Newfound coding potential of transcripts unveils missing members of human protein communities

Sebastien Leblanc[1,2], Marie A Brunet[1,2], Jean-François Jacques[1,2], Amina M Lekehal[1,2], Andréa Duclos[1], Alexia Tremblay[1], Alexis Bruggeman-Gascon[1], Sondos Samandi[1,2], Mylène Brunelle[1,2], Alan A Cohen[3], Michelle S Scott[1], Xavier Roucou[1,2,*]

[1]Department of Biochemistry and Functional Genomics, Université de Sherbrooke, Sherbrooke, Quebec, Canada.

[2] PROTEO, Quebec Network for Research on Protein Function, Structure, and Engineering.

[3]Department of Family Medicine, Université de Sherbrooke, Sherbrooke, Quebec, Canada.

*Corresponding author: Tel. (819) 821-8000x72240; E-Mail: xavier.roucou@usherbrooke.ca

Running title: Alternative proteins in communities

Keywords: alternative proteins, protein network, protein-protein interactions, pseudogenes, affinity purification-mass spectrometry

Abstract

Recent proteogenomic approaches have led to the discovery that regions of the transcriptome previously annotated as non-coding regions (i.e. UTRs, open reading frames overlapping annotated coding sequences in a different reading frame, and non-coding RNAs) frequently encode proteins (termed alternative proteins). This suggests that previously identified protein communities are partially incomplete since alternative proteins are not present in conventional protein databases. Here we incorporate this increased diversity in the re-analysis of a high throughput human network proteomics dataset thereby revealing the presence of 203 alternative proteins within 163 distinct communities associated with a wide variety of cellular functions and pathologies. We found 19 genes encoding both an annotated (reference) and an alternative protein interacting with each other. Of the 136 alternative proteins encoded by pseudogenes, 38 are direct interactors of reference proteins encoded by their respective parental gene. Finally, we experimentally validate several interactions involving alternative proteins. These data improve the blueprints of the human protein-protein interaction network and suggest functional roles for hundreds of alternative proteins.

Introduction

Cellular functions depend on myriads of protein communities acting in consort, and understanding cellular mechanisms on a large scale will require a relatively exhaustive catalog of protein-protein interactions. Hence, there have been major efforts to perform high throughput experimental mapping of physical interactions between human proteins (Luck *et al*, 2017). The methodologies involve binary interaction mapping using yeast 2-hybrid (Rolland *et al*, 2014), biochemical fractionation of soluble complexes combined with mass spectrometry (Wan *et al*, 2015), and affinity-purification coupled with mass-spectrometry (Huttlin *et al*, 2015, 2017; Liu *et al*, 2018).

In parallel to these experimental initiatives, computational tools were developed to help complete the human interactome (Keskin *et al*, 2016). Such tools are particularly useful for the identification of transient, cell-type or environmentally dependent interactions that escape current typical experimental protocols. Computational methods that can be used at large scales are created and/or validated using protein-protein interactions previously obtained experimentally (Keskin *et al*, 2016; Kovács *et al*, 2019). Thus, although computational tools complement experimental approaches, the experimental detection of protein-protein interactions is key to building a comprehensive catalog of interactomes.

The BioPlex network is the largest human proteome-scale interactome; initially, BioPlex 1.0 reporting 23744 interactions among 7668 proteins was followed by BioPlex 2.0, which forms the basis of the current study, with 56553 interactions reported involving 10961 proteins. Recent pre-print BioPlex 3.0 reached 118162 interactions among 14586 proteins in HEK293T cells

(Huttlin *et al*, 2017, 2015, 2020). The enrichment of interactors of roughly half of currently annotated (or reference) human proteins allowed the authors to functionally contextualize poorly characterized proteins, identify communities of tight interconnectivity, and find associations between disease phenotypes and these protein groups. In addition, pre-print BioPlex now provides a first draft of the interactome in HCT116 cells (Huttlin *et al*, 2020).

The experimental strategy behind BioPlex is based on the expression of each protein-coding open reading frame (ORF) present in the human ORFeome with an epitope tag, the affinity purification of the corresponding protein, and the confident identification of its specific protein interactors by mass spectrometry. The identification of peptides and proteins in each protein complex is performed using the Uniprot database. Hence, only proteins and alternative splicing-derived protein isoforms annotated in the Uniprot database can be detected. Using this common approach, the human interactome is necessarily made up of proteins already annotated in the Uniprot database, precluding the detection of novel unannotated proteins. Yet, beyond isoform derived proteomic diversity, multiple recent discoveries point to a general phenomenon of translation events of non-canonical ORFs in both eukaryotes and prokaryotes, including small ORFs and alternative ORFs (altORFs) (Brunet *et al*, 2020b; Orr *et al*, 2020). Typically, small ORFs are between 10 and 100 codons, while altORFs can be larger than 100 codons. Here, we use the term altORFs for non-canonical ORFs independently of their size. On average, altORFs are ten times shorter than conventional annotated ORFs but several thousands are longer than 100 codons (Samandi *et al*, 2017). AltORFs encode alternative proteins (altProts) and are found both upstream (i.e. 5'UTR) and downstream (i.e. 3'UTR) of the reference coding sequence as well as overlapping the reference coding sequence in a shifted reading frame within mRNAs (Fig 1A-B). Additionally, RNAs transcribed from long non-coding RNA genes and pseudogenes are systematically annotated as non-coding RNAs (ncRNAs); yet, they may also harbor altORFs and encode alternative proteins (Samandi *et al*, 2017). Consequently, the fraction of multi-coding or polycistronic human genes and of protein-coding "pseudogenes" may have been largely underestimated. AltORFs translation events are experimentally detected by ribosome profiling (Orr *et al*, 2020), a method that detects initiating and/or elongating ribosomes at the transcriptome wide level (Ingolia *et al*, 2019). Alternatively, large-scale mass spectrometry detection of alternative proteins requires first the annotation of altORFs and then *in-silico* translation of these altORFs to generate customized protein databases containing the sequences of the corresponding proteins (Delcourt *et al*, 2017). This integrative approach, termed proteogenomics, has emerged as a new research field essential to better capture the coding potential and the diversity of the proteome (Nesvizhskii, 2014; Ruggles *et al*, 2017).

The translation of altORFs genuinely expands the proteome, and proteogenomics approaches using customized protein databases allows for routine MS-based detection of altProts (Brunet *et al*, 2019; Delcourt *et al*, 2018). In order to uncover altProts otherwise undetectable using the UniProt database we re-analyzed the raw MS-data from the BioPlex 2.0 interactome with our OpenProt proteogenomics database.

OpenProt contains the sequences of proteins predicted to be encoded by all ORFs larger than 30 codons in the human transcriptome. This large ORFeome includes ORFs encoding proteins annotated by NCBI RefSeq, Ensembl and Uniprot, termed here reference proteins or refProts. It also includes still unannotated ORFs that encode novel isoforms sharing a high degree of similarity with refProts from the same gene. Finally, the third category of ORFs, termed altORFs, potentially encode altProts and shares no significant sequence similarity with a refProt from the same gene (Table 1). OpenProt is not limited by the three main assumptions that shape current annotations: (1) a single functional ORF in each mRNA, typically the longest ORF; (2) RNAs with ORFs shorter than 100 codons are typically annotated as ncRNAs; and (3) RNAs transcribed from genes annotated as pseudogenes are automatically annotated as ncRNAs. Thus, in addition to proteins present in NCBI RefSeq, Ensembl and Uniprot, OpenProt also contains the sequence for novel proteins, including novel isoforms and alternative proteins (Brunet *et al*, 2019, 2020c).

Using OpenProt, we were able to detect and map altProts within complexes of known proteins which increased protein diversity by including a higher number of small proteins. In addition, the data confirmed the significant contribution of pseudogenes to protein networks with 124 out of 280 altProts encoded by genes annotated as pseudogenes. We also detected many interacting proteins encoded either by the same gene or by a pseudogene and its corresponding parental gene. In sum, this work improves our knowledge of both the coding potential of the human transcriptome and the composition of protein communities by bringing diversity (i.e. small proteins) and inclusivity (i.e. proteins encoded in RNAs incorrectly annotated as ncRNAs) into the largest human protein-protein interaction (PPI) network to date.

Results

Re-analysis of BioPlex 2.0 mass spectrometry data and identification of preyed alternative proteins

We employed the OpenProt proteogenomics library in the re-analysis of high throughput AP-MS experiments from the BioPlex 2.0 network. Given the size of the OpenProt database (Fig 1C), the false discovery rate (FDR) for protein identification was adjusted from 1 % down to 0.001 % to mitigate against spurious identifications (Brunet *et al*, 2019). Such stringent FDR settings inevitably lead to fewer prey proteins identified; thus, our highly conservative methodology is likely to leave behind many false negatives. The BioPlex 2.0 network is built in a gene-centric manner in order to simplify the analysis by making abstraction of protein isoforms. In the current analysis, all refProts and their isoforms are also grouped under their respective gene, but results concerning altProts are necessarily given at the protein level.

In total, 434 unannotated proteins from 418 genes and 5669 refProts were identified in the re-analysis of raw MS data from the pull-down of 3033 refProts (baits), using a combination of multiple identification algorithms (Fig 1C). Since these identifications resulted from the re-analysis of raw MS data from BioPlex 2.0 with the OpenProt MS pipeline, we sought to determine the overlap between total sets of genes identified. RefProts from 4656 genes (or 86 % of total re-analysis results) were found in both the BioPlex 2.0 and in the present work (Fig EV1A), indicating that the re-analysis could reliably reproduce BioPlex results.

