProtKB (20) as fasta format (Fig. 1B).

3- Prediction of sPEP-canonical proteins interactions

Interaction predictions between sPEPs and canonical proteins with mimicINT. Initially, mimicINT is a workflow for microbe-host protein interactions inference we recently developed (18). It performs large-scale interaction inferences between microbe and human proteins by detecting putative molecular mimicry elements that can mediate the interactions with host proteins. These elements are host-like short linear motifs (Fig. 1D) *i.e.* SLiMs, extracted from the ELM database (21) and globular domains (Fig. 1E) predicted using the existing PFAM signatures (22). Overall, 7086 SLiM occurrences have been detected on sPEPs (using SLiMProb (23)) and filtered based on p-values computed by Monte-Carlo simulations; 28 Pfam signatures of globular domains have been detected on canonical proteins and sPEPs (using InterProScan (24)); templates of domain-domain interactions (DDIs, from 3DID (25)) and of domain-SLiM interactions (DMIs, from ELM) were used to infer 3938 DDIs (17 domains interacting) and 78776 DMIs (5455 SLiMs interacting) between 1816 sPEPs and 1603 canonical proteins (Fig. 1F). DMIs were then filtered based on domain scores computed by looking for Hidden Markov Models. Because sPEPs and canonical proteins belong to the same species, we may reasonably expect that human sPEPs display interfaces of interactions that resemble structures of the canonical proteins at the molecular level. Based on this assumption, interactions between sPEPs and canonical proteins have been inferred using mimicINT. sPEPs containing truncated domains (with 10 missing aa allowed), possibly impairing the ability of the domain to mediate the predicted interactions, were discarded (*i.e.* 310/336 sPEPs-containg domains).

Clusterisation of the sORFs sequences. As many sORFs overlap on the same transcript, sequences have been clustered with CD-Hit (26) (95% identity) to enable us to investigate further only the representative sequence of the cluster. Interactions of all cluster members have been transferred to the cluster representative (the sORF with the longest sequence)

4- Functional annotations of the sPEPs based on network clustering

Merging the sPEP interactions network with the canonical protein-protein interaction network. The sPEP-canonical protein interactions network has been merged with the canonical PPI network downloaded from MoonDB (27) (2021 update, unpublished release) and restricted to the canonical proteins expressed in monocytes according to the Human Proteome Atlas (19). For the sake of clarity, the resulting network is referred hereafter as the “merged interactome”.

Clustering of the "merged interactome". The largest connected component has been extracted from the "merged interactome" using python-igraph (v0.9.1). This component has been clustered with OCG (28) (default parameters). Each generated cluster has then been annotated either with the Gene Ontology (GO) biological process (BP) terms significantly enriched among the GO terms annotating the cluster proteins (hypergeometric test, BH corrected p-value < 0.00001) or with the GO terms annotating at least 50% of the annotated canonical proteins of the cluster, following a classical majority rule (29). In both cases, all cluster members, canonical proteins and sPEPs, inherited the annotation(s) of the cluster. Network visualization have been performed using Cytoscape (30).

5- sPEPs' interactor annotation enrichment analyses

Enrichment analyses have been performed using gProfiler (31) (parameters: correction method = 'Benjamini-Hochberg FDR'). False discovery rates (FDR) lower than 0.05 have been considered as significant. Simplification of GO:BP and REACTOME terms have been done using goSlim function from R package GSEAbase (doi:10.18129/B9.bioc.GSEABase) and the pathway hierarchy downloaded from the Reactome database (32). Representations of inferred GO terms annotations (Fig. 5) have been made with the R package 'rrvgo' (33).

Results

1- Interactions between sPEPs and canonical proteins have been inferred in monocytes

We built here the first large-scale network of sPEP-protein interactions in human monocytes. For this, the amino acid sequences of 10475 putative sORF-encoded peptides (sPEPs) identified by ribosome profiling in monocytes and collected from MetamORF (15), — a repository of unique sORFs identified by computational and experimental methods that we previously developed (Fig. 1A) — and the amino acid sequences of the 11404 canonical proteins constituting the monocyte proteome according to the Human Protein Atlas (Fig. 1B), were used for interaction predictions. With mimicINT (18), a computational method we previously proposed to infer PPIs from sequences based on interaction templates made of SLiMs and domains (see Methods, Fig. 1C-E), 154407 binary interactions between 4393 sPEPs and 3981 canonical proteins have been predicted in monocytes (two interactions involving the same couple of sPEPs and canonical protein interactors but mediated through two different sets of interfaces are counted as two in the count of total interactions whilst counted as one in the number of binary interactions). After

elimination of truncated domains and clusterization of the sequences (see Methods), 40% of the sPEPs (1816/4521) are predicted to interact with 15% of the canonical proteins (1717/11404) in monocytes, for a total of 49613 binary interactions (Table 1).

Tab. 1 : *Counts of inferred sPEP-Canonical Protein interactions*

Type of interactions	Total interactions	Binary interactions
Domain-domain interactions (DDIs)	3,938	2,089
Domain-SLiM interactions (DMIs)	78,776	47,524
All interactions	82,714	49,613

2- The SLiMs and domains mediating interactions in sPEPs are related to signaling and immunology processes

We here aim at exploring sPEP functions by studying their interactions. First, interfaces of interactions (domains and SLiMs) provide information about the molecular functions of the proteins that harbor them. Indeed, interfaces may mediate interactions with other proteins that notably allow them to take part in complexes or pathways, to be addressed to certain subcellular compartments or to be submitted to post-translational modifications. Hence, we investigated the most commonly used interfaces on sPEPs to predict their putative functions.

