Nuclei of HeLa cells interactomes unravel a network of ghost proteins involved in proteins translation

Tristan Cardon, Michel Salzet, Julien Franck, Isabelle Fournier

NUCLEI OF HELA CELLS INTERACTOMES UNRAVEL A NETWORK OF GHOST PROTEINS INVOLVED IN PROTEINS TRANSLATION

Tristan Cardon, Michel Salzet*, Julien Franck* and Isabelle Fournier*

Université de Lille, Inserm, U1192 - Laboratoire Protéomique, Réponse Inflammatoire et Spectrométrie de Masse (PRISM), F-59000 Lille, France

SUMMARY

Ghost proteins are issued from alternative Open Reading Frames (ORFs) and are missing a genome annotation. Indeed, historical filters applied for the detection of putative translated ORFs led to a wrong classification of transcripts considered as non-coding although translated proteins can be detected by proteomics. This Ghost (also called Alternative) proteome was neglected, and one major issue is to identify the implication of the Ghost proteins in the biological processes. In this context, we aimed to identify the protein-protein interactions (PPIs) of the Ghost proteins. For that, we re-explored a cross-link MS study performed on nuclei of HeLa cells using cross-linking mass spectrometry (XL-MS) associated with the HaltOrf database. Among 1679 cross-link interactions identified, 292 are involving Ghost Proteins. Forty-Four of these Ghost proteins are found to interact with 7 Reference proteins related to ribonucleoproteins, ribosome subunits and zinc finger proteins network. We, thus, have focused our attention on the heterotrimer between the RE/poly(U)-binding/degradation factor 1 (AUF1), the Ribosomal protein 10 (RPL10) and AltATAD2. Using I-Tasser software we performed docking models from which we could suggest the attachment of AUF1 on the external part of RPL10 and the interaction of AltATAD2 on the RPL10 region interacting with 5S ribosomal RNA as a mechanism of regulation of the ribosome. Taken together, these results reveal the importance of Ghost Proteins within known protein interaction networks.

*Correspondance: **Prof. Isabelle Fournier** (isabelle.fournier@univ-lille.fr), **Dr. Julien Franck** (julien.franck@univ-lille.fr) & **Prof Michel Salzet** (Michel.salzet@univ-lille.fr), Laboratoire Réponse Inflammatoire et Spectrométrie de Masse (PRISM), Inserm U1192 - Université de Lille, Faculté des Sciences, Campus Cité Scientifique, Bât SN3, 1er étage, F-59655 Villeneuve d'Ascq Cedex. Phone: +33 (0)3 20 43 41 94; Fax: +33 (0)3 20 43 40 54

KEYWORDS

Cross-linking mass spectrometry, Alternative proteins, Ghost proteins, Alternative Open Reading Frame, Translation, disuccinimidyl sulfoxide

1. INTRODUCTION

Advances in mass spectrometry (MS) instrumentation and bioinformatics tools have led to an exponential increase of MS-based proteomics strategy performances. It is now possible to get the identification and relative quantification of more than 10,000 proteins in 100 min as recently published by Meier et al.[1]. These MS-based shot-gun strategies were also extended to structural characterization of the identified proteins. Various approaches were proposed over the past 15 years in proteomics for measuring protein-protein interactions (PPIs). Among these are the affinity capture [2], proximity labeling methods such as Apex [3,4], BioID [5–7] and Virotrap [8]; and the cross-linking mass spectrometry (XL-MS) [9]. XL-MS is advantageously non-targeted, providing a global onset for systems biology. However, all these strategies rely on protein database interrogation. Therefore, only referenced proteins can be identified [10,11]. This is a clear limitation for discovery if the databases are not complete. Public databases are built on both measured proteins and predicted ones. Predicted proteins are deduced from genome information accordingly to well-defined rules of annotation but are not all experimentally validated. The rules used for predicting protein sequences include the number of codons, the type of sequence and the Kozak context, which predicts the ribosome binding capacity on an mRNA [10,12–14]. Indeed, only the longest open reading frame (ORF) (so-called reference ORF, RefORF) or protein-coding

sequences (CDSs) is considered per transcript in the databases (e.g. Ensembl [15] and GENCODE [16]), other ORFs being excluded from annotation [17]. In particular, short ORFs (sORFs) or small ORFs (smORFs) that do not respect the 100 codons (300 nucleotides) cut-off rule or the Kozak code, alternative ORFs (AltORF) remain unannotated. However, the proteome is more complex than initially expected and with recent advances in the field of genomics and MS-based proteomics with the high throughput sequencing technologies, it has been shown that traditional computational genome annotation algorithms have underestimated the number of coding sequences leaving out alternative promotors [18], alternative splicing [19], alternative polyadenylation [20] and ribosomal frameshifting [21]. There is, thus, a major challenge for genome annotation to reference all these new ORFs which are left out despite leading to proteins presenting biological activities. One difficulty, is to be able to distinguish in this rising number of ORFs, the ORFs which are translated into functional proteins (such as microproteins, micropetides or SEPs) from the small ORFs that are randomly present but not translated. The smORFs/sORFs/AltORFs are often distinguished from the RefORFs because they are shorter size leading to the translation of small proteins (<30kDa). In average proteins translated from AltORFs are 57 amino acids in size, when by contrast the RefORFs proteins are 344 amino acids [22–26]. Importantly, these microproteins are not proteoforms of annotated proteins but have a different primary structure. Different computational approaches were used to identify these novel coding ORFs and create new databases including the predicted "alternative" transcripts (HaltORF [27], OpenProt [23], smProt [28]). The proteins issued from these AltORFs are called Ghost or Alternative Proteins (AltProts). Interestingly, proteomics has largely contributed to experimentally evidence and validate the existence of AltProts. Indeed, RefProts and AltProts were both detected from various studies by bottom-up [29,30] and top-down [31,32] proteomics. Interestingly, these proteins were identified within the 15% of proteomics data remaining unmatched after database interrogation despite a good quality MS/MS spectra; thus bridging the gap between experimental and predicted data.

If the discovery of these AltProts was definitively a revolution in the approach to systems biology, there is a clear unmet goal to find out the functions of these proteins.

Absolute quantification by stable isotope-labelling and parallel reaction monitoring (PRM) was used to determine the levels of the two MIEF1 gene translational products, the reference MiD51 and the alternative MiD51 (AltMiD51) proteins, in two human cells lines and human colon tissues. This study has revealed a twofold higher expression of AltMiD51 compared to MiD51 [22] reinforcing the conviction that AltProts are major players in the regulation of biological systems. Studies have, indeed, demonstrated that AltProts can be important regulators in many fundamental events such as DNA repair [33], RNA decapping [34] , calcium homeostasis metabolism [35], mTor signaling pathway [36], muscle performance [37], myoblast formation [38] and mitochondria fission [22]. Recently, it was shown that unannotated Heat Shock Protein [39] and Cold Shock Protein [40] were identified in *E. coli* by means of MS based proteomics. Specific AltProts were also found to be involved in physiopathological mechanisms including cancer and Spinal Cord Injury [30,32,41]. One step forwards the function of AltProts, is the identification of their interactome, by measuring PPI to gather information on the signaling pathways they are involved in [42]. Several studies have recently highlighted the adequacy of large scale interactomics XL-MS method as a discovery tool for new interactions [43,44]. In this context, we were willing to re-explore, using the HaltOrf database [27], a dataset of XL-MS from HeLa cells nuclei previously published by Heck group [44]. This has lead us to demonstrate the ability of XL-MS technique to discover previously unrevealed AltProt-RefProt interactions.

2. MATERIAL AND METHODS

2.1. Ghost Protein Databases

The study was carried out using HaltORF database named "HS_GRCh38_altorf_20170421". This database is derived from the predicted *H. Sapiens* alternative proteins (release hg38, Assembly: GCF_000001405.26) which contains 182,709 entries. This database is a computer compilation of all putative proteins from noncoding regions of mRNA and ncRNA. Additional online databases such as "Ensembl" (https://www.ensembl.org) and "ref Seq" (https://www.ncbi.nlm.nih.gov/refseq) were also used to trace back the origin of the identified AltProts after HaltORF data interrogation. The AltProts originate from either

the 5 'and 3' UTR parts or from +2 or +3 reading frame shifts in the CDS of mature RNA; not following the Kozak frame despite the presence of a START and STOP codon. The HaltORF database was used in combination with the conventional RefProts database obtained from "UNIPROT".

2.2. Cell culture

The cells used in the analysis are derived from the HeLa line (ATCC). To summarizethe protocol described in the publication by F. Liu and al. ("*Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry*") [44], the cells are cultured in modified Dulbecco's Eagle environment with 10% fetal calf serum and 1% penicillin-streptomycin up to 80% confluence. The cells were then harvested by trypsinization and washed three times with PBS. After separation in the lysis membrane buffer and centrifugation, only the remaining nuclei were kept for the XL-MS. This nuclei fraction was then cross-linked with DSSO (1mM) with a 100 fold excess of cross-linker with respect to the protein quantity. The cross-linked proteins were then reduced, alkylated and digested by Lys-C/Trypsin mixture. The resulting cross-linked peptides were desalted on a Sep-Pak C18, dried and further enriched using SCX as previously described [44].

2.3. Cross-link Workflow

Data were extracted from the Chorus data repository (https://chorusproject.org) project I.D. number 890 and re-analyzed using Proteome Discoverer 2.2 (PD2.2) with the XLinkX node [45]. Interrogation of data was performed accordingly to the following workflow: first spectra were selected and DSSO was defined as cross-linker (characteristic mass 158.003765 Da). Then the workflow was divided into two paths. The first is dedicated to the cross-link identifications using the XLinkX Search as parameters Precursor Mass Tolerance: 10 ppm, FTMS fragment: 20 ppm, ITMS Fragment: 0.5 Da, search (database compiled AltProt + RefProt) and the validation was performed using percolator with a FDR set to 0.01. The second path is the total protein identification using SequestHT considering the following parameters: Trypsin

as enzyme, 2 missed cleavages, methionine oxidation as variable modification, DSSO hydrolyzed and carbamidomethylation of cysteins as static modification, Precursor Mass Tolerance: 10 ppm and Fragment mass tolerance: 0.6 Da. The validation was performed using Percolator with a FDR set to 0.01. A consensus workflow was then applied for the statistical arrangement. A de-isotope and TopX filter were used to determine the m/z-error with a selectivity around 10% FDR. The protein-protein interaction identifiers were displays in the xiNET software (http://crosslinkviewer.org) [46] and Cytoscape3.7.1 allowing for visualization of the partners and the number of recurrences of the same interaction.