Our stringent approach in the identification of altProts included the use of PepQuery to validate protein detection using a peptide-centric approach (Wen *et al*, 2019). This tool includes a step which verified that altProt-derived peptides were supported by experimental spectra that could not be better explained by peptides from refProts with any post-translational modification. In addition, peptides were screened for isobaric substitutions in order to reject dubious peptides that could match refProts (Choong *et al*, 2017). A total of 295 altProt identifications were validated with PepQuery including 136 altProts encoded by pseudogenes (Table EV1). MS-based identification of short proteins with a minimum of 2 unique suitable tryptic peptides remains an important challenge and the majority of short proteins are typically detected with a single unique peptide (Slavoff *et al*, 2013; Ma *et al*, 2014). Of the 295 altProts validated by PepQuery (Table EV2), 63 complied with the Human Proteome Project PE1 level for proteins with strong protein-level evidence, Guidelines v3.0 (Deutsch *et al*, 2019).

As expected, detected altProts were much shorter than refProts with a median size of 78 amino acids versus 474 (Fig 1D; Table EV1). AltORFs encoding the 295 detected and PepQuery-validated altProts were distributed among 1029 transcripts (Table EV1) and in addition to the 136 pseudogenes derived altProts, 38 were exclusively encoded by genes of non-coding biotypes (Fig 1E). A third were found in transcripts already encoding a refProt (Fig 1E), indicating that the corresponding genes are in fact either bicistronic (two non-overlapping ORFs) or dual-coding (two overlapping ORFs) (Table EV1). Of the altProts encoded by transcripts from genes of protein coding biotype, most were encoded by a frame-shifted altORF overlapping the annotated coding sequence or downstream of the annotated coding sequence in the 3'UTR (Fig 1F). The remaining altORFs were encoded by 5'UTRs or by transcripts annotated as non-coding but transcribed from those genes of protein coding biotype. From the localization of altORFs relative to the canonical CDS in the 107 mRNA from protein coding genes we conclude that 56 of those genes are in fact bicistronic and 51 are dual-coding (Table EV1). In addition, transcripts from 7 pseudogenes have been found to encode two altProts suggesting that 3 of them are in fact dual coding and 4 are bicistronic (Table EV1).

We collected protein orthology relationships from 10 species computed by OpenProt (Fig 1G). Although 100 altProts were specific to humans, a large number had orthologs in the mouse and chimpanzee, and 28 were even conserved through evolution since yeast. 167 altProts displayed at least one functional domain signature (InterProScan, version 5.14-53.0, (Mitchell *et al*, 2019)), further supporting their functionality (Table EV1).

Network assembly

After identification of prey proteins, CompPASS was used to compute semi-quantitative statistics based on peptide-spectral matches across technical replicates (Sowa *et al*, 2009). These metrics allow filtration of background and spurious interactions from the raw identifications of prey proteins to obtain high confidence interacting proteins (HCIP). To mitigate against the otherwise noisy nature of fast-paced high throughput approaches and to filter prey identifications down to the most confident interactions, we applied a Naïve Bayes classifier similar to CompPASS Plus (Huttlin *et al*, 2015). The classifier used representations of bait-prey pairs computed from detection statistics and assembled into a vector of 9 features as described by (Huttlin *et al*, 2015). High confidence interactions reported by BioPlex 2.0 served as target labels. HCIP classification resulted in the retention of 3.6 % of the starting set of bait-prey pairs identified (Fig EV1C). Notably, 815 baits from the original dataset were excluded after filtration because no confident interaction could be distinguished from background.

Following protein identifications and background filtration, the network was assembled by integrating all bait-prey interactions into one network (Fig 2A). All refProts and their isoforms were grouped under their respective gene, similar to the BioPlex analysis, but separate nodes are shown for altProts. In total, the re-analysis with OpenProt found 5650 prey proteins from the purification of 2218 bait proteins altogether engaged in 14029 interactions, the majority (59.1 %) of which were also reported by BioPlex 2.0 (Fig 2B). The average number of interactions per bait was 7.1. Among prey proteins, 280 altProts were found engaged in 347 interactions with 292 bait proteins.

Compared to BioPlex 2.0, a smaller total number of protein identification was expected because the OpenProt MS analysis pipeline is more stringent with a tolerance of 20 ppm on peak positions rather than 50 ppm and a 0.001 % protein FDR as opposed to 1 %. Indeed, we identified 14029 interactions in our reanalysis, compared to 56553 interactions reported by BioPlex 2.0 (Fig 2B). Among the 14029 interactions, 8288 (59.1 %) were also reported by BioPlex 2.0, and 7979 (56.8 %) were reported in the recently released (but not yet peer reviewed) BioPlex 3.0 (Fig 2B). Interestingly, 11329 interactions (20 %) from BioPlex 2.0 were not confirmed in BioPlex 3.0 using a larger number of protein baits, although the same experimental and computational methodologies were used (Fig 2B). This observation illustrates the challenge in the identification of protein-protein interactions with large-scale data given the relatively low signal to noise ratio in AP-MS data.

Network structural features and alternative protein integration

Network theoretic analysis confirmed that the OpenProt-derived network displayed the expected characteristics of natural networks. Variability in the number of interacting partners of a given protein in a network (node degree) is typically very wide and the degree distribution that characterizes this variation follows a power-law (Bianconi & Barabási, 2001). Similar to other protein networks, the degree distribution of the OpenProt-derived network also fitted a power-law, an indication that the vast majority of proteins have few connections and a minor fraction is highly connected (also called hubs) (Fig 2C). The degree of connectivity of altProts varied between 1 and 7 whereas that of refProt was between 1 and 84. On the one hand, since long and multidomain proteins are over-represented among hub proteins (Ekman *et al*, 2006), this difference may be explained by the fact that altProts in the network were on average 6 times shorter than refProts (Fig 1D). On the other hand, none of the altProts were used as baits which also explains their lower observed connectivity since average degree was 2.5 for preys but 7.1 for baits.

The mean degrees of separation between any two proteins in the OpenProt-derived network was 5 (Fig 2D), in agreement with the small-world effect that characterizes biological networks (Wagner & Fell, 2001).

Centrality analysis allows us to sort proteins according to their relative influence on network behaviour where the most central proteins tend to be involved in the most essential cellular processes (Jeong *et al*, 2001). Here, the eigenvector centrality measure indicates that altProts are found both at the network periphery connected to refProts of lesser influence as well as connected to central refProts of high influence (Fig 2E).

Known complexes from the CORUM database were mapped onto the network to assess the portion of complex subunits identified in the re-analysis (Table EV3). In most cases a majority were recovered (75 % of complexes showed ≥50 % recovery) (Fig 2F). We observed 50 altProts in the neighborhood of CORUM complex subunits that served as bait. Here multiple interesting patterns of altProt interactions were already noticeable: (1) altProts detected in the interactome of their respective refProts (Fig 2Gi), (2) altProts originating from pseudogenes and detected in the interactome of refProts encoded by the parental gene (Fig 2Gii-iii) and (3) altProts from protein coding genes or pseudogenes detected in network regions outside the immediate neighborhood of the related protein/gene (Fig 2Giv-vi).

The OpenProt-derived protein-protein interaction network displayed with a degree sorted circle layout showed that preyed altProts generally had a lower degree of connectivity compared to refProts (Fig 3A). This might be expected in part because no altProts were used as baits in the network, but also based on the limited range of binding capacity due to their smaller size. In order to investigate the local neighborhood of altProts, subnetworks were extracted by taking nodes within shortest path length of 2 and all edges between these for each altProt (here called second neighborhood). Notable altProts with high degree include OpenProt accessions IP_117582, a novel protein encoded by an altORF overlapping the reference coding sequence in the *BEND4* gene (Fig 3Ai), and IP_711679, encoded in a transcript of the *SLC38A10* gene currently annotated as a ncRNA (Fig 3Aii). Although these two altProts would not qualify as hub proteins per say, they seem to participate in the bridging of hubs from otherwise relatively isolated regions. Several other examples of altProts encoded by a lncRNA gene (Fig 3Aiii), in pseudogenes (Fig 3Aiv, v, vii, viii), and in protein-coding genes (Fig 3Avi, ix) integrate the network with a variety of topologies. One of these subnetworks features IP_710744, a recently discovered altProt and polyubiquitin precursor with 3 ubiquitin variants, encoded in the *UBBP4* pseudogene (Dubois *et al*, 2020). Theubiquitin variant Ubbp4^{A2} differs from canonical ubiquitin by one amino acid(T55S) and can be attached to target proteins (Dubois *et al*, 2020). Before network assembly this variant was identified reproducibly (across technical replicates) in the purification of 11 baits. Following HCIP identifications, only 3 interactions remained (Fig 3Aiv), likely because widespread identifications lead the Naïve Bayes classifier to assume non-specificity for those showing lower abundance. The 3 interactors include 2 ubiquitin ligases (*UBE2E2* (Q96LR5) and *UBE2E3* (Q969T4)) and *USP48* (Q86UV5), a peptidase involved in the processing of ubiquitin precursors.