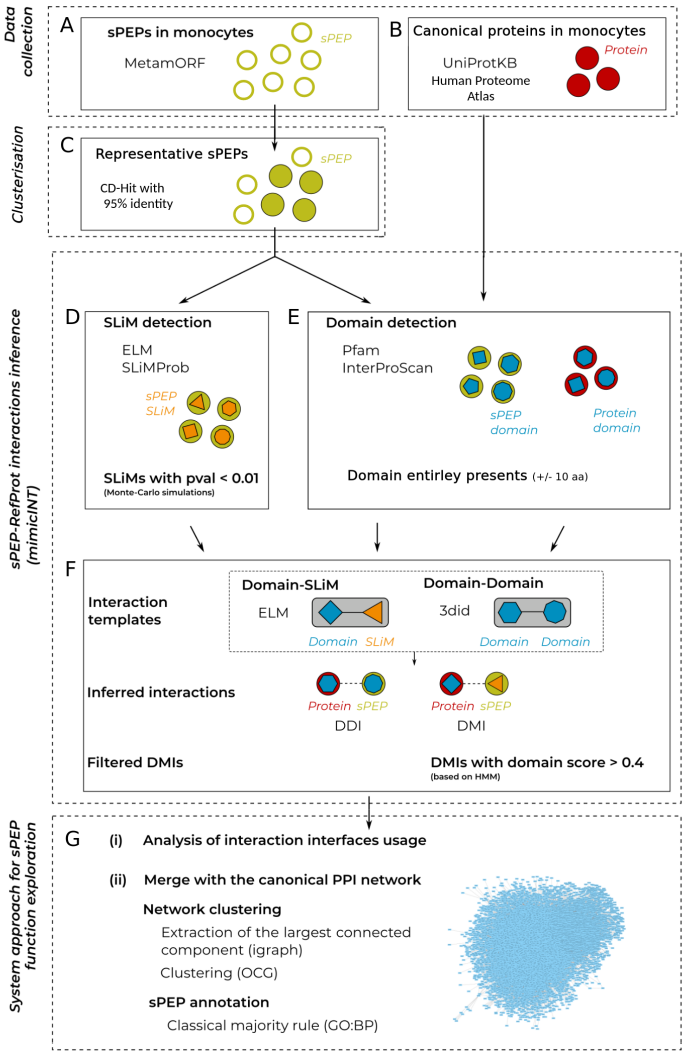

Fig. 1: *From sPEPs to function.* *(A) 10475 sPEPs sequences identified in monocytes have been collected from MetamORF and (B) 11417 canonical proteins sequences expressed in monocytes have been collected from UniProtKB. (C) After clustering of the sPEP sequences using CD-Hit, 4521 representative sequences of sPEPs are kept. (D-F) 49613 binary sPEP-canonical proteins interactions have been inferred using the mimicINT workflow. (G) A system approach has finally been used to explore the functions to which sPEPs participate.*

2.1- Most of the interaction interfaces on sPEPs are SLiMs related to housekeeping regulatory functions

We predicted 82714 SLiM-domain and domain-domain interactions corresponding to 49613 binary interactions between unique sPEPs and canonical proteins (Table 1). Moreover, a total of 7114 interfaces among which 7086 SLiMs and 28 domains have been identified on sPEPs (Table 2). In accordance with the

presence of multiple SLiMs on sPEPs, sPEP-protein interactions are mainly mediated by SLiMs. Indeed, SLiM occurrences constituted 99% of the interfaces of interactions harbored by sPEPs (7086/7114), leading to 95% of the predicted interactions (78776/82714), SLiM-domain interactions. Some sPEPs contain several SLiMs. Whereas 99% (1799/1816) of the sPEPs harbor at least one SLiM able to mediate interactions with canonical proteins, 77% of the sPEPs harbor between 1 and 3 of those. Domains correspond to 0,4% (28/7114) of the interaction interfaces on sPEPs. Overall, only 1% of the sPEPs contains domains able to mediate interactions (17/1816) with canonical proteins through 15 different domain-domain interaction templates. They account for 5% (3938/82714) of interactions between sPEPs and canonical proteins.

		sPEPs	Canonical Proteins
Counts in monocytes		4521	11,404
Interacting		1,816	1,717
Domains	#Occ.	28	21,258
	#Occ. interacting	17	651
	#Pfam domains interacting	15	227
SLiMs	#Occ.	7086	-
	#Occ. interacting	5455	-
	#ELM classes interacting	44	-

Tab. 2 : *Counts of domain and SLiM occurrences in sPEPs and Canonical Proteins*

Out of the six SLiM *motif types* defined in the ELM database (21) — ligand-binding sites (LIG), docking sites (DOC), subcellular targeting sites (TRG), post-translational modification sites (MOD), proteolytic cleavage sites (CLV) and degradation sites (DEG) —, the MOD class is preferentially used by the sPEPs to interact with canonical proteins (48%) whilst the DEG one is the less encountered (5%). This distribution differs from the one observed among canonical proteins, where the LIG and the CLV classes are the most and the fewest used, respectively (Fig. 2).