2.4. Modeling and prediction of interactions:

Structure modeling of Ghost Proteins (AltProts) and Reference Proteins (RefProts), were performed with the I-Tasser software [47] when protein structures were not available on Protein Data Bank (PDB) [48]. For both RefProts and AltProts the most stable models (C-Score between -5 and +2) were retained. Within the set of best predictions, only models which are in line with the distances expected for the DSSO cross-linker were considered and further examined. The prediction of protein-protein interactions were performed with the ClusPro software [49]. The RefProt was identified as a receiver and the AltProt as a ligand. The interaction model was carried out by docking the ligand on the receiver without cross-link restriction. ClusPro then generates multiple interaction models ranked in the order of stability. The selected models are still part of the Top5 "balanced" models taking into account the best compromise of stability. The selected interactions were then recreated with Chimera [50] to measure the distance between the atoms observed during the cross-link. The model is split between the ligand and the receptor to form two independent chains, the lysines found to be involved in interactions on PD2.2 and xiNET were then designated in order to identify the distance between the two points of the model. For example, the AltProt AltATAD2 model was generated from its amino acid sequence since it was never previously described the structural data could not be predicted by sequence homology and nature of these amino acids. The model was thus generated by I-Tasser with a C-Score of -3.66 in accordance with recommendations [-5; 2]. It was

observed that, AltATAD2 has a secondary structure composed of 4 alpha helices generating a tubular tertiary structure. Similarly, the AUF1 reference protein had no experimental model and needed to be carried out on I-Tasser. The generated model has a C-Score of -2.81 in agreement with the recommendations [-5; 2]. The second RefProt in interaction with AltATAD2, RPL10, has a public model which was obtained on PDB (reference number: 5aj0) [51]. The structure of RPL10 was performed by cryo electron microscopy. RPL10 is found in interaction with several ribosome proteins, forming the 60S subunit. In this model we also found the presence of several messengers and ribosomal RNAs. Thus, from this model, RPL10 could be isolated in order to generate the AltATAD2-RPL10 interaction. However, once this interaction has been obtained, the entire 60S ribosome model is used to correlate the position of AltATAD2 and to hypothesize the function.

3. RESULTS

3.1. Ghost Proteins revealed in nuclei of HeLas cells by XL-MS

Reprocessing of the PPIs from the nuclei of HeLas cells by XL-MS, revealed 1679 cross-link interactions (**Supp. Data 1**). Each of these interaction was determined with a minimum score of 20 and a cross-link workflow with FDR of 0.01, limiting the number of false positives. Among these 1679 cross-link interactions, 292 were found to involve Ghost Proteins (**Supp. Data 1, colored Ghost Proteins**) including 4 Ghost-Ghost proteins interactions (**Table 1**). In order to get a visual interpretation, the protein networks were generated under xiNET. To ease the data mining, it was possible to separate the interactions of two, three or more partners. Our interest is to focus on networks involving more than three partners thus facilitating the understanding of the involved signaling pathways. One of the most important network identified was highlighted, which represents the observed interactions between ribonucleoproteins, ribosome subunits, zinc finger proteins, in which RefProts and AltProts interact each other's (**Figure *1***). 44 Ghost Proteins in interaction with 7 ribonucleoproteins are observed in this specific network, reflecting the importance of Ghost Proteins in such interactions. Each Ghost Protein is identified by an "IP_" accession number and can be correlated to its transcript number and its associated gene (**Table 1).** This type of

annotation facilitates the identification of the RefProts associated with the mRNA presenting the translated AltProts. We were specifically interested in the networks where Ghost proteins interact with at least two RefProts. Among these, the ghost protein AltATAD2 was found to be in interaction with the RE/poly(U)-binding/degradation factor 1 (AUF1) and the Ribosomal protein 10 (RPL10).

3.2. Comparison of the identified networks for RefProts versus RefProts/AltProts.

To assess the influence of the AltProts on the identified cross-links and the protein networks, the identified interactions were compared with the interrogation of the RefProt database alone and with the combined RefProt/AltProt databases (**Figure 2**). This comparison shows that a large part of identified interactions are found both after using RefProts database alone and using the combined RefProts/AltProts database (yellow) and correspond to RefProts. It is also observed that a large number of protein interactions are added when the AltProt database is considered which is expected since the AltProt database is larger in size than the RefProt one (green). Finally, a non-negligible portion of RefProts that were identified with the RefProt database alone are not observed anymore when using the combination of the two databases (red). From these data, two main features are derived. The first is that in few cases, proteins initially identified as RefProts become attributed to AltProts by combination of the two databases. The second is that some of the RefProts identified are no longer observed with the combined database interrogation (**Figure 3A**). This highlight two important issues. One, is that somehow the current bioinformatics tools seems not to be well-suited to such large databases as the combination of RefProt/AltProt. Indeed the AltProt database has 182,709 entries when the RefProt is only 42,335 entries. In that situation, some of the RefProts fail to pass the FDR threshold. The second is that because some RefProts and AltProts can share a part of their amino acid sequences making proper identification of one or the other difficult. Indeed, if the peptides considered for the identification are only in the common region to the two proteins, and because the AltProt sequences are much smaller by comparison to the RefProts one, the identification weight in favour of the AltProts due to better sequence coverage. The representation of the number of interaction identified per score range

(**Figure 3B**) shows that interaction that were identified with both databases (RefProt/AltProt) are more confident that those identified only with one of the database (RefProt). These not surprisingly correspond to the proteins that are involved in larger network (**Figure 3A**) and identified with a larger number of peptides and interaction. The others (only identified in one interaction) present a relative similar score range. This correspond to proteins identified by only a single interaction. However, in general (**Figure 3**) the addition of the AltProt database bring a lot new information to the picture. To assess the veracity of the identify interactions, we have extracted some MS/MS spectra corresponding to the network involving the AltATAD2 protein. **Figure 4** provides examples of MS/MS spectra for two different interactions. For each interaction the CID and the ETD spectra are displayed with the proteins ID, the amino acid sequences and the cross-link sites. More MS/MS spectra can be found in the **Supp. Data 2**. The first interaction presented (**Figure 4A**) is an interaction between an AltProt and a RefProt which was identified with a score of 40.04. The CID spectrum mainly provides the exact mass of the two peptide chains after the CID cleavage of the DSSO. The annotation of the ETD spectrum show that both cleavages in the two peptide chains are observed and enable confident attribution of the cross-link site. The second example (**Figure 4B-C**) presents a case for which an interaction of the Q14103-4 (HNRNPD) RefProt is truly identified but the identification fails to provide the interacting partner with confidence. Indeed, this protein is found to interact with either an AltProt (IP_128579.1) (**Figure 4B**) or a RefProt (Q8TF62 i.e. ATP8B4) (**Figure 4C**) with scores passing the threshold (>20) on the two cases. Again, CID spectra provide the exact mass of the 2 peptide chains after CID cleavage of the cross-linker. The careful examination of the ETD spectra show that only 2-3 fragmentations (only 1 for the AltProt) are observed for the peptide chain which is not confidently identified. Despite the two proteins have no sequence homology (**Figure 4D**) the peptide MFMVDTKR (Mw_{mono}=1026.50 Da) of the AltProt with an oxydation of Methionine (+16) has the same molecular weight as the DLDDKYFK peptide (Mw_{mono}=1042.50 Da) of the RefProt. In that case, because no specific fragments are found by ETD on that peptide chain the interacting peptide is only identified by its exact mass. Since the AltProt sequence is much shorter than the RefProt, this positively weight on the identification score in favor of the AltProt (33.81) and lead to its preferential

identification. Except for these rare cases, most interaction were found to be trustworthy. For example, **Figure 5** presents the MS^2 spectra for the AltATAD2-RPL10 interaction. Here, the presence of fragments in the two peptides chains give better reliability to the identification.

3.3. AltATAD2 Partners

AltATAD2 is found in the CDS with a +2 ORF frame shift and presents a sequence of 139 amino acid residues for a theoretical molecular weight of 17,077 Da (**Figure 6**). The structure of this Ghost Protein, was predicted by I-Tasser, based on its amino acid sequence (**Figure 7**). The model with the best C-score was retained and used when performing docking by ClusPro2.0 (**Figure 7**). AltATAD2 is observed to interact with ARE/poly(U)-binding/degradation factor 1 (AUF1) and Ribosomal proteins (RPL10), two RefProts described in the literature to be involved in different signaling pathways. Docking was carried out between the AltATAD2-RPL10 and AltATAD2-AUF1 proteins, the Ghost Protein being always designated as the ligand of the refprot due to their size difference. For the refprot RPL10 and AUF1 the models were known from previous experiments thanks to structural studies and were retrieved from PDB. The *in-silico* interaction between AltATAD2 and RPL10 mainly shows, two binding sites for AltATAD2 on RPL10 (**Figure 7A**). The first binding sites is in the cavity of RPL10 and the second one at the periphery as part of the top 5 best electrostatic structures. These two models were chosen in the best generated models but also taking into account the molecular distance derived from the XL-MS using the DSSO cross linker which is <50Å. Similarly, the interaction between AltATAD2 and AUF1 gave two possible interaction sites between the partners *i.e.* one with the best electrostatic characteristics and the second with the best hydrophobic parameters and considering the distance XL-MS imposed by the cross-linker (**Figure 7B**). Finally, AltATAD2 is observed in interaction with these two refProts by fixing different regions. When assembling the docking of AtlATAD2-AUF19RPL10, AltATAD2-RPL10 and AltATAD2-AUF1 by "Match Making" of Chimera, the simultaneous fixation of AltATAD2 and AUF1/RPL10 was found to be feasible (**Figure 7C**) resulting in a possible hetero-dimer biological active complex.