After observing second neighborhoods of altProts we sought to evaluate the effect of altProt inclusion into local neighborhoods of refProts. To do so we computed the eigenvector centrality of each refProt within their own second neighborhood extracted from the assembled network with and without altProts. This analysis highlighted *ELP6* which undergoes a marked reduction in eigenvector centrality in its second neighbourhood (0.67 versus 0.56) when the altProt IP_688853 (encoded by the 'non-coding' gene AC092329.4) is included (Fig 3Bi,ii). This shows that node influence in this region of the network is poorly understood and that identifications of novel interactors may shed light over the recent association of this gene with tumorigenesis (Close *et al*, 2012).

In total, 45 pseudogene-encoded altProts were uncovered in the direct interactome of refProts from their respective parental genes (Table EV4, shortest path length of 1), of which 2 more examples are illustrated with more details in Fig 3C.

GAPDH is known to have a large number of pseudogenes (Liu *et al*, 2009). Yet protein products originating from 9 *GAPDH* pseudogenes were confidently identified in the purification of the canonical GAPDH protein (Fig 3Ci). Since the glycolytic active form of this enzyme is a tetramer, we conjecture that GAPDH tetramers may assemble from a heterogenous mixture of protein products from the parental gene and many of its pseudogenes. GAPDH is a multifunctional protein (Tristan *et al*, 2011); although different posttranslational modifications may explain in part how this protein switches function (Colell *et al*, 2009), it is possible that heterologous and homologous complexes contribute to GAPDH functional diversity. Especially given that 4 of the smallest protein products from *GAPDH* pseudogenes only contain the GAPDH NAD binding domain (IPR020828; IP_735797, IP_761275, IP_735800, IP_591881), the protein encoded by *GAPDHP1* only contains the GAPDH catalytic domain (IPR020829; IP_560713), while the largest proteins from *GAPDH* pseudogenes contain both domains (IP_557819, IP_672168, IP_3422225, IP_755869) (Table EV1). The *PHB1* subnetwork highlights an interaction between *PHB1* and *PHBP19*, one of the 21 *PHB* pseudogenes (Fig 3Bii). *PHB1 and PHB2* are paralogs and the proteins they encode, PHB1 and PHB2, heterodimerize; similar to GAPDH, the PHB1/PHB2 complex is multifunctional (Osman *et al*, 2009), and the dimerization of PHB1 or PHB2 with *PHBP19*-derived IP_762813, which also contains a prohibitin domain (IPR000163), may regulate the various activities of the complex.

We reasoned that pseudogene-derived altProts directly interacting with their parental gene-derived refProts (parental protein) may result from the generally high degree of sequence similarity, particularly for refProts known to multimerize. However, although a slight reduction of alignment scores was observed with an increase in degrees of separation, the 45 altProts directly interacting with parental protein display a large variety of sequence alignment scores (Fig 3Di). This suggests that direct interactions between pseudogene-derived altProts and their respective parental refProts involve other mechanisms in addition to sequence identity. Since 42 of the 45 altProts share between 1 and 7 InterPro entries with their respective parental proteins (Table EV4), protein domains may be an important mechanism driving these interactions.

The mean degrees of separation between a refProt and an altProt encoded in the same gene reveals two types of relationships (Fig 3Dii). 25 % (18) of altProt-refProt pairs have a degree of separation of 1, that is to say these altProts were found in the direct interactome of the corresponding refProt from the same gene. Hence, these protein pairs encoded by the same genes are clearly involved in the same function through direct or indirect physical contacts. Interestingly, 15 of these 18 altProts are encoded by dual-coding genes, i.e. with altORFs overlapping annotated CDSs. 75 % of altProt-refProt pairs follow a distribution of degrees of separation similar to the whole network (compare Fig 3Dii and 2D). This suggests that they are not more closely related than any other 2 proteins in the network despite shared transcriptional regulation.

Cluster detection reveals altProts as new participants in known protein communities

Biological networks are organised in a hierarchy of interconnected subnetworks called clusters or communities. To identify these communities, unsupervised Markov clustering (MCL) (Enright *et al*, 2002) was used similarly to methodology applied to BioPlex 2.0 (Huttlin *et al*, 2017). Partitioning of the network resulted in 1045 protein clusters, 163 of which contained at least one altProt (Fig 4A). The size of altProts in these communities varied between 29 to 269 amino acids indicating that protein length may not be a limiting factor in their involvement in functional groups. Links between clusters were drawn where the number of connections between members of cluster pairs was higher than expected (detailed in Materials and Methods).

In order to assign biological function to these clusters, and therefore generate testable hypotheses about the function of altProts detected among them, enrichment of gene ontology (GO) terms was computed for each community against the background of all human genes. Several communities of different sizes showing significant GO term enrichment are detailed in Fig 4B.

45 % of identified clusters showed GO term enrichment. The same analysis with the original BioPlex network showed 57 % of clusters with GO term enrichment; possibly because a higher number of protein identifications yielded a larger network and therefore a higher probability of significant enrichment.

The altProt IP_293201 from the gene *RNF215* was identified as a novel interactor of three subunits of the RNA exosome multisubunit complex (cluster #46), suggesting a possible role in RNA homeostasis. Clusters #214 and #369 included protein communities with essential activities: the large eukaryotic initiation factor EIF3 and the recently discovered KICSTOR complex, a lysosome-associated negative regulator of mTORC1 signaling (Wolfson *et al*, 2017, 1). At least one pseudogene encoded altProt was detected in each of these clusters. Intriguingly, altProts IP_790907 (cluster #214) and IP_602155 (cluster #369) interact with the parental proteins EIF3E and ITFG2, respectively. These altProts may either compete with the parental proteins to change the activity of the complexes, or function as additional subunits since each contains a relevant functional domain (initiation factor domain, IPR019382, and ITFG2 domain, PF15907, respectively). Several subunits of the spliceosome are present in cluster #15, a protein community that includes IP_637160, a novel interactor of SNRPA1, which contains a

U2A'/phosphoprotein 32 family A domain (IPR003603) where U2A' is a protein required for the spliceosome assembly (Caspary & Séraphin, 1998). Cluster #115 contains the two regulatory subunits of PKA, PRKAR1B and PRKAR2B, which form a dimer, and several A-kinase scaffold proteins that anchor this dimer to different subcellular compartments (Di Benedetto *et al*, 2008). Three altProts interacting with PRKAR2B are also present in this cluster. Interestingly, altProt IP_156019 is encoded by an altORF overlapping the canonical PRKAR2B coding sequence; hence, *PRKAR2B* is a dual-coding gene with both proteins, the refProt and the altProt, interacting with each other. The discovery of new altProts in known protein communities demonstrates a potential for the increase in our knowledge of biological complexes.

Disease association

The curated list of disease-gene associations published by DisGeNET relates 6,970 genes with 8,141 diseases in 32,375 associations (Piñero *et al*, 2020). After mapping this gene-disease association network onto our network of protein communities, 804 clusters of which 116 contained at least one altProt were found in association with 3,668 diseases (Fig 5A). The 116 gene-disease associations involving at least one altProt were distributed among 22 disease classes (Fig 5B). The distribution of disease-cluster associations involving altProts among the disease classes was similar to those involving refProts. Thus, no preferential association of altProts with certain disease classes could be observed.

A selection of subnetworks illustrates how altProts associate with different diseases (Fig 5C). *ADAM10* encodes a transmembrane refProt with metalloproteinase activity. Among protein substrates that are cleaved by ADAM10 and shed from cells, some act on receptors and activate signaling pathways important in normal cell physiology (Reiss & Saftig, 2009). Overexpression of this protease or increased shedding of tumorigenic proteoforms results in overactivation of signaling pathways and tumorigenesis (Murphy, 2008; Smith *et al*, 2020). IP_233890 is an altProt expressed from bicistronic *ADAM10* and its association with a subnetwork of transcription factors involved in tumorigenesis may further clarify the role of that gene in cancer (Fig 5Ci). Cluster #199 illustrates the association of a pair of refProt/altProt expressed from the same dual-coding gene, ZNF*408*, with three different diseases (Fig 5Cii). The implication of pseudogene-derived altProts is emphasized by the association of three of them with Acute Myelocytic Leukemia through their interaction with *ANXA2* (Fig 5C iii). Two of these interactions occur between a refProt from the parental gene and altProts encoded by two of its pseudogenes.

Cluster #133 relates proteins localized at the membrane with roles in intercellular signaling, development and organogenesis, as well as fatty acids transport proteins (Mahesh, 2013; Drazyk *et al*, 2019; Short *et al*, 2007, 1; Kim *et al*, 2020). AltProt IP_656413 associated with this cluster is coded by a pseudogene of the breakpoint cluster protein BCR, a Rho GTPase activating protein. IP_656413 is predicted to have a Rho GTPase activating protein domain InterProScan analysis (IPR000198) (Table EV1). Associations of this cluster with diseases both common (bronchial hypersensitivity) and rare (Fraser syndrome) highlight the potential of deeper protein coding annotations coupled with network proteomic studies to unveil novel members relevant to a wide array of pathological phenotypes. Characterization of the role of this altProt at the membrane, likely involved in intercellular signaling, may yield mechanistic insight surrounding associated pathologies.

Functional validation of protein-protein interactions involving an alternative protein

Interactions representative of the three following classes of complexes involving altProts were selected for further experimental validation: an altProt encoded by a dual-coding gene and interacting with the respective refProt, an altProt expressed from a pseudogene and interacting with the refProt encoded by the parental gene, and an altProt interacting with a refProt coded by a different gene.