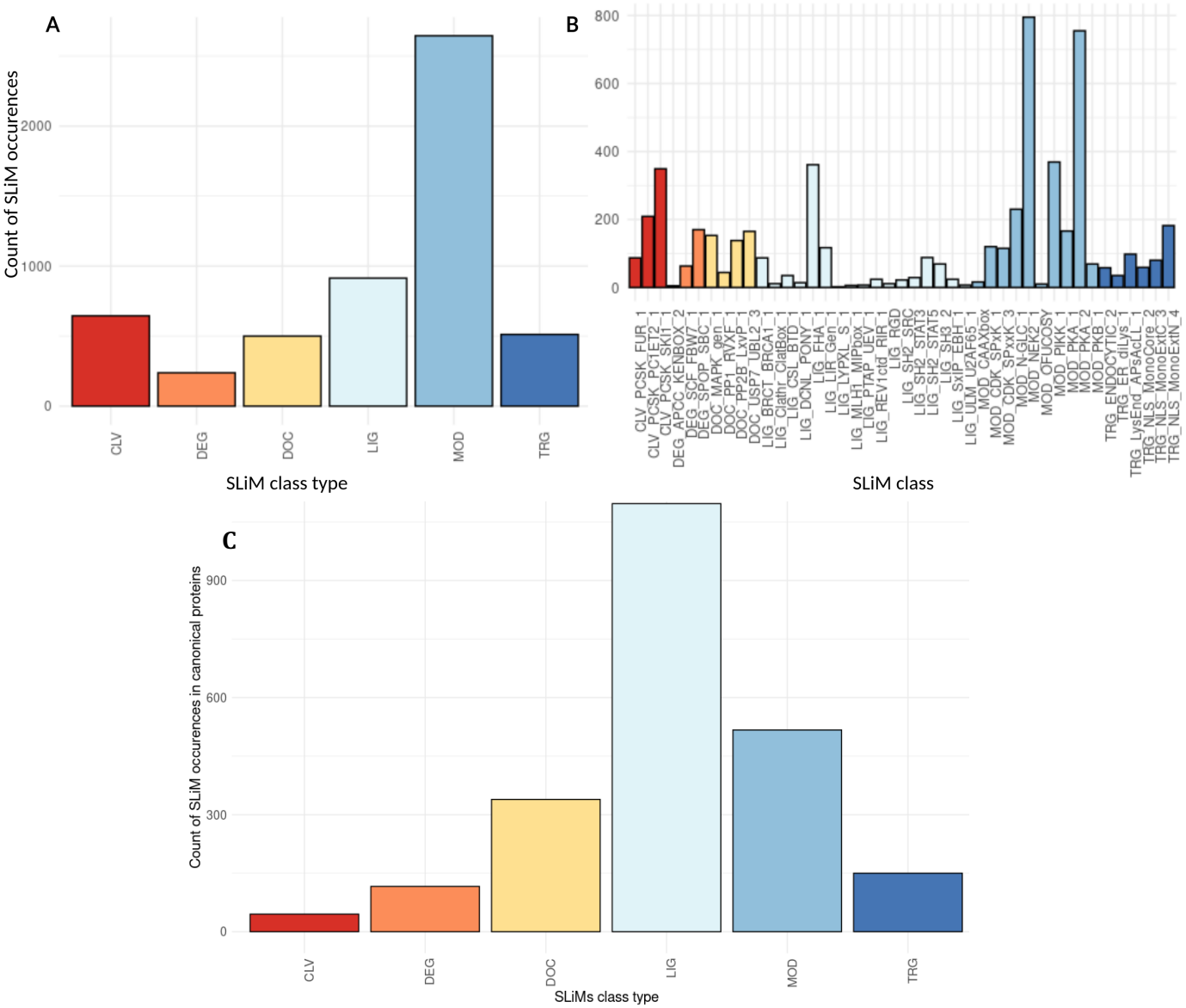

Fig. 2 : ***SLiM classes harbored by sPEPs.*** *For each SLiM class type, the count of occurrences has been computed (A), as well as the count of occurrences for each individual SLiM class (B). This distribution differs from the count of occurrence for SLiMs class types observed in canonical proteins according to the ELM database (only the true positive SLiM instances were analyzed) (C).*

As the *ELM classes* of SLiMs most commonly used to mediate interactions with canonical proteins are likely to provide insight about the biological processes in which the sPEPs are involved, we first only considered the 10 most commonly used ones. These classes are involved in cell cycle regulation, DNA repair, signaling, transport, transcriptional regulation and protein metabolism (Table 3) whereas the 15 interacting domains contained in sPEPs are mainly related to immunological responses, protein targeting, transcriptional

regulation and signaling (Table 4). Therefore, the analysis of the interaction interfaces comforts the hypothesis that sPEPs may be involved in signaling and in specific functions of monocytes, such as immunological responses. These findings are in line with our current knowledge of sPEP involvement notably in signaling and antigen presentation in eukaryotes (2,5,8,13,14).

SliM class	Site name (from ELM)	Function family	Pattern prob.	#Occ.[a]	#Interactions[b]
MOD_NEK2_1	NEK2 phosphorylation site	Cell cycle, DNA damage response	0.0097983	795	7950
MOD_PKA_2	PKA Phosphorylation site	Metabolism, signaling	0.0094575	755	22650
MOD_PIKK_1	PIKK phosphorylation site	DNA repair and checkpoint	0.0092301	369	1476
LIG_FHA_1	FHA phosphopeptide ligands	Cell cycle, DNA repair, transcriptional regulation	0.0086622	361	722
CLV_PCSK_SKI1_1	PCSK cleavage site	Proteolytic processing of neuropeptide, peptide hormone precursor	0.0068205	349	349
MOD_N-GLC_1	N-glycosylation site	Protein metabolism	0.0050178	230	230
CLV_PCSK_PC1ET2_1	PCSK cleavage site	Proteolytic processing of neuropeptide, peptide hormone precursor	0.0039028	209	627
TRG_NLS_MonoExtN_4	NLS classical Nuclear Localization Signals	Nuclear localization signal, transport	0.0012764	182	6916
DEG_SPOP_SBC_1	SPOP SBC docking motif	Protein degradation	0.000938	170	340
MOD$_{PKA_1}$	PKA Phosphorylation site	Metabolism, signalisation	0.0037418	166	4980

[a]: Number of occurrences of the SLiMs in sPEPs
[b]: Number of sPEPRIs mediated by the SLiMs

Tab. 3 : *Top 10 SLiM classes based on occurrence counts in sPEPs*

Pfam accession	Domain name	Function family	#Occ.[a]	#Interactions[b]
PF07654	Immunoglobulin C1-set domain	Immunology	1	837
PF00018	SH3 domain	Signal transduction	1	847
PF00036	EF hand	Protein targeting/transport	1	600
PF12162	STAT1 TAZ2 binding domain	Transcriptional activator	1	4
PF07145	Ataxin-2 C-terminal region	Protein targeting	1	4
PF00178	Ets-domain	Transcription factors	1	280
PF01023	S-100/ICaBP type calcium binding domain	Calcium binding	1	271
PF10584	Proteasome subunit A N-terminal signature	Catabolic proccess	1	102
PF06623	MHC-I C-terminus	Immunology	1	9
PF00008	EGF-like domain	Protein targeting	1	500
PF00048	Small cytokines (intecrine/chemokine), interleukin-8 like	Immunology	3	450
PF00098	Zinc knuckle	Gene regulation	1	26
PF10215	Oligosaccaryltransferase	Glycosylation of proteins	1	2
PF11627	Nuclear factor hnRNPA1	RNA-related activities	1	4
PF03951	Glutamine synthetase, beta-Grasp domain	Metabolism of nitrogen	1	2

[a]: Number of occurrences of the domain in sPEPs
[b]: Number of sPEPRIs mediated by the domain

Tab. 4 : *Domains in sPEPs*

2.2- The interacting partners of sPEPs inform on sPEPs 'functions.

To strengthen our hypotheses, we next investigated the biological process(es) in which the canonical proteins interacting with sPEPs are involved. To address this question, we looked for GO term, KEGG and Reactome pathway statistical enrichments among the annotations of the canonical proteins interacting with at least one sPEP.

First, SLiMs' interactors are significantly enriched in 1794 GO:BP terms (Fig. 3A-B), most of which are related with anatomical structure development (24% of the GO terms, including anatomical structure morphogenesis, multicellular organism development, cell morphogenesis, vasculature development,

nervous system development, etc.), immune system (immune response, inflammatory response, response to cytokine, leukocyte activation, etc.) and signaling (signaling, cell communication, intracellular signal transduction, regulation of signal transduction, etc.)