4. DISCUSSION

AltORFs were shown to lead to the translation of AltProts as demonstrated by their observation in the large scale proteomics data [30–32] when using appropriate databases. Very interestingly, the AltProts are also evidenced in the large scale XL-MS data and are found to be interacting with their RefProts counterparts. Observing the AltProts in their interacting network is definitely an approach to get closer to the function of these proteins. Indeed, large scale approach such as XL-MS will provide a global picture for many of these novel proteins without the requirement of developing antibody for each of these proteins as required by the antibody-based strategies. The PPIs highlighted for AltATAD2 is a good example. We have describe an interaction of AltATAD2 with both RPL10 and AUF1XL-MS. AUF1 is a heterogeneous nuclear ribonucleoprotein D (hnRNPD) which was among the first identified ARE-specific binding proteins (AUBPs) [52]. The AUBPs are complexes of proteins which are involved in the regulation of the AU-rich element (ARE) containing mRNAs. One of the limitations of the experiment here is the possible correlation between a found interaction and the time at which this interaction takes place. Here, XL-MS exhibits a global picture of the protein interaction network in the cell, not enabling to determine when an interaction occurs. As a result, the graphical representation obtained on xiNET gives a common interaction between the three proteins but fails to clarify if they are all together interacting at the same time. To access this information, one would need to phase the cells and performed XL-MS time course analyses. Therefore, several interpretations to this trimer interaction can be advanced. The first one is related to an independent interaction between AltATAD2-RPL10 and AltATAD2-AUF1. The AltATAD2-RPL10 interaction observed by XL-MS using DSSO is confirmed by 3D protein modeling and docking of AltATAD2 on RPL10. RPL10 structure was extracted via the online public model on PDB: 5aj0 from the study of Behrmann E. et al [51]. The docking performed on ClusPro highlights several possible fixation sites of AltATAD2 on RPL10; however, only two of them are redundant and in line with the distance limits imposed by the DSSO cross-linker. The first model attaches AltATAD2 at the periphery of RPL10, far from the region fixing the 5S ribosomal RNA. However, it has been shown that RPL10, by its external location on the ribosome, allows the grouping of the subunits

and the formation of an active ribosome. Moreover, its interaction with the 60S ribosomal export protein NMD3 would also be responsible for the migration of the peri-ribosome from the nucleus to the cytoplasm [53]. Thus, in this case the RPL10 interaction with AltATAD2 can be directly involved in this peri-ribosome migration. The second model locates AltATAD2 within the ribosome, and more precisely within the region of RPL10 interacting with the 5S ribosomal RNA. In that case, the protein could be involved in the regulation of the binding of the 5S ribosomal RNA (**Figure 8**). A previous study by cryoelectron microscopy (cryoEM) has demonstrated that the interaction of RPL10 participates in the ribosome constitution, integrating the proteins RPL5 and RPL11. However, it was shown that RPL10 was not essential for the ribosome formation and functionality. The attachment of AltATAD2 on RPL10 could explain RPL10 regulation function by blocking the 5S rRNA binding site. Another hypothesis is the possible cooperation of the interacting partners with the formation of a co-interaction between RPL10, AltATAD2 and AUF1. In this scenario, the interaction of AUF1 and RPL10 is not without consequence. Indeed AUF1, was previously described to have a dual function. It is an initiator of the mRNA degradation but as well a protein fixing to the ARE regions of the 3'UTR. On the other hand, RPL10 role is opposite to AUF1. RPL10 is involved in the ribosome assembly and, thus, in the regulation of the translation of mRNAs into proteins (**Figure 9)**. The co-interaction of AltATAD2 with AUF1 and RPL10 could be the first description of a regulation of the expression of the Ghost proteins resulting from non-coding regions such as the 3'UTR. Since Ghost proteins were not considered before, this could explain why this mechanism was not demonstrated before. Finally, we could hypothesize that the formation of the hetero-trimer, with the attachment of AUF1 on the external part of RPL10, could be involved in a mechanism of regulation of the ribosome. In this case, the RNA5S would fix onto the 60S subunit of the ribosome and activate the transcription. This mechanism would regulate ribosome activation by recruitment of AltATAD2 at the periphery of RPL10 via AUF1 leading to a fine regulation of protein translation. Recently, the so-called Ghost Protein "Nobody", derived from the non-coding RNA: LINC01420/LOC550643, was shown to be involved in the mRNA decapping signaling pathway by interacting with the decapping proteins 4 (EDC4) by multiple techniques including APEX (ascorbate enzymes peroxidase), Photo-cross-link and co-immunoprecipitation [34]. In summary,

this study confirms the involvement of Ghost Proteins in the regulation of mRNA expression. The demonstration of an interaction between AltProts and RefProts is a first clue demonstrating their effective role and function in cells. Ghost Proteins are active compounds actively participating to the cell regulation as the RefProts [30].

The proteomic community must widen its field of view to a world, going against the known Kozak dogma of the expression of the proteins, but existing and influencing the known and described models of today. Demonstrating the interaction of these proteins and their involvement in the signaling pathways within the cells is an important step forwards in understanding their functions. Herein, we demonstrate that the XL-MS non targeted large scale approach is useful in this demonstration by revealing the importance of the Ghost Proteins within the interaction networks of RefProts. Although various developments remain to be performed to improve cell interactomic inclusive to AltProts. As demonstrated here for a few percentage of proteins, the size of the total database used by aggregating the AltProt (182,709 entries) to the RefProt (42,335 entries) database, show some limitations of the actual interrogation tools due to the size of this database. Moreover, due to the important number of sequence in this total database and in the cases where only few fragments are observed for one of the peptide chains there is a not unneglectable probability that different peptides match with the same exact mass. This clearly highlight the importance, for such large scale data using such large database for interrogation, to be more stringent on the identification of the interaction and yet to manually check the MS/MS spectra corresponding to the interaction of interest. In general the rate of false positive interactions remain more elevated in the XL-MS approach and confident identification can only be obtained if strictly respecting specific guidelines as reported by t Iacobucci C. et al [54]. However, despite these few limitations, it is clear that large scale interactomics of AltProt will open the way to more complete systems biology pictures [43,55].

ACKNOWLEDGEMENTS

This research was supported by funding from Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation (MESRI), Institut National de la Santé et de la Recherche Médicale (Inserm) and Université de Lille.

AUTHOR CONTRIBUTIONS

Conceptualization, I.F., J.F. and M.S.; Methodology, I.F., J.F., T.C. and M.S.; Software, T.C.; Validation, I.F., J.F., T.C. and M.S.; Formal Analysis, T.C.; Investigation, I.F., J.F., T.C., and M.S.; Resources, I.F. and M.S.; Data curation, T.C.; Writing - Original Draft T.C. and M.S. Writing - Review & Editing, I.F. and M.S.; Supervision, I.F., J.F. and M.S.; Project Administration, I.F. and M.S.; Funding Acquisition, I.F., and M.S.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

[1] F. Meier, P.E. Geyer, S. Virreira Winter, J. Cox, M. Mann, BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes, Nat. Methods, 15 (2018) 440–448.

[2] A.-C. Gavin, M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J.M. Rick, A.-M. Michon, C.-M. Cruciat, M. Remor, C. Höfert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, et al., Functional organization of the yeast proteome by systematic analysis of protein complexes, Nature, 415 (2002) 141–147.

[3] J.D. Martell, T.J. Deerinck, Y. Sancak, T.L. Poulos, V.K. Mootha, G.E. Sosinsky, M.H. Ellisman, A.Y. Ting, Engineered ascorbate peroxidase as a genetically encoded reporter for electron microscopy, Nat. Biotechnol., 30 (2012) 1143–1148.

[4] S.S. Lam, J.D. Martell, K.J. Kamer, T.J. Deerinck, M.H. Ellisman, V.K. Mootha, A.Y. Ting, Directed evolution of APEX2 for electron microscopy and proximity labeling, Nat. Methods, 12 (2015) 51–54.

[5] K.J. Roux, D.I. Kim, M. Raida, B. Burke, A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells, J. Cell Biol., 196 (2012) 801–10.

[6] D.I. Kim, S.C. Jensen, K.A. Noble, B. KC, K.H. Roux, K. Motamedchaboki, K.J. Roux,

An improved smaller biotin ligase for BioID proximity labeling, Mol. Biol. Cell, 27 (2016) 1188–1196.

[7] E. Coyaud, C. Ranadheera, D. Cheng, J. Gonçalves, B.J.A. Dyakov, E.M.N. Laurent, J. St-Germain, L. Pelletier, A.-C. Gingras, J.H. Brumell, P.K. Kim, D. Safronetz, B. Raught, Global Interactomics Uncovers Extensive Organellar Targeting by Zika Virus, Mol. Cell. Proteomics, 17 (2018) 2242–2255.

[8] S. Eyckerman, K. Titeca, E. Van Quickelberghe, E. Cloots, A. Verhee, N. Samyn, L. De Ceuninck, E. Timmerman, D. De Sutter, S. Lievens, S. Van Calenbergh, K. Gevaert, J. Tavernier, Trapping mammalian protein complexes in viral particles, Nat. Commun., 7 (2016) 11416.

[9] A. Leitner, M. Faini, F. Stengel, R. Aebersold, Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines, (2016).

[10] K. Verheggen, H. Raeder, F.S. Berven, | Lennart Martens, H. Barsnes, M. Vaudel, Anatomy and evolution of database search engines-a central component of mass spectrometry based proteomic workflows, (2017).

[11] UniProt: a hub for protein information, Nucleic Acids Res., 43 (2015) D204–D212.

[12] M. Kozak, Rethinking some mechanisms invoked to explain translational regulation in eukaryotes, Gene, 382 (2006) 1–11.

[13] M. Kozak, Initiation of translation in prokaryotes and eukaryotes, Gene, 234 (1999) 187–208.

[14] M. Kozak, Regulation of Translation in Eukaryotic Systems, Annu. Rev. Cell Biol., 8 (1992) 197–225.

[15] B.L. Aken, P. Achuthan, W. Akanni, M.R. Amode, F. Bernsdorff, J. Bhai, K. Billis, D. Carvalho-Silva, C. Cummins, P. Clapham, L. Gil, C.G. Girón, L. Gordon, T. Hourlier, S.E. Hunt, S.H. Janacek, T. Juettemann, S. Keenan, M.R. Laird, I. Lavidas, et al., Ensembl 2017, Nucleic Acids Res., 45 (2017) D635–D642.

[16] J. Harrow, F. Denoeud, A. Frankish, A. Reymond, C.-K. Chen, J. Chrast, J. Lagarde, J.G. Gilbert, R. Storey, D. Swarbreck, C. Rossier, C. Ucla, T. Hubbard, S.E. Antonarakis, R. Guigo, GENCODE: producing a reference annotation for ENCODE, Genome Biol., 7 (2006) S4.

[17] D. Thierry-Mieg, J. Thierry-Mieg, AceView: a comprehensive cDNA-supported gene and transcripts annotation, Genome Biol., 7 Suppl 1 (2006) S12.1-14.

[18] R. V. Davuluri, Y. Suzuki, S. Sugano, C. Plass, T.H.-M. Huang, The functional consequences of alternative promoter use in mammalian genomes, Trends Genet., 24 (2008) 167–177.

[19] T.W. Nilsen, B.R. Graveley, Expansion of the eukaryotic proteome by alternative splicing, Nature, 463 (2010) 457–63.

[20] D.C. Di Giammartino, K. Nishida, J.L. Manley, Mechanisms and consequences of

alternative polyadenylation, Mol. Cell, 43 (2011) 853–66.