The dual-coding *FADD* gene expresses altProt IP_198808 in addition to the conventional FADD protein, and both proteins interact within the DISC complex (Fig 2Gi). We took advantage of a previous study aiming at the identification of the FADD interactome to test whether this altProt may also have been missed in this analysis because the protein database used did not contain altProt sequences (Eyckerman *et al*, 2016). In this work, the authors developed a new method called ViroTrap to isolate native protein complexes within extracellular virus-like particles to avoid artefacts of cell lysis in AP-MS. Among the baits under study FADD was selected to isolate the native FADD complex. First, we used the peptide-centric search engine PepQuery to directly test for the presence or the absence of IP_198808-derived specific peptides in the FADD complex datasets. Rather than interpreting all MS/MS spectra, this approach tests specifically for the presence of the queried peptides (Ting *et al*, 2015). Indeed, two unique peptides from IP_198808 were detected in each of the replicates of that study via PepQuery (Fig EV3A i,v). Second, we used a conventional spectrum-centric and database search analysis with the UniProt database to which was added the sequence of IP_198808. The altProt was identified in the FADD interactome (Fig EV3B) with 4 unique peptides (Fig EV3A i,iii,iv,v). In transfected cells, FADD formed large filaments (Fig 6A, right), previously labelled Death Effector Filaments (Siegel *et al*, 1998). IP_198808 co-localized in the same filaments in the nucleus, while the cytosolic filaments contained FADD only. Finally, this interaction was validated by co-immunoprecipitation (Fig 6A, left). These proteomics, microscopic and biochemical approaches confirmed the interaction between the two proteins encoded in dual-coding *FADD*.

Next, we selected 2 pairs of interactions of an altProt expressed from a pseudogene with a refProt expressed from the corresponding parental gene. The interaction between altProt IP_624363 encoded in the *EEF1AP24* pseudogene and EEF1A1 (Fig 3Av) was confirmed by co-immunoprecipitation (Fig 6B, left). Both proteins also displayed strong co-localization signals (Fig 6B, right). In order to validate the interaction between *PHBP19*-encoded IP_762813 and PHB1, we performed two experiments. First, PHB1-GFP co-immunoprecipitated with IP_762813 (Fig 6C, left). Second, we performed independent AP-MS experiments for both IP_762813 and PHB1 in HEK293 cells. We confirmed the presence of PHB1 in the interactome of IP_762813 and the presence of IP_762813 in the interactome of PHB1 (Fig 6D, right). Interestingly, we observed shared interactors between IP_762813 and PHB1 (IRS4 (O14654), ATP1A1 (P05023) and XPO1 (O14980)), as well as interactors specific to each. Prey-prey interactions from STRING also showed a certain interconnectivity of both interactomes, whilst each retained unique interactors (Fig EV3C).

The altProt IP_117582 encoded in the *BEND4* gene is one of the most central and most connected alternative proteins in our network (Fig 3A). The interaction with RPL18 was tested and confirmed by co-immunoprecipitation (Fig 6D, left), and their co-localization was also confirmed by immunofluorescence (Fig 6D, right).

Discussion

The discovery of unannotated altProts encoded by ORFs localized in “non-coding” regions of the transcriptome raises the question of the function of these proteins. The translation of altProts may result from biological translational noise producing non-bioactive molecules. Alternatively, altProts may play important biological roles (Orr *et al*, 2020). Here, we addressed the issue of the functionality of altProts by testing their implication in protein-protein interactions. We have reanalyzed the Bioplex 2.0 proteo-interactomics data using the proteogenomics resource OpenProt which provides customized databases for all ORFs larger than 30 codons in 10 species (Brunet *et al*, 2019, 2020c). Under stringent conditions, a total of 295 prey altProts were detected, of which 280 could be confidently mapped in the network of 292 bait refProts. 136 altProts are expressed from pseudogenes, 121 from dual-coding and bicistronic genes, and 38 from transcripts annotated as ncRNA but should in fact be protein-coding. In addition to revealing new members of protein communities, this study lends definitive support to the functionality of hundreds of altProts and provides avenues to investigate their function.

The detection of 295 altProts under stringent conditions confirms the hindrance introduced by three assumptions of conventional annotations: (1) eukaryotic protein-coding genes are monocistronic; (2) RNAs transcribed from genes annotated as pseudogenes are ncRNAs; and (3) ncRNAs are annotated as such based on non-experimental criteria, including the largely used 100 codons minimal length (Dinger *et al*, 2008). The persistence of these assumptions in conventional genomic annotations limits the repertoire of proteins encoded by eukaryotic genomes (Brunet *et al*, 2018). It remains possible that functional altORFs in regions of the transcriptome annotated as non-coding are exceptions and that a large fraction of genes and

RNAs comply with current assumptions. However, an ever-increasing number of proteogenomics studies demonstrate that thousands of altORFs and their corresponding proteins are translated (Samandi *et al*, 2017; Chen *et al*, 2020).

Conventional annotations introduce some confusion by opting to create a new gene entry within a previously annotated gene where a novel protein product has been reported or where novel transcripts have been mapped, rather than annotate a second ORF in the initial gene. The result is that some genomic regions have been assigned a second gene in the same orientation, nested within a previously annotated gene. This is the case for pseudogene *ENO1P1* (Ensembl: ENSG00000244457; genomic location: chr1: 236,483,165-236,484,468 (GRCh38.p13)) which overlaps the protein coding gene *EDARADD* (Ensembl: ENSG00000186197; genomic location: chr1:236,348,257-236,502,915 (GRCh38.p13)) which also encodes altProt IP_079312. Thus, as a result of this annotation, a pseudogene (*ENO1P1*) is nested within a protein-coding gene (*EDARADD*). Similarly, a second protein-coding gene termed *AL022312.1* (Ensembl: ENSG00000285025; genomic location: chr22: 39,504,231-39,504,443 (GRCh38.p13)) was added within the protein-coding *MIEF1* gene (Ensembl: ENSG00000100335; genomic location: chr22:39,499,432-39,518,132 (GRCh38.p13)) to annotate the recently discovered altORF upstream of the *MIEF1* CDS (Samandi *et al*, 2017; Vanderperre *et al*, 2013). We suggest that recognizing the polycistronic nature of some human genes to be able to annotate multiple protein-coding sequences in the same gene is more straightforward than annotating additional small genes nested in longer genes in order to comply with monocistronic annotations.

The involvement of 280 altProts in 347 of the 14029 protein-protein interactions in the current network (or 2.5 %) represents a sizable number of previously missing nodes and edges and contributes to the understanding of network topology. The impact of altProt inclusion on network structure is revealed by the bridging role many seem to play between interconnected regions (Fig 3Ai-ix). This linkage of otherwise independent complexes introduces major changes to network structure shown to be related to biological system state (e.g. cell type) (Huttlin *et al*, 2020). Results from the current analysis are thus anticipated to yield insight regarding molecular function and mechanisms of protein complexes in the contexts of cell type and other suborganismally defined states (Huttlin *et al*, 2020). Indeed, the presence of altProts in protein communities associated with known function and/or diseases makes it possible to generate testable hypotheses regarding their role in physiological and pathological mechanisms (Leblanc & Brunet, 2020).

An important observation stemming from the current study is that many pseudogenes encode one altProt in the network, including some encoding 2 altProts. Strikingly, several altProts expressed from pseudogenes interact with their respective parental protein. This suggests that pseudogene-encoded altProts are functional paralogs and that their incorporation into homomeric protein complexes of the parental protein could modulate or change the activity of the parental complex. Such function would be reminiscent of the role of homomers and heteromers of paralogs in the evolution of protein complexes in yeast, allowing structural and functional diversity (Marchant *et al*, 2019; Pereira-Leal *et al*, 2007). The GAPDH subnetwork with its 9 pseudogene-encoded altProts is particularly striking. Besides its canonical function in glycolysis, GAPDH displays a variety of different functions in different subcellular locations, including apoptosis, DNA repair, regulation of RNA stability, transcription, membrane fusion, and cytoskeleton dynamics (Colell *et al*, 2009; Sirover, 2012; Tristan *et al*, 2011). We propose that the incorporation of different paralog subunits in this multimeric complex results in the assembly of different heteromeric complexes and may at least in part entail such functional and localization diversity. This hypothesis is in agreement with the speculation that the diversity of functions associated with GAPDH correlates with the remarkable number of GAPDH pseudogenes (Liu *et al*, 2009).

Among the 274 genes encoding the 280 altProts inserted in the network, 18 encode refProt/altProt pairs that specifically interact with each other, which implies that these pairs are involved in the same function. Such functional cooperation between a refProt and an altProt expressed from the same eukaryotic gene confirms previous observations in humans (Samandi *et al*, 2017; Chen *et al*, 2020; Bergeron *et al*, 2013; Klemke *et al*, 2001). Dual-coding genes are common in viruses (Chirico *et al*, 2010) and proteins expressed from viral overlapping ORFs often interact (Pavesi *et al*, 2018). The general tendency of physical or functional interaction between two proteins expressed from the same gene should help decipher the role of newly discovered proteins provided that functional characterization of the known protein is available. Molecular mechanisms behind the functional cooperation of such protein pairs remain to be explored.