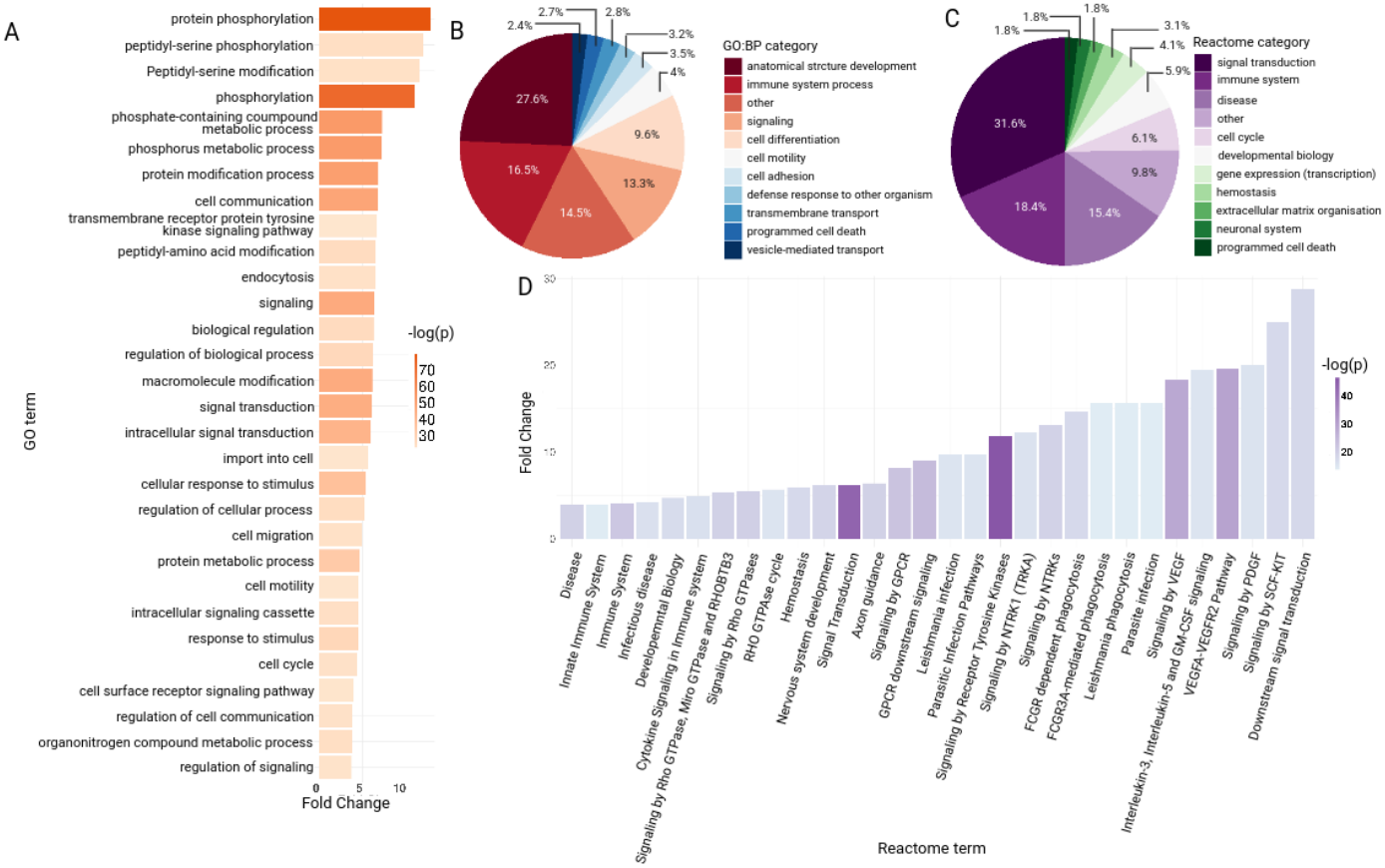

Fig. 3 : ***Representation of the functional enrichment of SLiMs' interactors.*** *(A) Representation of the 30 most enriched GO:BP terms and (B) percentage of GO:BP enriched terms according to GO slim terms. (D) Representation of the 30 most enriched Reactome terms and (C) percentage of Reactome terms enriched according to their root terms.*

Domains'interactors are significantly enriched in 2476 GO:BP terms (Fig. 4A-B) mainly related to the same categories. The most statistically enriched functions of proteins interacting with both types of interfaces are related to phosphorylation and phosphorus metabolism (protein phosphorylation, phosphate-containing compound metabolic process, phosphorus metabolic process, protein autophosphorylation, etc.) This link with signal transduction is confirmed by the strong enrichment of both interactor types among Reactome signaling pathways (Fig.3C-D, Fig.4C-D, more than 31% among which signal transduction, signaling by VEGF, GPCR downstream signaling, signaling by Rho GTPases for the SLiMs' interactors and

signal transduction, signaling by GPCR, Signaling by Interleukins and Signaling by Receptor Tyrosine Kinases for domains' interactors). Immune system and disease related terms are also enriched.