[21] N.M. Wills, J.F. Atkins, The potential role of ribosomal frameshifting in generating aberrant proteins implicated in neurodegenerative diseases, RNA, 12 (2006) 1149–53.

[22] V. Delcourt, M. Brunelle, A. V. Roy, J.-F. Jacques, M. Salzet, I. Fournier, X. Roucou, 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 (2018) 2402–2411.

[23] M.A. Brunet, M. Brunelle, J.-F. Lucier, V. Delcourt, M. Levesque, F. Grenier, S. Samandi, S. Leblanc, J.-D. Aguilar, P. Dufour, J.-F. Jacques, I. Fournier, A. Ouangraoua, M.S. Scott, F.-M. Boisvert, X. Roucou, OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes, Nucleic Acids Res., (2018).

[24] S. Samandi, A. V Roy, V. Delcourt, J.-F. Lucier, J. Gagnon, M.C. Beaudoin, B. Vanderperre, M.-A. Breton, J. Motard, J.-F. Jacques, M. Brunelle, I. Gagnon-Arsenault, I. Fournier, A. Ouangraoua, D.J. Hunting, A.A. Cohen, C.R. Landry, M.S. Scott, X. Roucou, Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins, Elife, 6 (2017) e27860.

[25] H. Mouilleron, V. Delcourt, X. Roucou, Death of a dogma: eukaryotic mRNAs can code for more than one protein, Nucleic Acids Res., 44 (2016) 14–23.

[26] V. Delcourt, A. Staskevicius, M. Salzet, I. Fournier, X. Roucou, Small Proteins Encoded by Unannotated ORFs are Rising Stars of the Proteome, Confirming Shortcomings in Genome Annotations and Current Vision of an mRNA, Proteomics, 18 (2018) e1700058.

[27] B. Vanderperre, J.-F. Lucier, X. Roucou, HAltORF: a database of predicted out-of-frame alternative open reading frames in human, Database (Oxford)., 2012 (2012) bas025.

[28] Y. Hao, L. Zhang, Y. Niu, T. Cai, J. Luo, S. He, B. Zhang, D. Zhang, Y. Qin, F. Yang, R. Chen, SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci, Brief. Bioinform., 19 (2018) 636–643.

[29] E. Le Rhun, M. Duhamel, M. Wisztorski, J.-P. Gimeno, F. Zairi, F. Escande, N. Reyns, F. Kobeissy, C.-A. Maurage, M. Salzet, I. Fournier, C. Henkel, H. Peter, Evaluation of non-supervised MALDI mass spectrometry imaging combined with microproteomics for glioma grade III classification ☆, (2016).

[30] B. Vanderperre, J.-F. Lucier, C. Bissonnette, J. Motard, G. Tremblay, S. Vanderperre, M. Wisztorski, M. Salzet, F.-M. Boisvert, X. Roucou, H. Steen, M. Mann, D. Licatalosi, R. Darnell, R. Davuluri, Y. Suzuki, S. Sugano, C. Plass, T. Huang, T. Nilsen, et al., Direct Detection of Alternative Open Reading Frames Translation Products in Human Significantly Expands the Proteome, PLoS One, 8 (2013) e70698.

[31] V. Delcourt, J. Franck, J. Quanico, J.-P. Gimeno, M. Wisztorski, A. Raffo-Romero,

F. Kobeissy, X. Roucou, M. Salzet, I. Fournier, Spatially-Resolved Top-down Proteomics Bridged to MALDI MS Imaging Reveals the Molecular Physiome of Brain Regions, Mol. Cell. Proteomics, 17 (2018) 357–372.

[32] V. Delcourt, J. Franck, E. Leblanc, F. Narducci, Y.-M. Robin, J.-P. Gimeno, J. Quanico, M. Wisztorski, F. Kobeissy, J.-F. Jacques, X. Roucou, M. Salzet, I. Fournier, Combined Mass Spectrometry Imaging and Top-down Microproteomics Reveals Evidence of a Hidden Proteome in Ovarian Cancer, EBioMedicine, 21 (2017) 55–64.

[33] S.A. Slavoff, J. Heo, B.A. Budnik, L.A. Hanakahi, A. Saghatelian, A human short open reading frame (sORF)-encoded polypeptide that stimulates DNA end joining, J. Biol. Chem., 289 (2014) 10950–7.

[34] N.G. D'Lima, J. Ma, L. Winkler, Q. Chu, K.H. Loh, E.O. Corpuz, B.A. Budnik, J. Lykke-Andersen, A. Saghatelian, S.A. Slavoff, A human microprotein that interacts with the mRNA decapping complex, Nat. Chem. Biol., 13 (2017) 174–180.

[35] C. Lee, J. Zeng, B.G. Drew, T. Sallam, A. Martin-Montalvo, J. Wan, S.-J. Kim, H. Mehta, A.L. Hevener, R. de Cabo, P. Cohen, The Mitochondrial-Derived Peptide MOTS-c Promotes Metabolic Homeostasis and Reduces Obesity and Insulin Resistance, Cell Metab., 21 (2015) 443–454.

[36] A. Matsumoto, A. Pasut, M. Matsumoto, R. Yamashita, J. Fung, E. Monteleone, A. Saghatelian, K.I. Nakayama, J.G. Clohessy, P.P. Pandolfi, mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide, Nature, 541 (2017) 228–232.

[37] D.M. Anderson, K.M. Anderson, C.-L. Chang, C.A. Makarewich, B.R. Nelson, J.R. McAnally, P. Kasaragod, J.M. Shelton, J. Liou, R. Bassel-Duby, E.N. Olson, A micropeptide encoded by a putative long noncoding RNA regulates muscle performance, Cell, 160 (2015) 595–606.

[38] P. Bi, A. Ramirez-Martinez, H. Li, J. Cannavino, J.R. McAnally, J.M. Shelton, E. Sánchez-Ortiz, R. Bassel-Duby, E.N. Olson, Control of muscle formation by the fusogenic micropeptide myomixer, Science (80-.)., 356 (2017) 323–327.

[39] P. Yuan, N.G. D'Lima, S.A. Slavoff, Comparative Membrane Proteomics Reveals a Nonannotated E coli Heat Shock Protein, Biochemistry, 57 (2018) 56–60.

[40] N.G. D'Lima, A. Khitun, A.D. Rosenbloom, P. Yuan, B.M. Gassaway, K.W. Barber, J. Rinehart, S.A. Slavoff, Comparative Proteomics Enables Identification of Nonannotated Cold Shock Proteins in *E coli*, J. Proteome Res., 16 (2017) 3722–3731.

[41] E. Le Rhun, M. Duhamel, M. Wisztorski, F. Zairi, C.A. Maurage, I. Fournier, N. Reyns, M. Salzet, METB-07CLASSIFICATION OF HIGH GRADE GLIOMA USING MATRIX-ASSISTED LASER DESORPTION/IONIZATION MASS SPECTROMETRY IMAGING (MALDI MSI): INTERIM RESULTS OF THE GLIOMIC STUDY, Neuro. Oncol., 17 (2015) v136.3-v136.

[42] Z. Ning, B. Hawley, C.-K. Chiang, D. Seebun, D. Figeys, Detecting Protein–Protein Interactions/Complex Components Using Mass Spectrometry Coupled Techniques, in: Humana Press, New York, NY, 2014: pp. 1–13.

[43] O. Klykov, B. Steigenberger, S. Pektaş, D. Fasci, A.J.R. Heck, R.A. Scheltema, Efficient and robust proteome-wide approaches for cross-linking mass spectrometry, Nat. Protoc., (n.d.).

[44] F. Liu, D.T.S. Rijkers, H. Post, A.J.R. Heck, Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry, Nat. Methods, 12 (2015) 1179–1184.

[45] H. Li, B. Lei, W. Xiang, H. Wang, W. Feng, Y. Liu, S. Qi, Differences in Protein Expression between the U251 and U87 Cell Lines, Turk. Neurosurg., 27 (2017) 894–903.

[46] C.W. Combe, L. Fischer, J. Rappsilber, xiNET: cross-link network maps with residue resolution, Mol. Cell. Proteomics, 14 (2015) 1137–47.

[47] Y. Zhang, I-TASSER server for protein 3D structure prediction, BMC Bioinformatics, 9 (2008) 40.

[48] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, Nucleic Acids Res., 28 (2000) 235–242.

[49] S.R. Comeau, D.W. Gatchell, S. Vajda, C.J. Camacho, ClusPro: a fully automated algorithm for protein-protein docking, Nucleic Acids Res., 32 (2004) W96–W99.

[50] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, T.E. Ferrin, UCSF Chimera?A visualization system for exploratory research and analysis, J. Comput. Chem., 25 (2004) 1605–1612.

[51] E. Behrmann, J. Loerke, T. V Budkevich, K. Yamamoto, A. Schmidt, P.A. Penczek, M.R. Vos, J. Bürger, T. Mielke, P. Scheerer, C.M.T. Spahn, Structural snapshots of actively translating human ribosomes, Cell, 161 (2015) 845–57.

[52] G. Brewer, An A+U-Rich Element RNA-Binding Factor Regulates c-myc mRNA Stability In Vitro, 1991.

[53] A.W. Johnson, E. Lund, J. Dahlberg, Nuclear export of ribosomal subunits, Trends Biochem. Sci., 27 (2002) 580–5.

[54] C. Iacobucci, A. Sinz, To Be or Not to Be? Five Guidelines to Avoid Misassignments in Cross-Linking/Mass Spectrometry, Anal. Chem., 89 (2017) 7832–7835.

[55] F. Liu, P. Lössl, R. Scheltema, R. Viner, A.J.R. Heck, Optimized fragmentation schemes and data analysis strategies for proteome-wide cross-link identification, Nat. Commun., 8 (2017).

TABLE AND FIGURE LEGENDS

Table 1: Identification of inter-cross-links Ghost Proteins, with a maximum identification score of 50.91 and a minimum of 26.54 these identifications are found among the RefProt-RefProt / Ghost Proteins interactions.

Table 2: List of AltProt-RefProt interactions identify in the network (color code is the same as in Figure1). For each AltProt the transcript number and gene name from Ensembl database associated with the RNA is given. Each interaction observed in the subdivisions of Figure 1 is identified.

Figure 1: Interaction network obtained from the XL-MS experiments using a RefProt/AltProt database for data interrogation issued from the combination of the RefProt (Uniprot) and the AltProt (HaltProt) databases. 44 AltProts are found in the network to be interacting with 10 ribonucleoproteins, 3 zinc finger proteins and 2 ribosomal proteins. The network is subdivided into 7 fractions allowing the annotation of AltProts and RefProts in interactions (see Table 2).