Furthermore, several pairs of proteins encoded by the same gene but acting in distant parts of the network have also been identified. Could these altProts be a source of cross talk between functional modules under the same regulation at the genetic level, but multiplexed at the protein function level?

The current study shows that the 280 altProts incorporated in the network differ from refProts by their size (6 times smaller in average) but do not form a particular class of gene products;

rather they are members of common communities present throughout the proteomic landscape. Initial serendipitous detection of altProts subsequently called for proteogenomics approaches which widened discoveries via systematic and large-scale detection (Peeters & Menschaert, 2020; Brunet *et al*, 2020b). System resilience and biodiversity have long been linked in the ecology literature (Peterson *et al*, 1998); by analogy the increased proteomic diversity due to altProts could be a contributing factor to this effect in cellular systems. To find out the extent to which altProts play widespread and important biological functions will require more studies in functional genomics.

Materials & Methods

Reanalysis of AP-MS data

Files obtained from the authors of the BioPlex 2.0 contained the results of 8,364 affinity purification-mass spectrometry (AP-MS) experiments using 3033 bait proteins (tagged with GFP) in 2 technical replicates or more barring missing replicates and corrupted files (Huttlin *et al*, 2017, 2015). Files were converted from RAW to MGF format using Proteowizard 3.0 and searched with SearchGUI 2.9.0 using an ensemble of search engines (Comet, OMSSA, X!Tandem, and MS-GF+). Search parameters were set to a precursor ion tolerance of 4.5 ppm and fragment ion tolerance of 20 ppm, trypsin digestion with a maximum of 2 missed cleavages, and variable modifications included oxidation of methionine and acetylation of N termini. The minimum and maximum length for peptides were 8 and 30 amino acids respectively. Search results were aggregated using PeptideShaker 1.13.4 with a 0.001 % protein level false discovery rate (FDR) as described previously (Brunet *et al*, 2019). The protein library contained a non redundant list of all reference proteins from Uniprot (release 2019_03_01), Ensembl (GRCh38.95), and RefSeq (GRCh38.p12) (134477 proteins) in addition to all alternative protein (488956 proteins) and novel isoforms (68612 proteins) predictions from OpenProt 1.6. AltProt identifiers throughout the current article are accessions from OpenProt starting with “IP_”. The library was concatenated with reversed sequences for the target decoy approach to spectrum matching.

Validation of altProt identifications

Novel protein identifications were supported by unique peptides. An additional peptide centric approach was used to validate that spectra supporting such peptides could not be better explained by peptides from refProts with post-translational modifications. PepQuery allows the search of specific peptides in spectra databases using an unrestricted modification search option (Wen *et al*, 2019). All possible peptide modifications from UniMod artifact and post translational modifications were considered when ensuring unicity of spectral matches (downloaded March 2020) (Dm & Js, 2004).

AltProt sequences with peptides validated with PepQuery have been submitted to the Uniprot Knowledge Base.

Obtaining spectral counts

Because altProts are smaller than refProts they have a lower number of uniquely identifying peptides. For this reason altProts with at least one unique peptide across multiple replicates were considered, but only refProts identified with at least two unique peptides across multiple replicates were retained for downstream analysis. Spectra shared among refProts were counted in the total spectral count of each protein. Spectra assigned to altProts were counted only if unique to the protein or shared with another altProt. Spectra shared between an altProt and at least one refProt were given to the refProt. RefProt spectral counts were combined by gene following the methodology of the original study; however, it was necessary to keep altProts separate as many are encoded by genes that already contain a refProt or other altProts.

Interactions scoring

Following protein identifications, high confidence interacting proteins (HCIPs) were identified following the method outlined in the original study (Huttlin *et al*, 2015). Briefly, the CompPASS R package was first used to compute statistical metrics (weighted D-score, Z score, and entropy) of prey identification based on peptide spectrum match (PSM) counts. The results from CompPASS were then used to build a vector of 9 features (as described in (Huttlin *et al*, 2015)) for each candidate bait-prey pair which were passed to a Naive Bayes classifier (CompPASS Plus) tasked with the discrimination of HCIP from background identifications. The original study also included a class for wrong identification, but since decoy information was unavailable and because our approach employs a FDR three orders of magnitudes lower in the identification step, a third class was not deemed necessary. The classifier was trained in cross-validation fashion using 96 well plate batches as splits and protein-protein interactions from the original study as target labels for true interactors.

Threshold selection was implemented considering the Jaccard overlap (equation i), recall (equation ii), precision and F1 score (equation iv) metrics between networks resulting from the re-analysis and the original study. The main differences between the OpenProt derived re-analysis and BioPlex 2.0 lie in the total spectral counts resulting from the use of different search algorithms and more stringent FDR. It was thus important to tune model threshold selection to maximally reproduce results from the original study (Figure EV1B). A threshold of 0.045 was selected as it compromised well between optimal Jaccard overlap, F score, and precision (Fig EV1A).

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad \text{(i)}$$

$$precision = \frac{|A \cap B|}{|A|} \qquad \text{(ii)}$$

$$recall = \frac{|A \cap B|}{|B|} \qquad \text{(iii)}$$

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall} \qquad \text{(iv)}$$

A: set of OpenProt derived protein-protein interactions

B: set of BioPlex 2.0 protein-protein interactions

Network assembly and structural analysis

Bait-prey pairs classified as HCIP were combined into an undirected network using genes to represent refProt nodes and OpenProt protein accessions to represent altProt nodes. The Networkx 2.5 Python package was used for network assembly and all network metrics calculations.

The power law fit to the degree distribution was computed with the discreet maximum likelihood estimator described by (Clauset *et al*, 2009).

A list of known protein complexes from CORUM 3.0 (Giurgiu *et al*, 2019) (core complexes, downloaded March 2020) was mapped onto the resulting network to assess the validity of identified interactions (Table EV3). Only complexes in which at least two subunits corresponded to baits present in the network were selected for downstream analyses. The portion of subunits identified in the direct neighbourhood of baits was computed for each complex.

Patterns of interactions involving altProt and refProts

We aimed to assess the relationship between pseudogene-derived altProts and their corresponding refProts from parental genes, in terms of their sequence similarity and their degrees of separation in the network. Parent genes of pseudogenes were selected via the psiCUBE resource (Sisu *et al*, 2014) combined with manual curation using Ensembl. Needleman Wunch global alignment algorithm (with BLOSUM62 matrix) as implemented by the sciki-bio Python package (version 0.5.5) was used as a similarity measure between protein sequences.

To assess degrees of separation, shortest path lengths were computed both for altProt-refProt pairs of pseudogene-parental gene and altProt-refProt pairs encoded by the same gene. For the former, when the refProt was not present in the network, or when no path could be computed between nodes, the shortest path length was computed using a mapping of either the BioPlex 2.0 or BIOGRID networks (Stark *et al*, 2006).

Community detection via clustering

A Python implementation of the markov clustering (MCL) algorithm (https://github.com/GuyAllard/markov_clustering) was used to partition the network into clusters of proteins (Enright *et al*, 2002). Various values of the inflation parameter between 1.5 and 2.5 were attempted and, similarly to the original study, a value of 2.0 was selected as it compared favorably with known protein complexes. Only clusters of 3 proteins or higher were retained yielding a total of 1045 clusters. Connections between clusters were determined by calculating enrichment of links between proteins in pairs of clusters using a hypergeometric test with alpha value set to <0.05 and a Benjamini-Hochberg corrected FDR of 1 %. A total of 266 pairs of clusters were found to be significantly connected.

Disease association

A list of 32,375 disease-gene associations curated by DisGeNET (downloaded March 2020) was mapped onto the network of 1045 protein communities. A disease was associated with a cluster when it was deemed enriched in genes associated with the disease as calculated by hypergeometric testing, with alpha value set to <0.01 Benjamini-Hochberg corrected FDR of 1 %.

Gene Ontology Enrichment

Gene Ontology term enrichments for both altProt second neighborhoods and protein clusters were computed using the GOAtools Python package (version 1.0.2). Count propagation to parental terms was set to true, alpha value to 0.05, with a Benjamini-Hochberg corrected FDR of 1 %.

Cloning and antibodies

All nucleotide sequences were generated by the Bio Basic Gene Synthesis service, except for pcDNA3-FLAG-FADD, a kind gift from Jaewhan Song (Addgene plasmid # 78802 ; http://n2t.net/addgene:78802 ; RRID:Addgene_78802). IP_117582, IP_624363, and IP_762813 were all tagged with 2 FLAG (DYKDDDDKDYKDDDDK) at their C-terminal. IP_198808 was tagged with eGFP at its C-terminal. All altProt coding sequences were subcloned into a pcDNA3.1- plasmid. The coding sequences of RPL18, eEF1A1 and PHB were derived from their canonical transcript (NM_000979.3, NM_001402.6, NM_001281496.1 respectively). RPL18 and PHB were tagged with eGFP at their C-terminal and eEF1A1 was tagged with eGFP at its N-terminal. All refProt coding sequences were subcloned into a pcDNA3.1- plasmid.