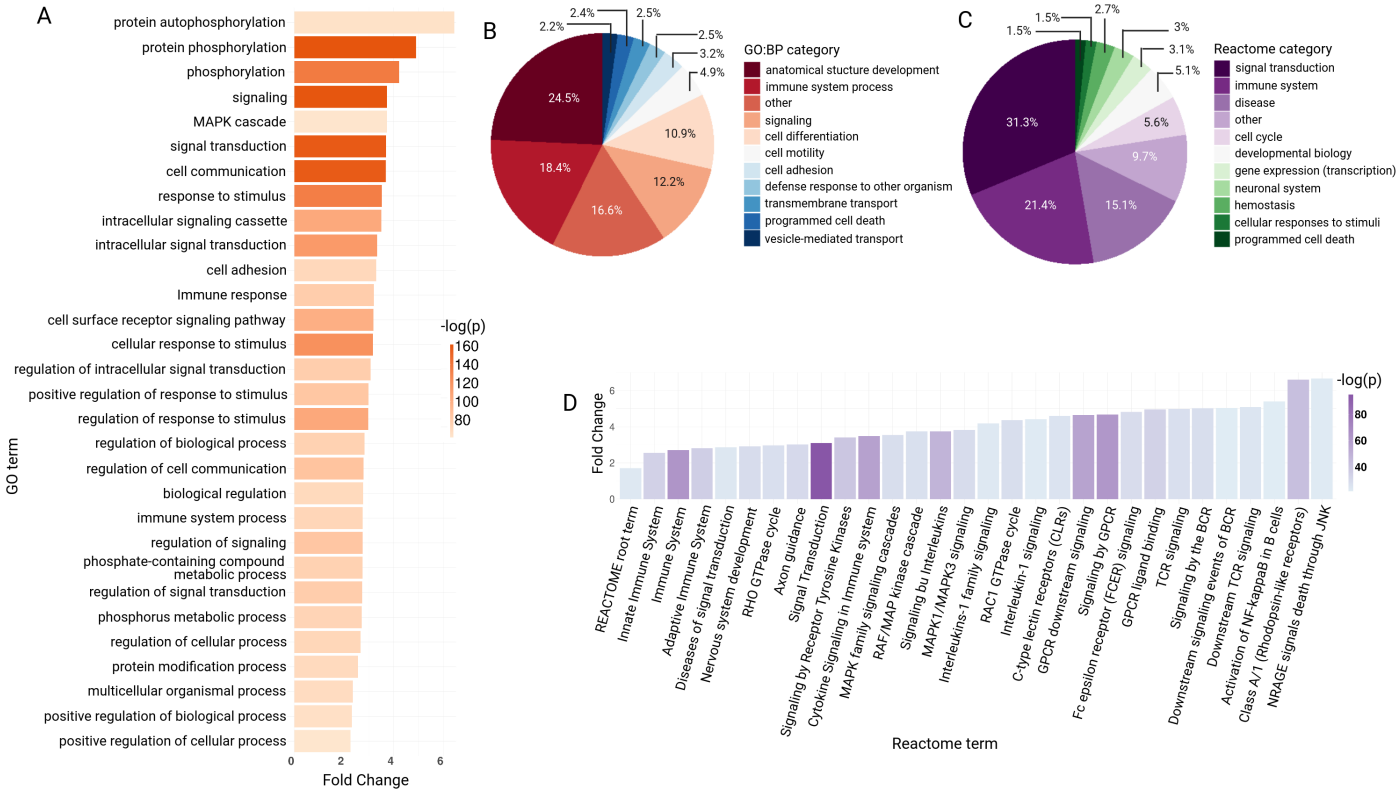

Fig. 4 : ***Representation of the functional enrichment of domains' interactors.*** *(A) Representation of the 30 most enriched GO:BP terms and (B) percentage of GO:BP enriched terms according to GO slim terms. (D) Representation of the 30 most enriched Reactome terms and (C) percentage of Reactome categories enriched.*

3- **sPEPs are major regulatory peptides**

Proteins involved in the same complexes or metabolic pathways are known to cluster in the canonical PPI network, and the topology of the PPI network has been successfully exploited in the past to perform assignment of cellular functions to uncharacterized proteins (17). Consequently, our second objective was to take advantage of the sPEP-protein interaction to predict sPEP functions by including those in the human PPI network. Then, we identified overlapping clusters in the merged network using our algorithm OCG (28) (see Methods).

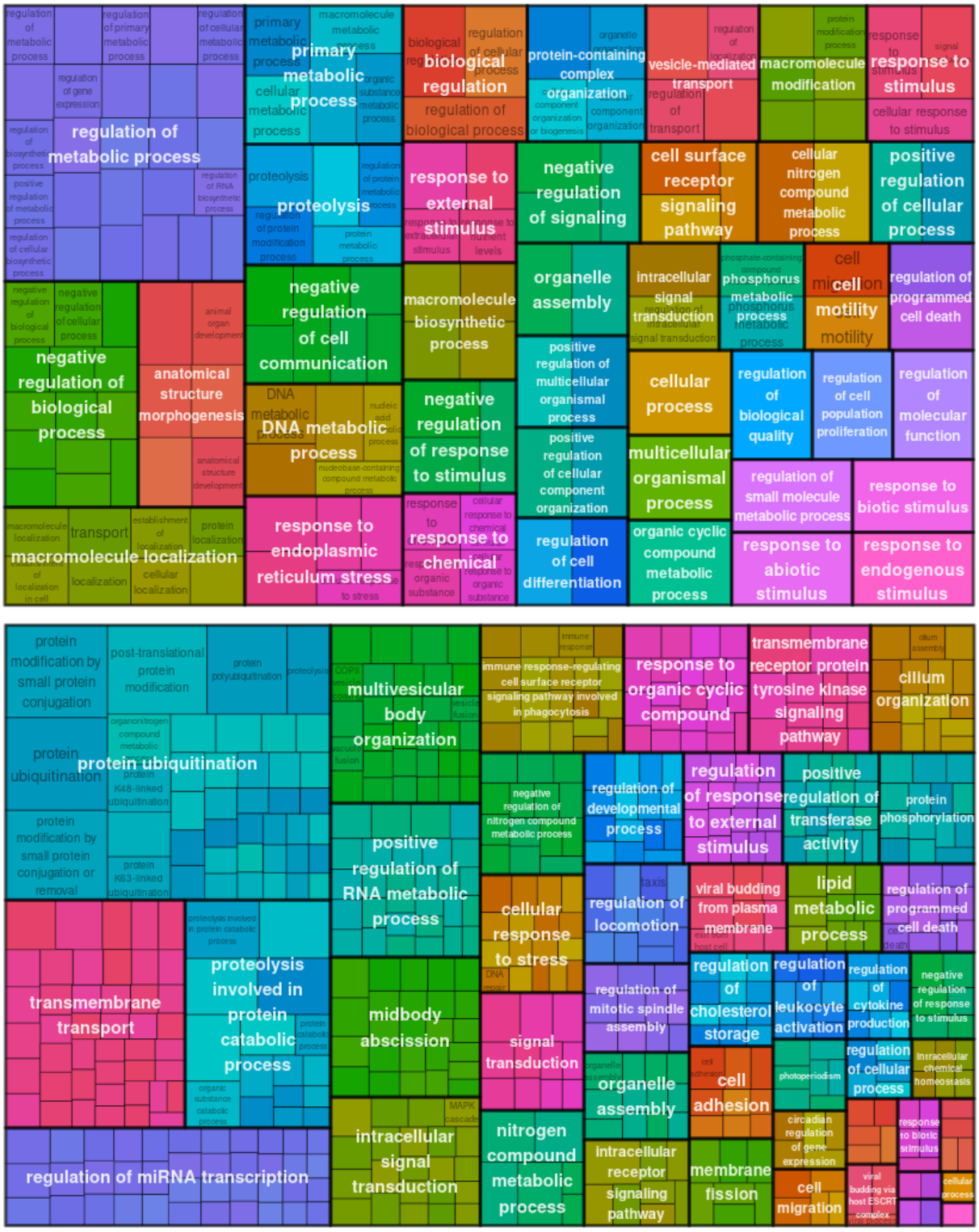

Fig. 5 : ***Annotations of all network clusters containing at least one sPEP*** *(A) according to the majority rule ≥ 50%, (B) with terms statistically over-represented (corrected pvalue ≤ 10^{-5}). Each compartment is representing the proportion of clusters annotated to this term and its child terms.*