Figure 2: Cytoscape description of the interaction map obtained from XL-MS data. Data analysis comparison for the RefProt or the combined RefProts/AltProts databases using DyNet apps. In green are the nodes and edges found using the combined RefProts/AltProts databases, in red the identification specific to the RefProt database and in yellow identifications obtained with both the RefProt and the combined RefProts/AltProts databases

Figure 3: Identified interaction obtained from XL-MS data by data interrogation with the RefProt database alone or the combination of the RefProts/AltProts databases **(A)** Global mapping of all interactions. Red indicates interactions identified with the RefProt database alone, green with the combined RefProt/AltProt alone and yellow with both RefProt and RefProt/AltProt Databases **(B)** Distribution of the identification scores for each cross-link as a function of the number of identified interaction.

Figure 4: MS/MS spectra (CID and ETD) with their annotation for identified interaction between RefProt and AltProt. CID/ETD MS^2 spectra of the identified interaction of **(A)** the RefProt Q14103-4 (HNRNPD) with the AltProt IP_297459.1, **(B)** the RefProt Q14103-4 (HNRNPD) with the AltProt IP_128579 and **(C)** the RefProt Q14103-4 (HNRNPD) with the RefProt Q8TF62 (ATP8B4). **(D)** Sequences alignment of the RefProt Q14103-4 (HNRNPD) and the AltProt IP_128579 showing that the 2 proteins do not share sequence homology.

Figure 5: CID/ETD MS^2 spectra and their annotation of the identified interaction between AltATAD2 and this interacted protein RPL10.

Figure 6: Focus on AltATAD2 protein **(A)** In the interactome network, AltATAD2 is observed to be in interaction with 2 different partners **(B)** AltATAD2 amino acid sequence and schematic representation of its sequence location in the mature RNA. AltATAD2 is found in the CDS with a +2 ORF shift **(C)** Nucleic acid sequence encoding AltATAD2 within the ATAD2 mRNA.

Figure 7: 3D modeling of the interactions between AltATAD2 and the RefProts AUF1 or RPL10. **(A)** Models predicted by ClusPro2.0 for the AltATAD2/RPL10 interaction. These two models are part of the TOP5 predictions and are in agreement with the distance restrictions imposed by the XL-MS. Surface modeling was also performed to manually control the likelihood of the result **(B)** Predicted models and 3D surface presentations for the AltATAD2/AUF1 interaction selecting predictions with the highest scores in good agreement with XL-MS **(C)** 3D model of the co-interaction between AUF1-AltATAD2-RPL10. 3D modeling was used to check that AUF1 and RPL10 are not confused in space.

Figure 8: Implementation of AltATAD2 on the 3D modeling of RPL10 and ribosome 60S obtained by cryoEM **(A)** AltATAD2 is found to be interacting at the periphery of RPL10,

thus meeting no other subunit of the 60S ribosome or 5S rRNA **(B)** On the second position AltATAD2 is observed in the space used by the 5S rRNA. However AltATAD2 does not merge with the position of other subunit of the ribosome 60S. This confirms the ability of AltATAD2 to get into this position.

Figure 9: Schematic representation of the different hypothesized configurations for the co-interaction of RPL10-AltATAD2 and AUF1. All these steps could sequentially exist at different time point to regulate the transcription and the translation. **(A)** AltATAD2 in internal position on RPL10 prevents the binding of the ribosomal RNA5S. A decrease in binding of RNA5S on the 60S subunit of the ribosome leads to a decrease in the protein translation. **(B)** AltATAD2 at the outer position on RPL10 allows the formation of the RPL10-AltATAD2-AUF1 complex. In this configuration the RNA5S can fix onto the 60S subunit of the ribosome and activate the transcription. This mechanism would regulate ribosome activation by recruitment of AltATAD2 at the periphery of RPL10 by AUF1 leading to a fine regulation of protein translation. **(C)** Lastly the interaction of the RPL10-AltATAD2-AUF1 complex takes place in the 3'UTR region and leads to the recruitment of the sub-unit 60S at the ARE to activate the translation of AltProts present in this region.

Supplementary data 1: Reprocessing of the protein-protein interactions in the nuclei of HeLas cells by XL-MS-MS method, revealed 1679 cross-link interactions .Each of these interactions determined with a minimum score of 20 and a cross-link workflow with FDR of 0.01, limiting the number of false positives. Among these 1679 cross-link interactions, 292 involved Ghost Proteins (in green)

Supplementary data 2: Examples of CID/ETD MS^2 spectra and their annotation of various identified interaction between either RefProts and AltProts or RefProts and RefProts. AltProts interaction identified are not found to be less confident than RefProts ones.

Score	Protein1	PepPos1	PepSeq1	LinkPos1	Protein2	PepPos2	PepSeq2	LinkPos2
50.91	IP_243260.1	5	APRPGNWKWQRR	8	IP_202369.1	28	VGNKSR	4
28.74	IP_222735.1	61	RENKVCVSTWQK	4	IP_093889.1	46	QRAKS	4
28.73	IP_145224.1	2	TIKTKHMIK	5	IP_177042.1	8	LTSRKR	5
26.54	IP_210743.1	24	GGLKTSRDSR	4	IP_138860.1	1	MPATDGKCK	7

Table 1

Ribosomal protein

Ribonucleoprotein

Other RefProt

1				TR	GN	Gene Description	Gene Name
RefProt		RPL10				60S ribosomal protein L10	
		IP_118801.1		NM_001144756.1	10886	neuropeptide FF receptor 2	NPFFR2
		IP_166911.1		NM_014109.3	29028	ATPase family, AAA domain containing 2	ATAD2
2				TR	GN	gene name	
RefProt		HNRNPAB				heterogeneous nuclear ribonucleoprotein A/B	
		IP_066105.1		NM_001002912.4	127254	glutamate rich 3	ERICH3
		IP_135883.1		XR_427728.1	102724275		
		IP_134928.1		NM_014594.1	30832	Zinc finger protein 354C	ZNF354C
		IP_263009.1		NR_028337.1	400624	long intergenic non-protein coding RNA 1973	LINC01973
		IP_257296.1		NM_001160423.1	10642	insulin like growth factor 2 mRNA binding protein 1	IGF2BP1
3				TR	GN	gene name	
RefProt		RBMX				RNA binding motif protein X-linked	
		IP_094161.1		NM_001204.6	659	bone morphogenetic protein receptor type 2	BMPR2
		IP_249315.1		NM_018146.2	55178	RNA methyltransferase like 1	RNMTL1
RefProt		HNRNPA0				heterogeneous nuclear ribonucleoprotein A0	
		IP_303885.1		NM_001163280.1	27336	HIV-1 Tat specific factor 1	HTATSF1
4				TR	GN	gene name	
RefProt		HNRNPA1				heterogeneous nuclear ribonucleoprotein A1	
	HNRNPA2B1					heterogeneous nuclear ribonucleoprotein A2/B1	
		IP_210711.1		NM_022658.3	3224	homeobox C8	HOXC8
		IP_226921.1		NM_001281734.1	53349	zinc finger FYVE-type containing 1	ZFYVE1
		IP_146439.1		NM_001002255.1	387082	small ubiquitin-like modifier 4	SUMO4
		IP_086141.1		NM_032208.2	84168	ANTXR cell adhesion molecule 1	ANTXR1
	IP_124600.1			NR_034075.1	100499177	THAP9 antisense RNA 1	THAP9-AS1
	IP_214370.1			NM_001286262.1	255394	t-complex 11 like 2	TCP11L2
	IP_118616.1			NM_214711.3	401137	proline rich 27	PRR27
	IP_250819.1			NM_198154.1	339168	Transmembrane protein 95	TMEM95
	IP_155900.1			NM_019042.3	54517	Pseudouridylate synthase 7 homolog	PUS7
	IP_174099.1			NM_014290.2	23424	tudor domain containing 7	TDRD7
	IP_097241.1			NM_022817.2	8864	period circadian regulator 2	PER2
RefProt		HNRNPD				heterogeneous nuclear ribonucleoprotein D	
		IP_150105.1		NM_000535.5	5395	postmeiotic segregation increased 2	PMS2
		IP_297459.1		NM_000381.3	4281	midline 1	MID1
		IP_128579.1		NM_000046.3	411	Arylsulfatase B	ARSB
5				TR	GN	gene name	
RefProt		HNRNPC				heterogeneous nuclear ribonucleoprotein C (C1/C2)	
	VDAC2					voltage dependent anion channel 2	
			HNRNPM			heterogeneous nuclear ribonucleoprotein M	
		IP_233089.1		NM_001194998.1	22995	centrosomal protein 152	CEP152
		IP_270709.1		NM_003437.3	7695	Zinc finger protein 136	ZNF136
		IP_297440.1		NM_001256944.1	1183	chloride voltage-gated channel 4	CLCN4
		IP_154990.1		NM_001287054.1	7586	zinc finger with KRAB and SCAN domains 1	ZKSCAN1
		IP_113552.1		NR_103821.1	442075	EMC3 antisense RNA 1	EMC3-AS1
		IP_062261.1		NM_032884.4	84970	chromosome 1 open reading frame 94	C1orf94
		IP_279198.1		NM_052925.2	114823	Leukocyte receptor cluster (LRC) member 8	LENG8
		IP_233136.1		NM_001193489.1	9728	SECIS binding protein 2 like	SECISBP2L
	IP_092512.1			NM_003659.3	8540	alkylglycerone phosphate synthase	AGPS
		IP_217585.1		XR_429051.1	102724196		
			IP_257543.1	XM_006721752.1	201191		
			IP_086598.1	NM_015470.2	26056	RAB11 family interacting protein 5	RAB11FIP5
RefProt		ANXA2				annexin A2	
		IP_085590.1		NM_001122964.2	57223	protein phosphatase 4 regulatory subunit 3B	PPP4R3B
		IP_078517.1		NM_001024226.1	375	ADP ribosylation factor 1	ARF1
6				TR	GN	gene name	
RefProt		AHNAK				AHNAK nucleoprotein	
		IP_123046.1		NM_001166373.1	55016	membrane associated ring-CH-type finger 1	MARCH1
		IP_238136.1		NR_026647.1	791115	Prader-Willi region non-protein coding RNA 2	PWRN2
		IP_094777.1		NM_001039538.1	4133	microtubule associated protein 2	MAP2
		IP_074702.1		NM_000721.3	777	calcium voltage-gated channel subunit alpha1 E	CACNA1E
		IP_124534.1		NR_046377.1	728040	long intergenic non-protein coding RNA 2499	LINC02499
7				TR	GN	gene name	
RefProt		HIST1H1C				histone cluster 1 H1 family member c	
		IP_062513.1		NM_012199.2	26523	Argonaute RISC catalytic component 1	AGO1