Cell culture, transfections and immunofluorescence

HEK293 and HeLa cultured cells were routinely tested negative for mycoplasma contamination (ATCC 30–1012K). Transfections, immunofluorescence, confocal analyses were carried out as previously described (Brunet *et al*, 2020a). Briefly, transfection was carried with jetPRIME®, DNA and siRNA transfection reagents (VWR) according to the manufacturer’s protocol. To note, only 0.1 μg of pEGFP DNA versus 3 μg IP_198808-GFP was used for transfection in 100 mm petri dishes to compensate for its higher transfection and expression efficiency. Cells were fixed in 4 % paraformaldehyde for 20 mins at 4°C, solubilized in 1 % Triton for 5 mins and incubated in blocking solution (10 % NGS in PBS) for 20 mins. The primary antibodies were diluted in the blocking solution as follows: anti-Flag (Sigma, F1804) 1/1000. The secondary antibodies were diluted in the blocking solution as follows: anti-mouse Alexa 647 (Cell signaling 4410S) 1/1000. All images were taken on a Leica TCS SP8 STED 3X confocal microscope.

Affinity Purification and western blots

Immunoprecipitation experiments via GFP-Trap (ChromoTek, Germany) were carried out as previously described (Samandi *et al*, 2017), while experiments via Anti-FLAG® M2 Magnetic Beads (M8823, Sigma) were conducted according to the manufacturer's protocol with minor modifications. Briefly, HEK293 cells were lysed in the lysis buffer (150 mM NaCl, 50 mM Tris pH 7.5, 1 % Triton, 1 x EDTA-free Roche protease inhibitors) and incubated on ice for 30 mins prior to a double sonication at 12 % for 3 seconds each (1 min on ice between sonications). The cell lysates were centrifuged, the supernatant was isolated and the protein content was assessed using BCA assay (Pierce). Anti-FLAG beads were conditioned with the lysis buffer. 20 µL of beads were added to 1 mg of proteins at a final concentration of 1 mg/mL and incubated overnight at 4°C. Then, the beads were washed 5 times with the lysis buffer (twice with 800 µL and twice with 500µL) prior to elution in 45 µL of Laemmli buffer and boiled at 95°C for 5 min. For co-immunoproecipitation of PHB1-GFP and RPL18-GFP, stringent wash were done with modified lysis buffer (250 mM NaCl + 20 µg/ml peptide FLAG (F3290 Sigma)) prior to elution with 200µg/ml peptide FLAG. Eluates were loaded onto 10 % SDS-PAGE gels for western blotting of GFP and FLAG tagged proteins. 40 µg of input lysates were loaded into gels as inputs. Western blots were carried out as previously described (Brunet *et al*, 2020a). The primary antibodies were diluted as follows: anti-Flag (Sigma, F7425) 1/1000 and anti-GFP (Santa Cruz, sc-9996) 1/8000. The secondary antibodies were diluted as follows: anti-mouse HRP (Santa Cruz sc-516102) 1/10000 and anti-rabbit HRP (Cell signaling 7074S) 1/10000.

Affinity Purification Mass Spectrometry (AP-MS)

For interactome analysis by mass spectrometry, HEK293 cells at a 70 % confluence were transfected with GFP-tagged PHB or with FLAG-tagged PHBP19 (IP_762813). 24h after transfection, cells were rinsed twice with PBS, and lysed in the AP lysis buffer (150 mM NaCl, 50 mM Tris-HCl and 1 % Triton). Protein concentration was evaluated with a BCA dosage and 1 mg of total protein was incubated at 4 °C for 4 hours with agarose GFP beads (ChromoTek, Germany) for PHB-GFP or with magnetic FLAG beads (Sigma, M8823) for IP_762813-FLAG. The beads were pre-conditioned with the AP lysis buffer. The beads were then washed twice with 1 mL of AP lysis buffer, and 5 times with 5 mL of 20 mM NH4HCO3 (ABC). Proteins were eluted and reduced from the beads using 10 mM DTT (15 mins at 55 °C), and then treated with 20 mM IAA (1 hour at room temperature in the dark). Proteins were digested overnight by adding 1 µg of trypsin (Promega, Madison, Wisconsin) in 100 µL ABC at 37 °C overnight. Digestion was quenched using 1 % formic acid and the supernatant was collected. Beads were washed once with acetonitrile/water/formic acid (1/1/0.01 v/v) and pooled with supernatant. Peptides were dried with a speedvac, desalted using a C18 Zip-Tip (Millipore Sigma, Etobicoke, Ontario, Canada) and resuspended into 30 µl of 1 % formic acid in water prior to mass spectrometry analysis.

Mass spectrometry analysis of in-house affinity purifications

Peptides were separated in a PepMap C18 nano column (75 µm × 50 cm, Thermo Fisher Scientific). The setup used a 0–35 % gradient (0–215 min) of 90 % acetonitrile, 0.1 % formic acid at a flow rate of 200 nL/min followed by acetonitrile wash and column re-equilibration for a total gradient duration of 4 h with a RSLC Ultimate 3000 (Thermo Fisher Scientific, Dionex). Peptides were sprayed using an EASYSpray source (Thermo Fisher Scientific) at 2 kV coupled to a quadrupole-Orbitrap (QExactive, Thermo Fisher Scientific) mass spectrometer. Full-MS spectra within a m/z 350–1600 mass range at 70,000 resolution were acquired with an automatic gain control (AGC) target of 1e6 and a maximum accumulation time (maximum IT) of 20 ms.

Fragmentation (MS/MS) of the top ten ions detected in the Full-MS scan at 17,500 resolution, AGC target of 5e5, a maximum IT of 60 ms with a fixed first mass of 50 within a 3 m/z isolation window at a normalized collision energy (NCE) of 25. Dynamic exclusion was set to 40 s. Mass spectrometry RAW files were searched with the Andromeda search engine implemented in MaxQuant 1.6.9.0. The digestion mode was set at Trypsin/P with a maximum of two missed cleavages per peptides. Oxidation of methionine and acetylation of N-terminal were set as variable modifications, and carbamidomethylation of cysteine was set as fixed modification. Precursor and fragment tolerances were set at 4.5 and 20 ppm respectively. Files were searched using a target-decoy approach against UniprotKB (Homo sapiens, SwissProt, 2020-10 release) with the addition of IP_762813 sequence for a total of 20360 entries. The false discovery rate (FDR) was set at 1 % for peptide-spectrum-match, peptide and protein levels. Only proteins identified with at least two unique peptides were kept for downstream analyses.

Highly confident interacting proteins (HCIPs) scoring of in-house affinity purifications

Protein interactions were scored using the SAINT algorithm. For each AP-MS, experimental controls were used: GFP alone transfected cells for PHB-GFP AP and mock transfected cells for IP_762813-2F AP. For the PHB-GFP AP, controls from the Crapome repository (Mellacheruvu *et al*, 2013) corresponding to transient GFP-tag expression in HEK293 cells, pulled using camel agarose beads were used. These controls are: CC42, CC44, CC45, CC46, CC47, and CC48. For the IP_762813-FLAG AP, controls from the Crapome repository (Choi et al, 2011) corresponding to transient FLAG-tag expression in HEK293 cells, pulled using M2-magnetic beads were used. These controls are: CC55, CC56, CC57, CC58, CC59, CC60 and CC61. The fold-change over the experimental controls (FC_A), over the Crapome controls (FC_B) and the SAINT probability scores were calculated as follows. The FC_A was evaluated using the geometric mean of replicates and a stringent background estimation. The FC_B was evaluated using the geometric mean of replicates and a stringent background estimation. The SAINT score was calculated using SAINTexpress, using experimental controls and default parameters. Proteins with a SAINT score above 0.8, a FC_A and a FC_B above 1,5 were considered HCIPs.

Network visualisation of in-house affinity purifications

The network was built using Python scripts (version 3.7.3) and the Networkx package (version 2.4). The interactions from the STRING database were retrieved from their protein links downloadable file. Only interactions with a combined score above 750 were kept.

Data Availability

The datasets and computer code produced in this study are available in the following databases:

- Protein interaction AP-MS data for both IP_762813 and PHB1 in HEK293 cells were deposited to the ProteomeXchange Consortium via the PRIDE (Perez-Riverol *et al*, 2016) partner repository with the dataset identifier PXD022491.
- Jupyter notebooks containing the analyses are available in the GitHub repository created for this project (https://github.com/Seb-Leb/altProts_in_communities).

Acknowledgements

We thank the Gygi lab for providing mass spectrometry (MS) datasets and particularly Ed Huttlin for helpful email exchanges. XR, MSS and AAC are members of the Fonds de Recherche du Québec Santé (FRQS)-supported Centre de Recherche du Centre Hospitalier Universitaire de Sherbrooke. This research was supported by CIHR grants MOP-137056 and MOP-136962, and by a Canada Research Chair in Functional Proteomics and Discovery of Novel Proteins to X.R. We thank the team at Calcul Québec and Compute Canada for their support with the use of the supercomputer mp2 from Université de Sherbrooke. We thank Darel Hunting for critically reviewing the manuscript.

Author contributions

Conceptualization: XR, SL and MAB. Experiments in Fig 1-5, EV1, EV2, data visualization, all Tables: SL. Naive Bayes classifier and interaction scoring: AAC, MSS, SL. Experiments in Fig 6: AML, AD, AT, ABG, MAB and JFJ. Experiments in Fig EV3: MAB and JFJ. Writing_original draft: XR and SL. Writing_review&editing: AAC, JFJ, MAB, MSS, SL, SS. Resources, funding acquisition, project administration: XR. SS and MB initiated the project and mentored SL.

Conflict of interest

Authors report no conflict of interest.