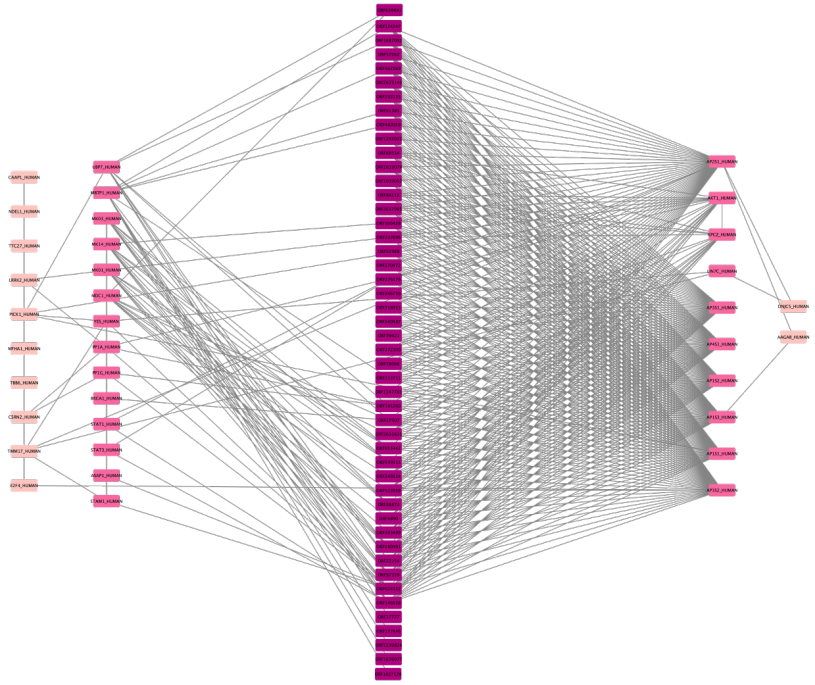

Fig. 6 : *Example of cluster statistically enriched in sPEPs.* *Cluster 302 contains 84 proteins among which 48 are sORFs (corrected p-val 2,38.10^{-16}). Purple nodes = sORFs; dark pink nodes = first neighbors; light pink = second neighbors. This cluster has been annotated with a majority rule ≥ 50%, the functions inferred to the proteins of this cluster are localization, cellular localization, transport, establishment of localization, signal transduction, positive regulation of cellular process, cellular component organization, positive regulation of biological process, response to stimulus, cellular component organization or biogenesis, negative regulation of biological process, regulation of biosynthetic process, regulation of macromolecule metabolic process and regulation of metabolic process. None of these terms are statistically enriched*.

Clusters are then annotated with the Biological Process (GO:BP) terms either designating at least half of the canonical proteins of the cluster (majority rule) or with statistically over-represented GO:BP terms among those annotating the annotated cluster proteins (Fig.5). Among the 368 clusters composing the network, 201 (*i.e.* 54%) contain at least one sPEP and their size varies from 7 to 970 proteins per cluster (Fig.S1). These clusters are mainly annotated to metabolic and signaling functions, regulation of different cellular processes and localization/transport related functions (Fig. 5). Only 6 clusters of very large size

(containing from 84 to 826 proteins), annotated to terms related to signaling, localization and metabolism, are statistically enriched in sPEPs (*e.g.* Fig. 6). Cluster annotations are then transferred to the sPEPs they contain. By doing so, we predicted the function(s) of the 1816 sPEPs contained in the network (Fig.7). As clusters are annotated with several GO:BP terms, sPEPs are multi-annotated as well (39 terms/sPEP on average). The annotation using GO:BP terms are in accordance with the annotation using Reactome terms. For example, the cluster 81 has been annotated with both GO:BP and Reactome terms related to RNA metabolic process (nucleic acid metabolic process/transcription initiation at RNA polymerase II promoter for GO:BP and RNA Polymerase II Pre-transcription Events/RNA Polymerase II Transcription Initiation/mRNA Splicing - Major Pathway for Reactome). It is to note that although numerous, these terms describe related functions for a single sPEP. From the distribution of the terms annotating all the sPEPs (Fig.7), 50% of the inferred terms relate to metabolism, 30% to the regulation of different cellular processes, 15% to signaling and 3% to transport/localization. A vast majority of sPEPs has been annotated with such general terms (1714 sPEPs are annotated with 'primary metabolic process', 1647 with 'positive regulation of metabolic process', 837 with 'intracellular signal transduction' and 544 with 'vesicle mediated pathway'), but looking at the exact terms inferred to our sPEPs allows getting a more precise knowledge of their specific functions (Fig.S2). Strikingly, the diversity of the regulation in which sPEPs are predicted to be involved suggests they are major regulatory peptides.

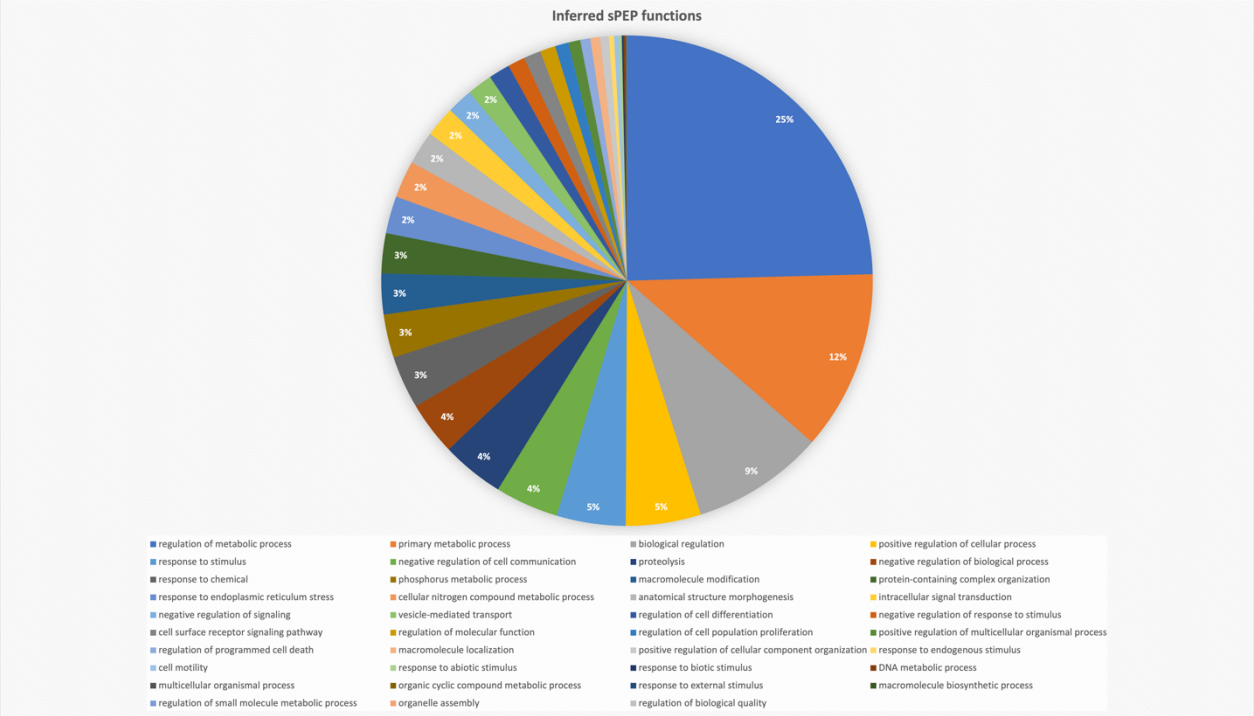

Fig. 7 ***: Inferred sPEP functions.*** *Percentages in the pie slices < 2% are not shown. Functions are ordered in the legend horizontally in a decreasing order.*

Discussion

To our knowledge, this study is the first to present a network of sPEP-canonical proteins interactions in *H. sapiens* as well as GO term annotations for human sPEPs at a large-scale. 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, signaling or immunology processes. Then, we investigated the topology of the sPEP-canonical protein interaction network 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, regulation of different processes, signal transduction and localization. Overall, our results suggest that most of the sPEPs are likely to be involved in biological processes both central to the cell and related to specialized biological functions such as immunological responses.