Table 2

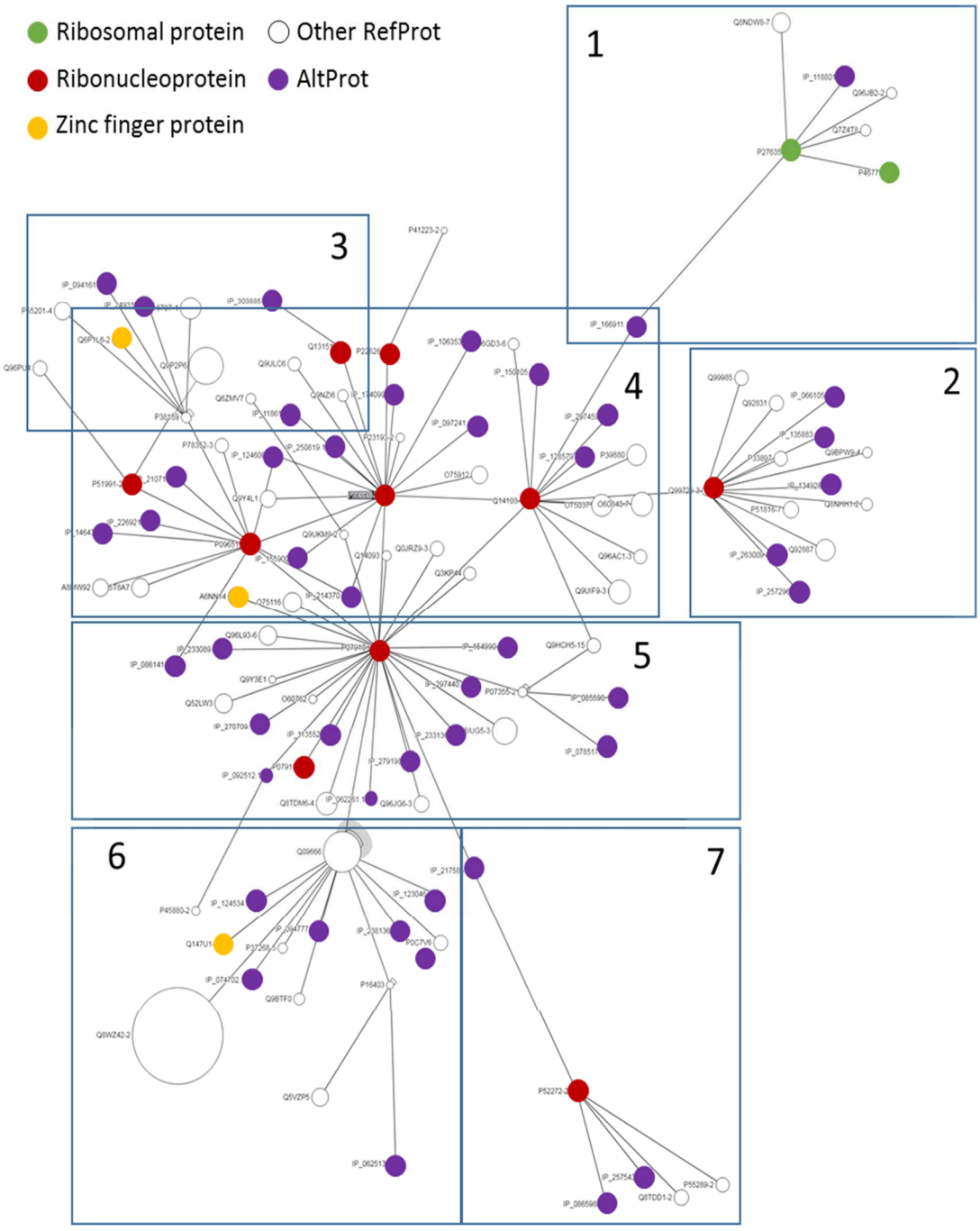
Ribosomal protein
Other RefProt
Ribonucleoprotein
AltProt
Zinc finger protein
1
2
3
4
5
6
7

Figure 1

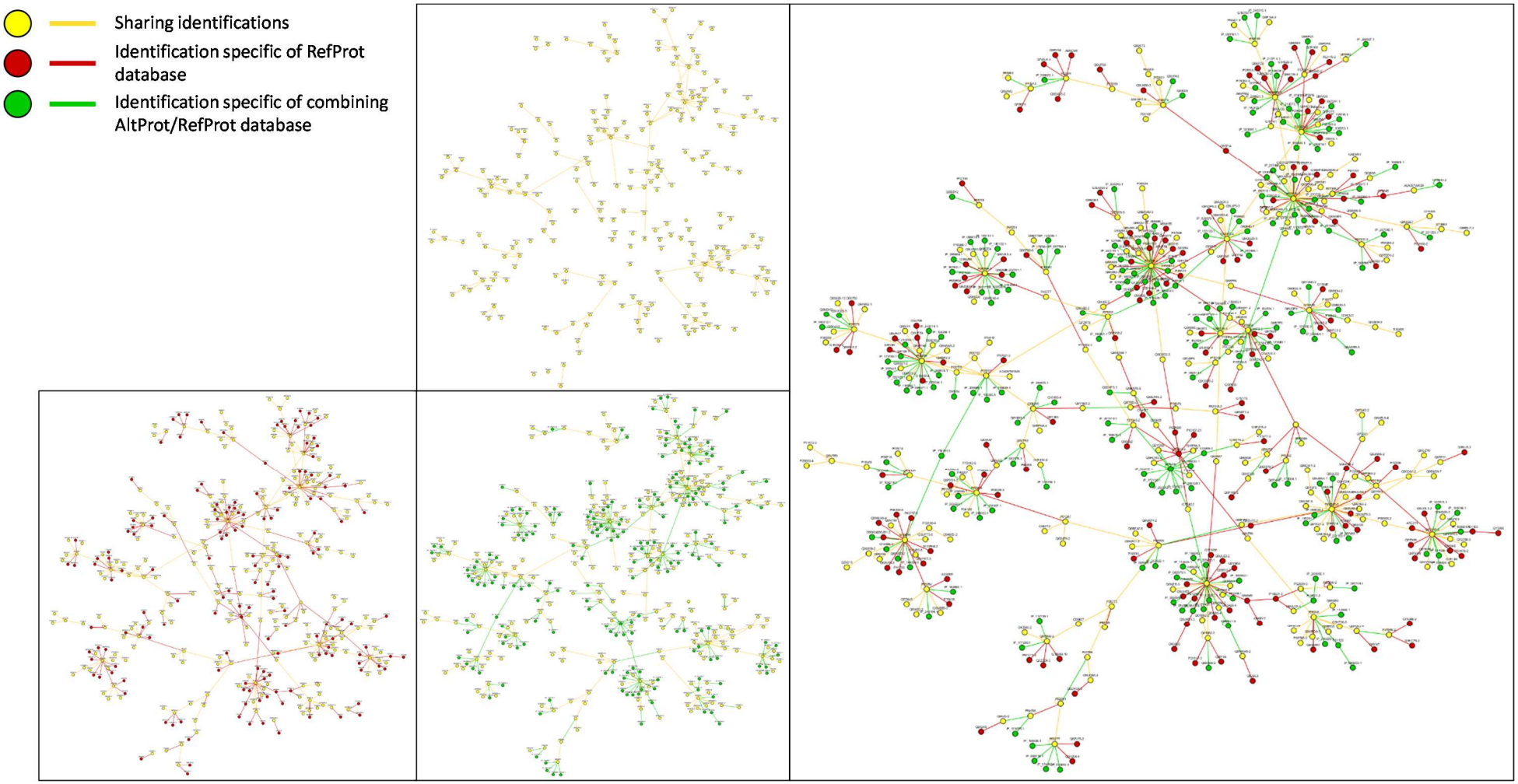
Sharing identifications
Identification specific of RefProt database
Identification specific of combining AltProt/RefProt database

Figure 2


A
B
140,00
120,00
100,00
80,00
60,00
40,00
20,00
0,00
number of interaction
20-30
30-40
40-50
50-60
60-70
70-80
80-90
90-100
100-110
110-120
120-130
130-140
140-150
150-160
160-170
170-185
Scores

Figure 3

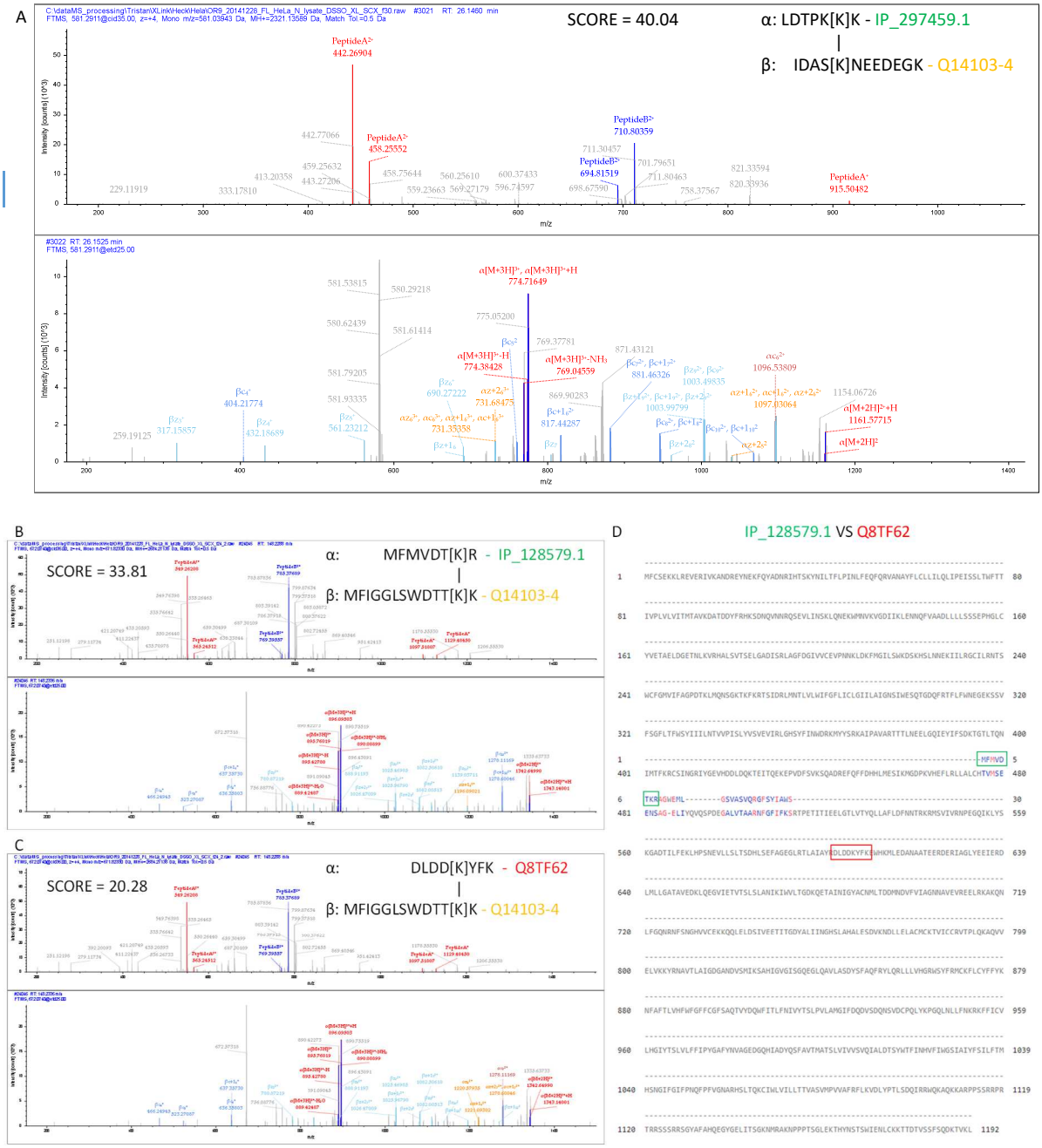
A
SCORE = 40.04
α: LDTPK[K]K - IP_297459.1
β: IDAS[K]NEEDEGK - Q14103-4
B
SCORE = 33.81
α: MFMVDT[K]R - IP_128579.1
β: MFIGGLSWDTT[K]K - Q14103-4
C
SCORE = 20.28
α: DLDD[K]YFK - Q8TF62
β: MFIGGLSWDTT[K]K - Q14103-4
D
IP_128579.1 VS Q8TF62

Figure 4

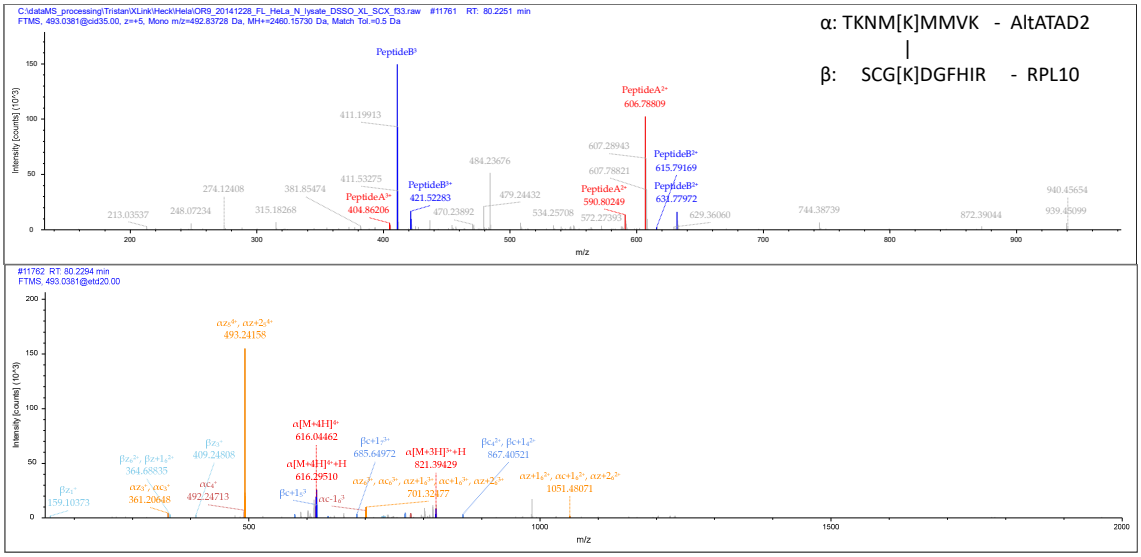
α: TKNM[K]MMVK - AltATAD2
β: SCG[K]DGFHIR - RPL10
PeptideB³
PeptideA²⁺
606.78809
PeptideB²⁺
615.79169
PeptideB²⁺
631.77972
PeptideA²⁺
590.80249
PeptideB³⁺
421.52283
PeptideA³⁺
404.86206
411.19913
411.53275
484.23676
607.28943
607.78821
479.24432
470.23892
534.25708
572.27393
629.36060
744.38739
872.39044
940.45654
939.45099
274.12408
381.85474
315.18268
248.07234
213.03537
Intensity [counts] (10^3)
m/z
#11762 RT: 80.2294 min
FTMS, 493.0381@etd20.00
493.24158
616.04462
616.29510
685.64972
821.39429
701.32477
867.40521
1051.48071
409.24808
492.24713
361.20648
364.68835
159.10373
Intensity [counts] (10^3)
m/z

Figure 5

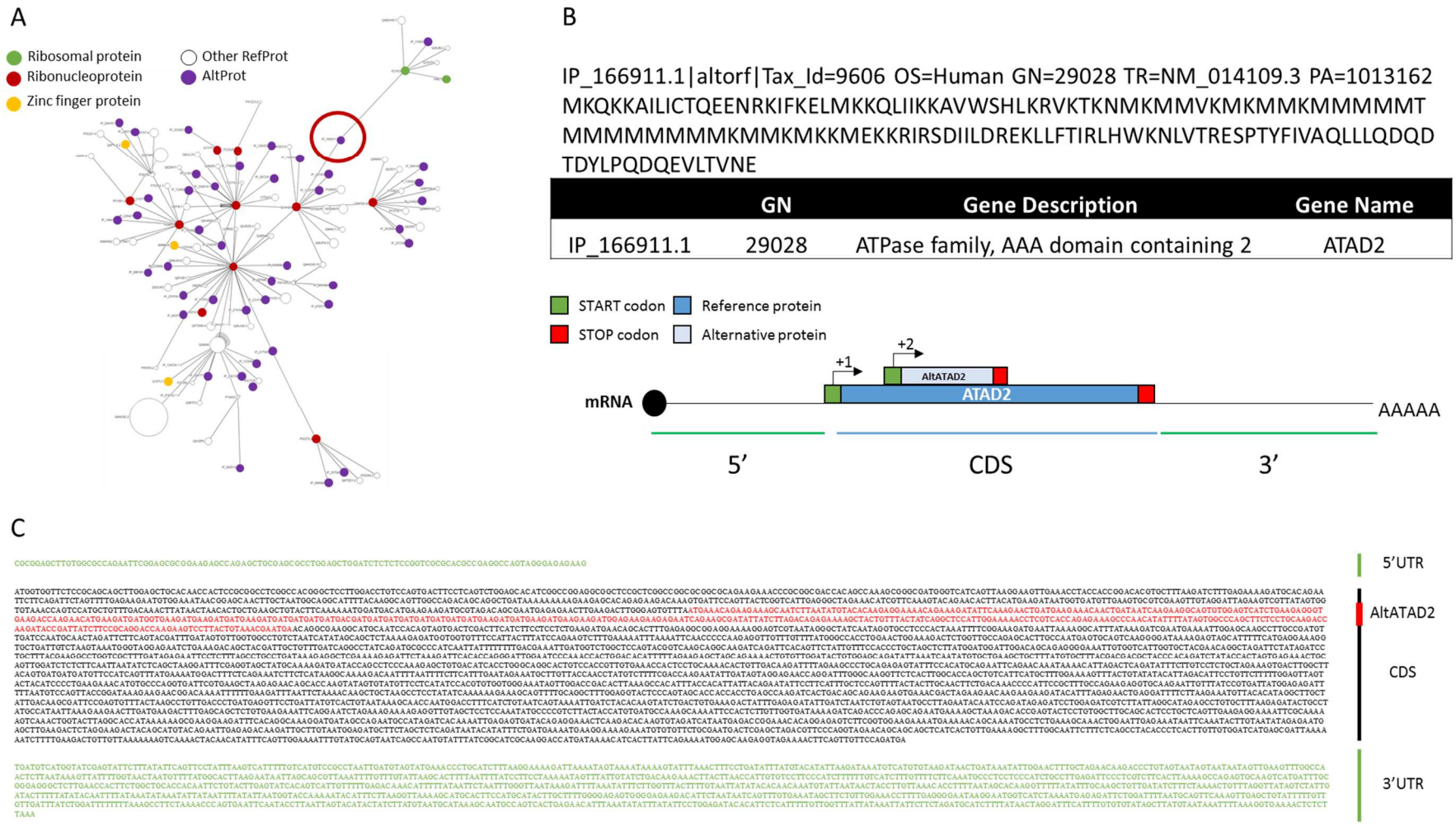
A
Ribosomal protein
Ribonucleoprotein
Zinc finger protein
Other RefProt
AltProt
B
IP_166911.1|altorf|Tax_Id=9606 OS=Human GN=29028 TR=NM_014109.3 PA=1013162
MKQKKAILICTQEENRKIFKELMKKQLIIKKAVWSHLKRVKTKNMKMMVKMKMMKMMMMMT
MMMMMMMMKMMKMKKMEKKRIRSDIILDREKLLFTIRLHWKNLVTRESPTYFIVAQLLLQDQD
TDYLPQDQEVLTVNE
GN
Gene Description
Gene Name
IP_166911.1
29028
ATPase family, AAA domain containing 2
ATAD2
START codon
STOP codon
Reference protein
Alternative protein
+1
+2
AltATAD2
ATAD2
mRNA
AAAAA
5’
CDS
3’
C
5’UTR
AltATAD2
CDS
3’UTR