However, this study has been performed exclusively on human monocyte data, and our findings have been discussed in the scope of this particular species and cell type. 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 occupancy is not necessarily associated with an effective translation of a functional protein (12,35).

The interactions of sPEPs with canonical proteins 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-canonical protein 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 consider the subcellular location of canonical proteins and sPEPs nor the accessibility of the interaction interfaces for the interactors, making it prone to over-estimation of interactions. During the course of this study, the first dataset of experimentally determined interactions of sPEPs has been released (16). Although this could have allowed us to assess the quality of our predictions, the absence of common sPEPs between the two datasets (60 vs. 1816 sPEPs) hinders any comparison. A future development of the project in which the interactomes of sPEPs translated in other tissues are investigated should allowed us

to reach this goal.

Overall, our findings on the function of the sPEPs underline their importance for the regulation of cellular processes. Indeed sPEPs, although overlooked so far, should now be considered as novel major regulation actors.

Availability and Implementation

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).

Acknowledgments

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

Funding Information

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). MS master student stipend is taken from a Human Frontier Science Program grant to CB (RGP004/2023).

References

1. Andrews SJ, Rothnagel JA. Emerging evidence for functional peptides encoded by short open reading frames. Nat Rev Genet. mars 2014;15(3):193-204.
2. Couso JP, Patraquim P. Classification and function of small open reading frames. Nat Rev Mol. juill 2017;18(9):575-89.
3. Olexiouk V, Crappé J, Verbruggen S, Verhegen K, Martens L, Menschaert G. sORFs.org: a repository of small ORFs identified by ribosome profiling. Nucleic Acids Res. janv 2016;44(D1):D324-9.
4. Tharakan R, Sawa A. Minireview: Novel Micropeptide Discovery by Proteomics and Deep Sequencing Methods. Front Genet. 2021;12:651485.
5. Vitorino R, Guedes S, Amado F, Santos M, Akimitsu N. The role of micropeptides in biology. Cell Mol Life Sci. avr 2021;78(7):3285-98.
6. Plaza S, Menschaert G, Payre F. In Search of Lost Small Peptides. Annu Rev Cell Dev Biol. oct 2017;33(1):391-416.
7. Pueyo JI, Magny EG, Couso JP. New peptides under the s(ORF)ace of the genome. Trends Biochem Sci. août 2016;41(8):665-78.
8. Renz PF, Valdivia-Francia F, Sendoel A. Some like it translated: small ORFs in the 5'UTR. Exp Cell Res. 1 nov 2020;396(1):112229.
9. Hinnebusch AG, Ivanov IP, Sonenberg N. Translational control by 5'-untranslated regions of eukaryotic mRNAs. Science. 17 juin 2016;352(6292):1413-6.

10. Cabrera-Quio LE, Herberg S, Pauli A. Decoding sORF translation – from small proteins to gene regulation. RNA Biol. nov 2016;13(11):1051-9.
11. Brunet MA, Leblanc S, Roucou X. Reconsidering proteomic diversity with functional investigation of small ORFs and alternative ORFs. Exp Cell Res. 1 août 2020;393(1):112057.
12. Gray T, Storz G, Papenfort K. Small Proteins; Big Questions. J Bacteriol. 18 janv 2022;204(1):e0034121.
13. Laumont CM, Daouda T, Laverdure JP, Bonneil É, Caron-Lizotte O, Hardy MP, et al. Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. Nat Commun. janv 2016;7:10238.
14. Erhard F, Halenius A, Zimmermann C, L'Hernault A, Kowalewski DJ, Weekes MP, et al. Improved Ribo-seq enables identification of cryptic translation events. Nat Methods. mars 2018;15(5):363-6.
15. Choteau SA, 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. Database (Oxford). 22 juin 2021;2021:baab032.
16. Sandmann CL, Schulz JF, Ruiz-Orera J, Kirchner M, Ziehm M, Adami E, et al. Evolutionary origins and interactomes of human, young microproteins and small peptides translated from short open reading frames. Mol Cell. 16 mars 2023;83(6):994-1011.e18.
17. Brun C, Chevenet F, Martin D, Wojcik J, Guénoche A, Jacq B. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. Genome Biol. 2003;5(1):R6.
18. Choteau SA, Cristianini M, Maldonado K, Drets L, Boujeant M, Brun C, et al. mimicINT: a workflow for microbe-host protein interaction inference [Internet]. bioRxiv; 2022 [cité 20 mars 2024]. p. 2022.11.04.515250. Disponible sur: https://www.biorxiv.org/content/10.1101/2022.11.04.515250v1
19. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, et al. Towards a knowledge-based Human Protein Atlas. Nat Biotechnol. déc 2010;28(12):1248-50.
20. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research. 6 janv 2023;51(D1):D523-31.
21. Kumar M, Gouw M, Michael S, Sámano-Sánchez H, Pancsa R, Glavina J, et al. ELM—the eukaryotic linear motif resource in 2020. Nucleic Acids Research. 8 janv 2020;48(D1):D296-306.
22. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Res. 8 janv 2021;49(D1):D412-9.
23. Edwards RJ, Paulsen K, Aguilar Gomez CM, Pérez-Bercoff Å. Computational Prediction of Disordered Protein Motifs Using SLiMSuite. Methods Mol Biol. 2020;2141:37-72.
24. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 1 mai 2014;30(9):1236-40.
25. Mosca R, Céol A, Stein A, Olivella R, Aloy P. 3did: a catalog of domain-based interactions of known three-dimensional structure. Nucleic Acids Res. janv 2014;42(Database issue):D374-379.
26. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics [Internet]. déc 2012;28(23). Disponible sur: http://dx.doi.org/10.1093/bioinformatics/bts565
27. Ribeiro DM, Briere G, Bely B, Spinelli L, Brun C. MoonDB 2.0: an updated database of extreme multifunctional and moonlighting proteins. Nucleic Acids Res. 8 janv 2019;47(D1):D398-402.
28. Becker E, Robisson B, Chapple CE, Guénoche A, Brun C. Multifunctional proteins revealed by overlapping clustering in protein interaction network. Bioinformatics. 1 janv 2012;28(1):84-90.
29. Chapple CE, Robisson B, Spinelli L, Guien C, Becker E, Brun C. Extreme multifunctional proteins identified from a human protein interaction network. Nat Commun. 2015;6:7412.
30. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. nov 2003;13(11):2498-504.
31. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2 juill 2019;47(W1):W191-8.
32. Milacic M, Beavers D, Conley P, Gong C, Gillespie M, Griss J, et al. The Reactome Pathway Knowledgebase 2024. Nucleic Acids Res. 5 janv 2024;52(D1):D672-8.
33. Sayols S. rrvgo: a Bioconductor package for interpreting lists of Gene Ontology terms. MicroPubl Biol. 2023;2023.
34. Makarewich CA, Olson EN. Mining for Micropeptides. Trends Cell Biol. sept 2017;27(9):685-96.
35. UniProt: the universal protein knowledgebase in 2021 - PubMed [Internet]. [cité 22 mars 2024]. Disponible sur: https://pubmed.ncbi.nlm.nih.gov/33237286/

Supplementary figures:

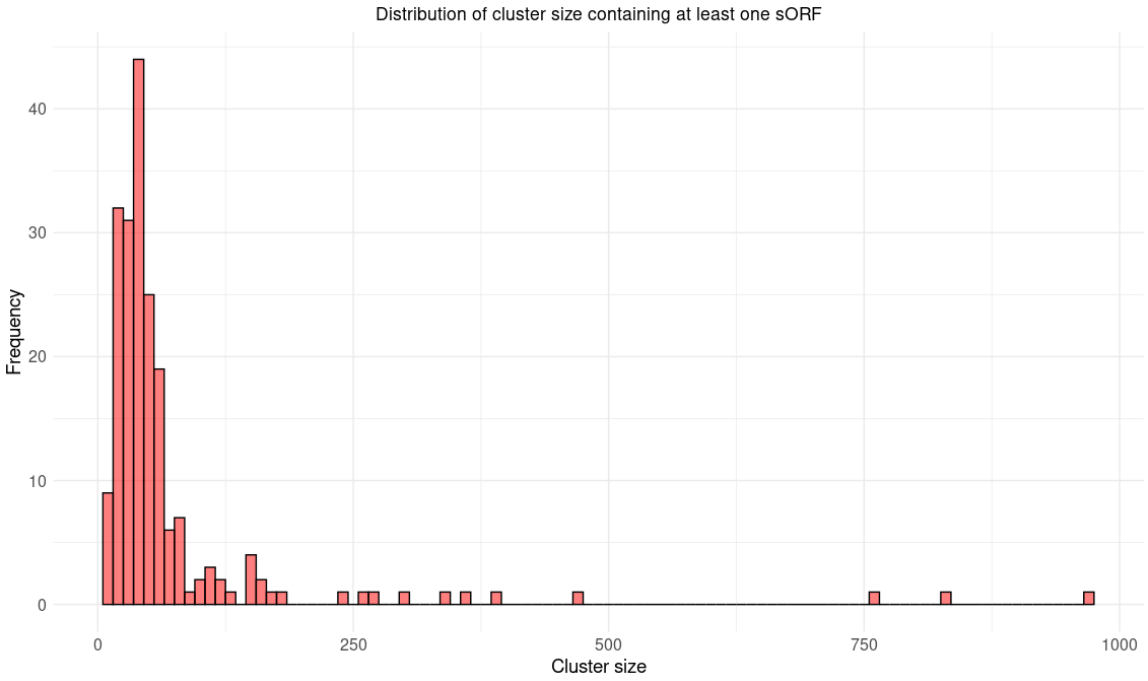

Fig. S1 ***: Distribution of cluster size.*** *Size of the 201 clusters containing at least one sPEP, ranging from 7 up to 970 proteins .*

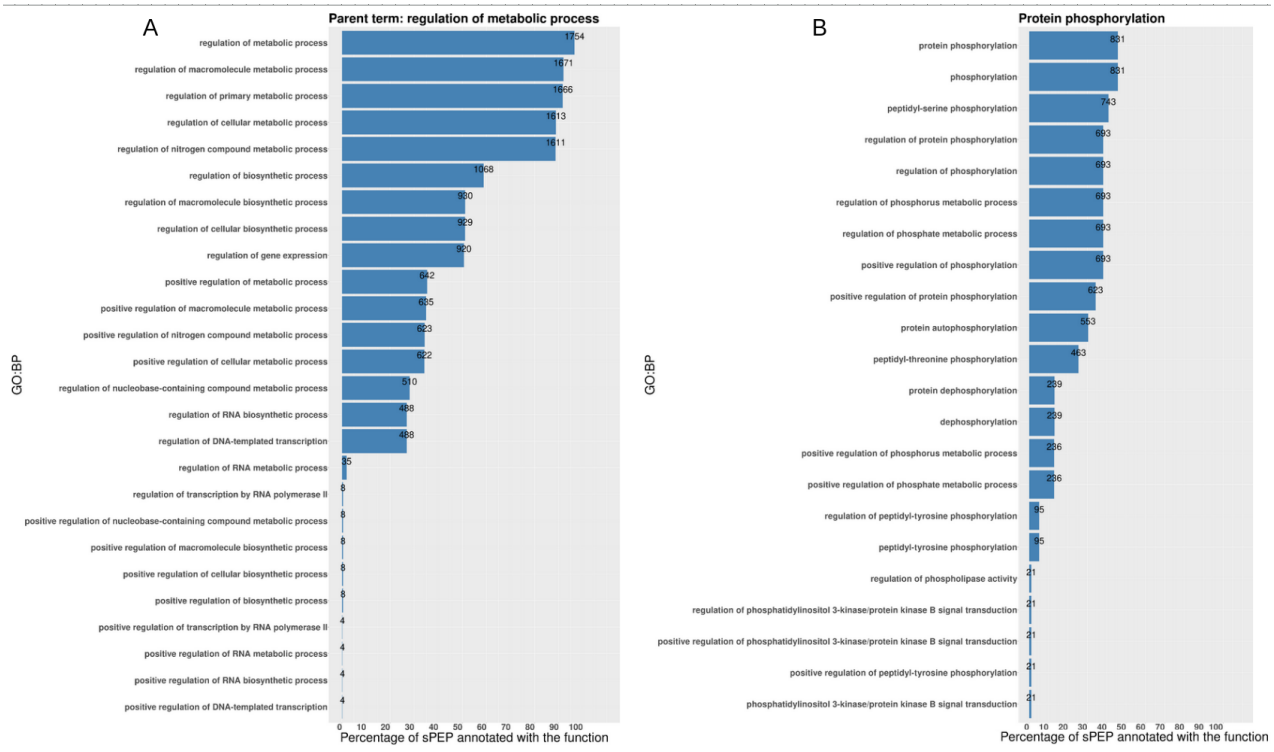

Fig. S2 ***: Percentage of exact terms inferred to sPEPs.*** *Percentage and count of sPEPs annotated with terms corresponding to the parent terms "regulation of metabolic process (A)" and "protein phosphorylation" (B). The percentage of sPEPs annotated with the function are represented in the x axis, and count of sPEPs annotated for each child term is shown overlapping with the bar.*