Figure 6

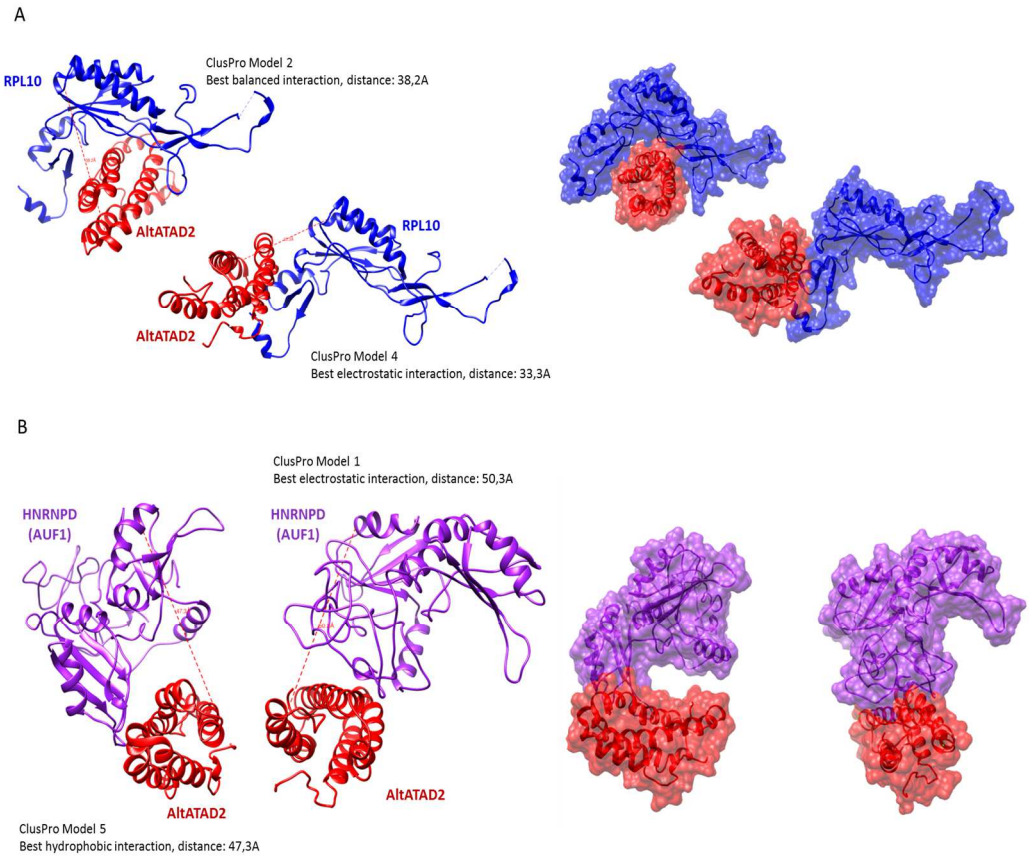
A
RPL10
ClusPro Model 2
Best balanced interaction, distance: 38,2A
AltATAD2
RPL10
AltATAD2
ClusPro Model 4
Best electrostatic interaction, distance: 33,3A
B
ClusPro Model 1
Best electrostatic interaction, distance: 50,3A
HNRNPD
(AUF1)
HNRNPD
(AUF1)
AltATAD2
AltATAD2
ClusPro Model 5
Best hydrophobic interaction, distance: 47,3A
C

Figure 7

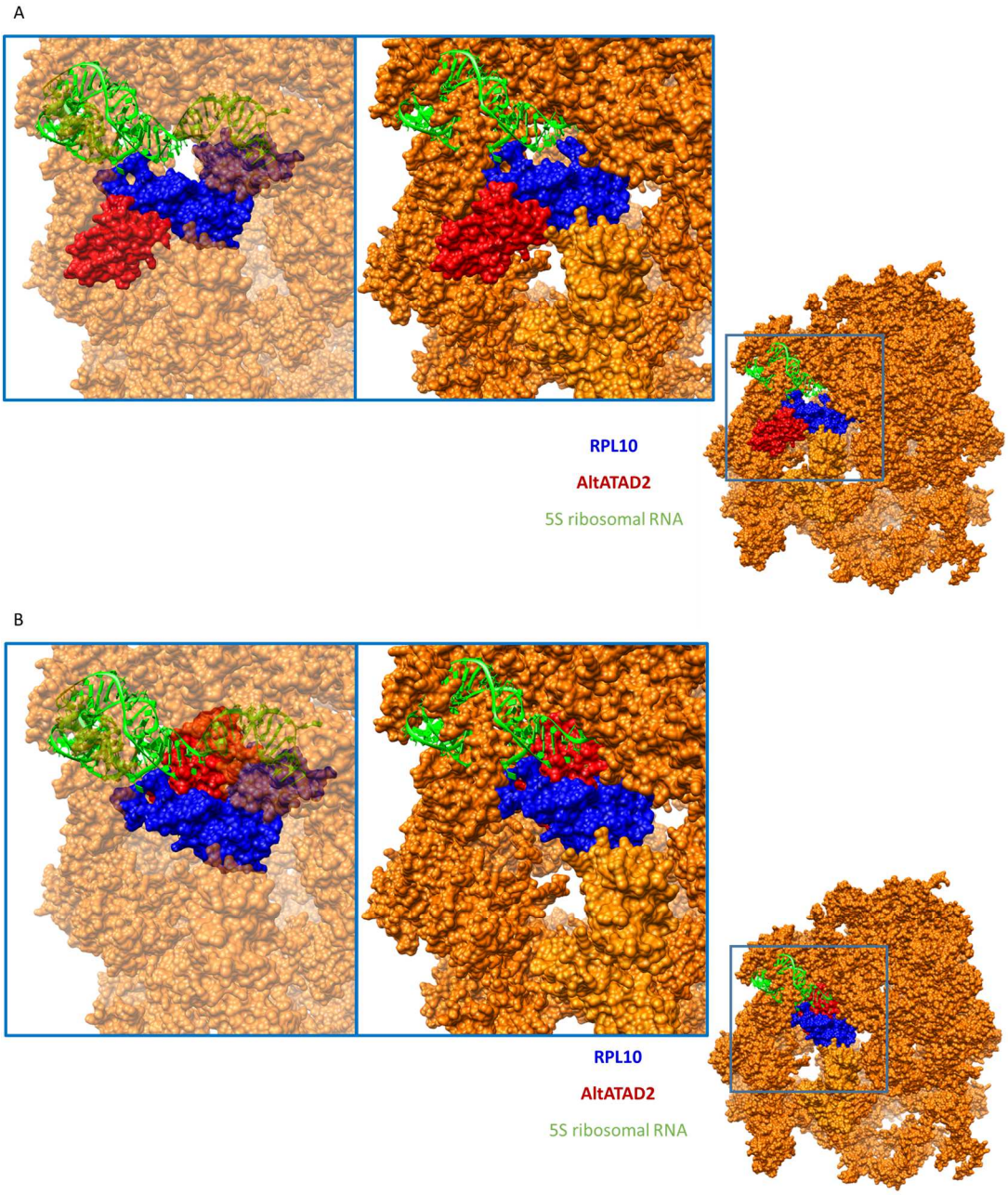
A
RPL10
AltATAD2
5S ribosomal RNA
B
RPL10
AltATAD2
5S ribosomal RNA

Figure 8

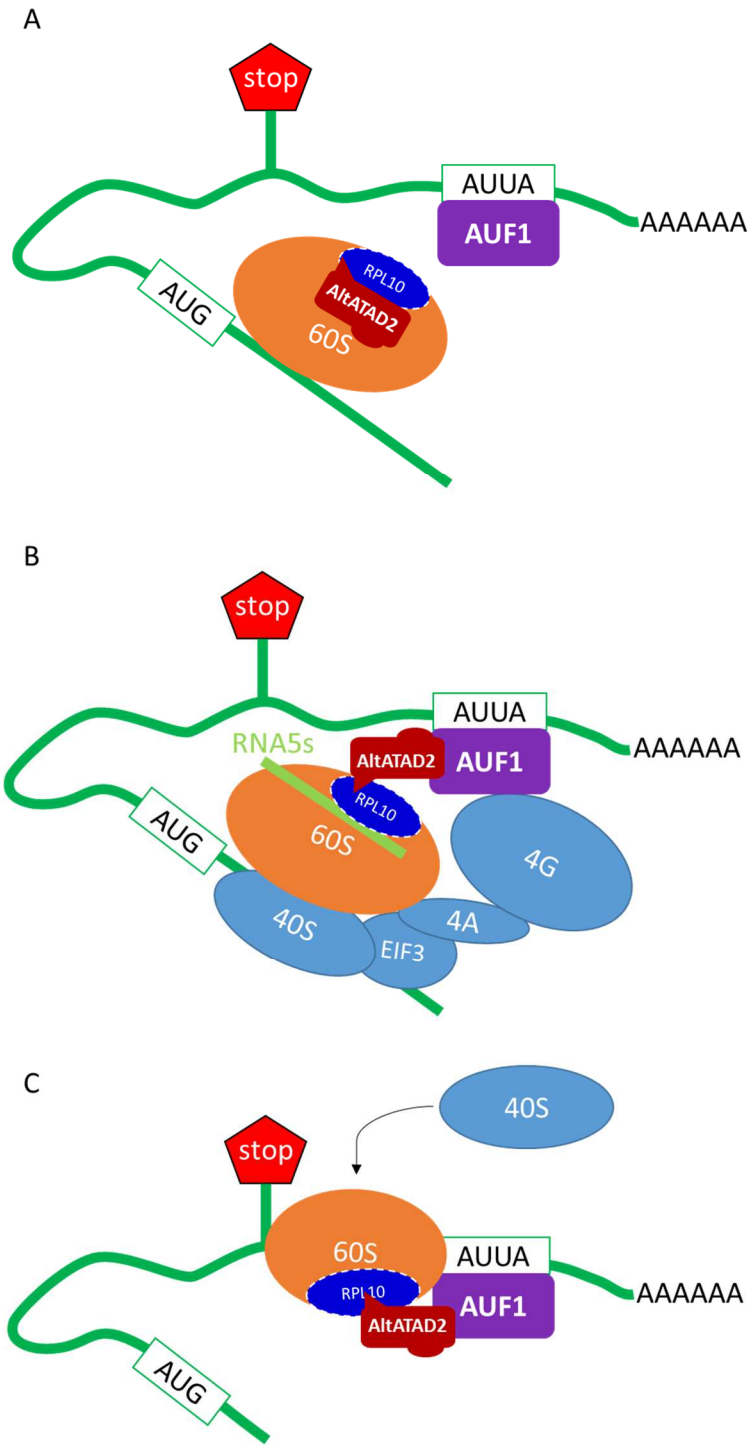
A
stop
AUUA
AUF1
AAAAAA
AUG
RPL10
AltATAD2
60S
B
stop
AUUA
RNA5s
AltATAD2
AUF1
AAAAAA
RPL10
AUG
60S
4G
40S
4A
EIF3
C
40S
stop
60S
AUUA
RPL10
AUF1
AAAAAA
AltATAD2
AUG

Figure 9

Graphical abstract

Schematic representation of the Ghost protein AltATAD2 interaction found in the XL-MS data. ALtATAD2-RPL10 interaction would be involved to in the formation of the 60S ribosome and it's binding of 5S rRNA. Moreover, the co-interaction of AUF1-AltATAD2-RPL10 is hypothesized to play a role in the possible regulation of the expression of the AltProts in the 3'UTR region of the mRNAs.