References

Bergeron D, Lapointe C, Bissonnette C, Tremblay G, Motard J & Roucou X (2013) An out-of-frame overlapping reading frame in the ataxin-1 coding sequence encodes a novel ataxin-1 interacting protein. *J Biol Chem* 288: 21824–21835

Bianconi G & Barabási A-L (2001) Bose-Einstein Condensation in Complex Networks. *Phys Rev Lett* 86: 5632–5635

Brunet MA, Brunelle M, Lucier J-F, Delcourt V, Levesque M, Grenier F, Samandi S, Leblanc S, Aguilar J-D, Dufour P, *et al* (2019) OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucleic Acids Res* 47: D403–D410

Brunet MA, Jacques J-F, Nassari S, Tyzack GE, McGoldrick P, Zinman L, Jean S, Robertson J, Patani R & Roucou X (2020a) FUS gene is dual-coding with both proteins united in FUS-mediated toxicity. *bioRxiv*: 848580

Brunet MA, Leblanc S & Roucou X (2020b) Reconsidering proteomic diversity with functional investigation of small ORFs and alternative ORFs. *Exp Cell Res* 393: 112057

Brunet MA, Levesque SA, Hunting DJ, Cohen AA & Roucou X (2018) Recognition of the polycistronic nature of human genes is critical to understanding the genotype-phenotype relationship. *Genome Res*

Brunet MA, Lucier J-F, Levesque M, Leblanc S, Jacques J-F, Al-Saedi HRH, Guilloy N, Grenier F, Avino M, Fournier I, *et al* (2020c) OpenProt 2021: deeper functional annotation of the coding potential of eukaryotic genomes. *Nucleic Acids Res*

Caspary F & Séraphin B (1998) The yeast U2A'/U2B complex is required for pre-spliceosome formation. *EMBO J* 17: 6348–6358

Chen J, Brunner A-D, Cogan JZ, Nuñez JK, Fields AP, Adamson B, Itzhak DN, Li JY, Mann M, Leonetti MD, *et al* (2020) Pervasive functional translation of noncanonical human open reading frames. *Science* 367: 1140–1146

Chirico N, Vianelli A & Belshaw R (2010) Why genes overlap in viruses. *Proc Biol Sci* 277: 3809–3817

Choong W-K, Lih T-SM, Chen Y-J & Sung T-Y (2017) Decoding the Effect of Isobaric Substitutions on Identifying Missing Proteins and Variant Peptides in Human Proteome. *J Proteome Res* 16: 4415–4424

Clauset A, Shalizi CR & Newman MEJ (2009) Power-law distributions in empirical data. *SIAM Rev* 51: 661–703

Close P, Gillard M, Ladang A, Jiang Z, Papuga J, Hawkes N, Nguyen L, Chapelle J-P, Bouillenne F, Svejstrup J, *et al* (2012) DERP6 (ELP5) and C3ORF75 (ELP6) Regulate Tumorigenicity and Migration of Melanoma Cells as Subunits of Elongator. *J Biol Chem* 287: 32535–32545

Colell A, Green DR & Ricci J-E (2009) Novel roles for GAPDH in cell death and carcinogenesis. *Cell Death Differ* 16: 1573–1581

Delcourt V, Franck J, Leblanc E, Narducci F, Robin Y-M, Gimeno J-P, Quanico J, Wisztorski M, Kobeissy F, Jacques J-F, *et al* (2017) Combined Mass Spectrometry Imaging and Top-down Microproteomics Reveals Evidence of a Hidden Proteome in Ovarian Cancer. *EBioMedicine* 21: 55–64

Delcourt V, Staskevicius A, Salzet M, Fournier I & Roucou X (2018) 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: e1700058

Deutsch EW, Lane L, Overall CM, Bandeira N, Baker MS, Pineau C, Moritz RL, Corrales F, Orchard S, Van Eyk JE, *et al* (2019) Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 3.0. *J Proteome Res* 18: 4108–4116

Di Benedetto G, Zoccarato A, Lissandron V, Terrin A, Li X, Houslay MD, Baillie GS & Zaccolo M (2008) Protein kinase A type I and type II define distinct intracellular signaling compartments. *Circ Res* 103: 836–844

Dinger ME, Pang KC, Mercer TR & Mattick JS (2008) Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput Biol* 4: e1000176

Dm C & Js C (2004) Unimod: Protein modifications for mass spectrometry. *Proteomics* 4

Drazyk AM, Tan RYY, Tay J, Traylor M, Das T & Markus HS (2019) Encephalopathy in a Large Cohort of British Cerebral Autosomal Dominant Arteriopathy With Subcortical Infarcts and Leukoencephalopathy Patients. *Stroke* 50: 283–290

Dubois M-L, Meller A, Samandi S, Brunelle M, Frion J, Brunet MA, Toupin A, Beaudoin MC, Jacques J-F, Lévesque D, *et al* (2020) UBB pseudogene 4 encodes functional ubiquitin variants. *Nat Commun* 11: 1306

Ekman D, Light S, Björklund AK & Elofsson A (2006) What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae? *Genome Biol* 7: R45

Enright AJ, Van Dongen S & Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30: 1575–1584

Eyckerman S, Titeca K, Van Quickelberghe E, Cloots E, Verhee A, Samyn N, De Ceuninck L, Timmerman E, De Sutter D, Lievens S, *et al* (2016) Trapping mammalian protein complexes in viral particles. *Nat Commun* 7: 11416

Giurgiu M, Reinhard J, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C & Ruepp A (2019) CORUM: the comprehensive resource of mammalian protein complexes-2019. *Nucleic Acids Res* 47: D559–D563

Huttlin EL, Bruckner RJ, Navarrete-Perea J, Cannon JR, Baltier K, Gebreab F, Gygi MP, Thornock

A, Zarraga G, Tam S, *et al* (2020) Dual Proteome-scale Networks Reveal Cell-specific Remodeling of the Human Interactome. *bioRxiv*: 2020.01.19.905109

Huttlin EL, Bruckner RJ, Paulo JA, Cannon JR, Ting L, Baltier K, Colby G, Gebreab F, Gygi MP, Parzen H, *et al* (2017) Architecture of the human interactome defines protein communities and disease networks. *Nature* 545: 505–509

Huttlin EL, Ting L, Bruckner RJ, Gebreab F, Gygi MP, Szpyt J, Tam S, Zarraga G, Colby G, Baltier K, *et al* (2015) The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* 162: 425

Ingolia NT, Hussmann JA & Weissman JS (2019) Ribosome Profiling: Global Views of Translation. *Cold Spring Harb Perspect Biol* 11

Jeong H, Mason SP, Barabási A-L & Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature* 411: 41

Keskin O, Tuncbag N & Gursoy A (2016) Predicting Protein-Protein Interactions from the Molecular to the Proteome Level. *Chem Rev* 116: 4884–4909

Kim H-K, Bhattarai KR, Junjappa RP, Ahn JH, Pagire SH, Yoo HJ, Han J, Lee D, Kim K-W, Kim H-R, *et al* (2020) TMBIM6/BI-1 contributes to cancer progression through assembly with mTORC2 and AKT activation. *Nat Commun* 11: 4012

Klemke M, Kehlenbach RH & Huttner WB (2001) Two overlapping reading frames in a single exon encode interacting proteins—a novel way of gene usage. *EMBO J* 20: 3849–3860

Kovács IA, Luck K, Spirohn K, Wang Y, Pollis C, Schlabach S, Bian W, Kim D-K, Kishore N, Hao T, *et al* (2019) Network-based prediction of protein interactions. *Nat Commun* 10: 1240

Leblanc S & Brunet MA (2020) Modelling of pathogen-host systems using deeper ORF annotations and transcriptomics to inform proteomics analyses. *Comput Struct Biotechnol J* 18: 2836–2850

Liu X, Salokas K, Tamene F, Jiu Y, Weldatsadik RG, Öhman T & Varjosalo M (2018) An AP-MS- and BioID-compatible MAC-tag enables comprehensive mapping of protein interactions and subcellular localizations. *Nat Commun* 9: 1188

Liu Y-J, Zheng D, Balasubramanian S, Carriero N, Khurana E, Robilotto R & Gerstein MB (2009) Comprehensive analysis of the pseudogenes of glycolytic enzymes in vertebrates: the anomalously high number of GAPDH pseudogenes highlights a recent burst of retrotrans-positional activity. *BMC Genomics* 10: 480

Luck K, Sheynkman GM, Zhang I & Vidal M (2017) Proteome-Scale Human Interactomics. *Trends Biochem Sci* 42: 342–354

Ma J, Ward CC, Jungreis I, Slavoff SA, Schwaid AG, Neveu J, Budnik BA, Kellis M & Saghatelian A (2014) Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. *J Proteome Res* 13: 1757–1765

Mahesh PA (2013) Unravelling the role of ADAM 33 in asthma. *Indian J Med Res* 137: 447–450

Marchant A, Cisneros AF, Dubé AK, Gagnon-Arsenault I, Ascencio D, Jain H, Aubé S, Eberlein C, Evans-Yamamoto D, Yachie N, *et al* (2019) The role of structural pleiotropy and regulatory evolution in the retention of heteromers of paralogs. *eLife* 8

Mellacheruvu D, Wright Z, Couzens AL, Lambert J-P, St-Denis NA, Li T, Miteva YV, Hauri S, Sardiu ME, Low TY, *et al* (2013) The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat Methods* 10: 730–736

Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, Brown SD, Chang H-Y, El-Gebali S, Fraser MI, *et al* (2019) InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res* 47: D351–D360

Murphy G (2008) The ADAMs: signalling scissors in the tumour microenvironment. *Nat Rev Cancer* 8: 929–941

Nesvizhskii AI (2014) Proteogenomics: concepts, applications and computational strategies. *Nat Methods* 11: 1114–1125

Orr MW, Mao Y, Storz G & Qian S-B (2020) Alternative ORFs and small ORFs: shedding light on the dark proteome. *Nucleic Acids Res* 48: 1029–1042

Osman C, Merkwirth C & Langer T (2009) Prohibitins and the functional compartmentalization of mitochondrial membranes. *J Cell Sci* 122: 3823–3830

Pavesi A, Vianelli A, Chirico N, Bao Y, Blinkova O, Belshaw R, Firth A & Karlin D (2018) Overlapping genes and the proteins they encode differ significantly in their sequence composition from non-overlapping genes. *PLOS ONE* 13: e0202513

Peeters MKR & Menschaert G (2020) The hunt for sORFs: A multidisciplinary strategy. *Exp Cell Res* 391: 111923

Pereira-Leal JB, Levy ED, Kamp C & Teichmann SA (2007) Evolution of protein complexes by duplication of homomeric interactions. *Genome Biol* 8: R51

Perez-Riverol Y, Xu Q-W, Wang R, Uszkoreit J, Griss J, Sanchez A, Reisinger F, Csordas A, Ternent T, Del-Toro N, *et al* (2016) PRIDE Inspector Toolsuite: Moving Toward a Universal Visualization Tool for Proteomics Data Standard Formats and Quality Assessment of ProteomeXchange Datasets. *Mol Cell Proteomics MCP* 15: 305–317

Peterson G, Allen CR & Holling CS (1998) Ecological Resilience, Biodiversity, and Scale. *Ecosystems* 1: 6–18

Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F & Furlong LI (2020) The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res* 48: D845–D855

Reiss K & Saftig P (2009) The 'a disintegrin and metalloprotease' (ADAM) family of sheddases: physiological and cellular functions. *Semin Cell Dev Biol* 20: 126–137

Rolland T, Taşan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, Yi S, Lemmens I, Fontanillo C, Mosca R, *et al* (2014) A proteome-scale map of the human interactome network. *Cell* 159: 1212–1226

Ruggles KV, Krug K, Wang X, Clauser KR, Wang J, Payne SH, Fenyö D, Zhang B & Mani DR (2017) Methods, Tools and Current Perspectives in Proteogenomics. *Mol Cell Proteomics MCP* 16: 959–981

Samandi S, Roy AV, Delcourt V, Lucier J-F, Gagnon J, Beaudoin MC, Vanderperre B, Breton M-A, Motard J, Jacques J-F, *et al* (2017) Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. *eLife* 6

Short K, Wiradjaja F & Smyth I (2007) Let's stick together: the role of the Fras1 and Frem proteins in epidermal adhesion. *IUBMB Life* 59: 427–435

Siegel RM, Martin DA, Zheng L, Ng SY, Bertin J, Cohen J & Lenardo MJ (1998) Death-effector Filaments: Novel Cytoplasmic Structures that Recruit Caspases and Trigger Apoptosis. *J Cell Biol* 141: 1243–1253

Sirover MA (2012) Subcellular dynamics of multifunctional protein regulation: mechanisms of GAPDH intracellular translocation. *J Cell Biochem* 113: 2193–2200

Sisu C, Pei B, Leng J, Frankish A, Zhang Y, Balasubramanian S, Harte R, Wang D, Rutenberg-Schoenberg M, Clark W, *et al* (2014) Comparative analysis of pseudogenes across three phyla. *Proc Natl Acad Sci* 111: 13361–13366

Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, Levin JZ, Karger AD, Budnik BA, Rinn JL & Saghatelian A (2013) Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat Chem Biol* 9: 59–64

Smith TM, Tharakan A & Martin RK (2020) Targeting ADAM10 in Cancer and Autoimmunity. *Front Immunol* 11: 499

Sowa ME, Bennett EJ, Gygi SP & Harper JW (2009) Defining the human deubiquitinating enzyme interaction landscape. *Cell* 138: 389–403

Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A & Tyers M (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 34: D535-539

Ting YS, Egertson JD, Payne SH, Kim S, MacLean B, Käll L, Aebersold R, Smith RD, Noble WS & MacCoss MJ (2015) Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data. *Mol Cell Proteomics MCP* 14: 2301–2307

Tristan C, Shahani N, Sedlak TW & Sawa A (2011) The diverse functions of GAPDH: views from different subcellular compartments. *Cell Signal* 23: 317–323

Vanderperre B, Lucier J-F, Bissonnette C, Motard J, Tremblay G, Vanderperre S, Wisztorski M, Salzet M, Boisvert F-M & Roucou X (2013) Direct detection of alternative open reading frames translation products in human significantly expands the proteome. *PloS One* 8: e70698

Wagner A & Fell DA (2001) The small world inside large metabolic networks. *Proc R Soc Lond B Biol Sci* 268: 1803–1810

Wan C, Borgeson B, Phanse S, Tu F, Drew K, Clark G, Xiong X, Kagan O, Kwan J, Bezginov A, *et al* (2015) Panorama of ancient metazoan macromolecular complexes. *Nature* 525: 339–344

Wen B, Wang X & Zhang B (2019) PepQuery enables fast, accurate, and convenient proteomic validation of novel genomic alterations. *Genome Res* 29: 485–493

Wolfson RL, Chantranupong L, Wyant GA, Gu X, Orozco JM, Shen K, Condon KJ, Petri S, Kedir J, Scaria SM, *et al* (2017) KICSTOR recruits GATOR1 to the lysosome and is necessary for nutrients to regulate mTORC1. *Nature* 543: 438–442

Figure legends

Figure 1 - Analysis overview and identification of alternative proteins in the human interactome.

A-B The classical model of RNA transcript coding sequence annotation includes only one reference open reading frame (ORF) on mRNAs encoding a reference protein (refProt) and no functional ORF within ncRNAs (A), while the alternative translation model considers multiple proteins encoded in different reading frames in the same transcript including refProts and alternative proteins (altProt)(B).

C Our re-analysis pipeline of high throughput AP-MS experiments from BioPlex 2.0 employs stringent criteria to ensure confident identification of both protein detection and interaction detection. Of the 434 altProts initially identified in the dataset, 280 joined the network of protein interactions after filtration.

D AltProts are in general shorter than reference proteins. Boxes represent the inter quartile range (IQR) marked at the median and the whiskers are set at 1.5*IQR over and under the 25th and 75th percentiles.

E Identified altProts (295) were encoded by transcripts (455) of a variety of biotypes. 121 of identified altProts are encoded by transcripts of protein coding biotype, 136 by transcripts of pseudogenes, and 38 exclusively by transcripts of non-coding biotype (ncRNA).

F AltORFs found encoded by transcripts from genes of protein coding biotype are most often overlapping the canonical CDS or localized downstream in the 3’UTR. A significant fraction of altORFs also localize in ncRNAs of protein coding genes. CDS: coding region, UTR: untranslated region (non-coding).

G Orthology data across 10 species from OpenProt 1.6 for detected altProts.

Figure 2 - Interaction mapping and network features of protein-protein interactions.

A The largest component of the network assembled from the OpenProt based re-analysis of high throughput affinity purification mass spectrometry data from BioPlex 2.0.

B A venn diagram of bait-prey interactions identified with the OpenProt derived re-analysis, BioPlex 2.0, and BioPlex 3.0 shows a significant overlap despite the smaller overall size of the re-analysis results (due to stringent filtration). It should also be noted that alternative proteins were not present in the BioPlex 2.0 analytical pipeline which accounts for part of the gap in overlap.

C The degree distribution (distribution of node connectivity) follows a power law as demonstrated by a discrete maximum likelihood estimator fit. The great majority of proteins have a small number of connections while a few are highly connected (often called hubs).

D The distribution of degrees of separation between all protein pairs (i.e. the length of the shortest path between all pairs of proteins) indicates that the network fits small-world characteristics.

E Alternative proteins were found diffusely throughout the network and across the spectrum of eigenvector centrality (EVC) (dark lines). EVC is a relative score that indicates the degree of influence of nodes on the network; here, altProts display involvement in both influential and peripheral regions.

F Known protein complexes from the CORUM 3.0 resource (Giurgiu *et al*, 2019) were mapped onto the network. Subunit recovery rate confirms the overall validity of the interactions confidently identified by the pipeline. All CORUM core complexes for which at least two subunits appear as baits in the network were considered.

G Selected CORUM complexes are shown with the addition of altProts found in the interaction network of baited subunits. Black edges indicate detection in the re-analysis, grey edges indicate those only reported by CORUM.

Figure 3 - Specific features of protein-protein interactions involving preyed alternative proteins.

A Degree-sorted circular layout of the OpenProt derived full network separated by bait and preys. Direct neighbors and neighbors of neighbors (here called second neighborhood) were extracted for each altProt. Second neighborhoods of alternative proteins display a variety of topologies with some acting as bridges (iv, vi,vii,ix) and others embedded in interconnected regions (i-iii, v). Larger nodes represent the proteins for which the second neighborhood was extracted.

B Second neighborhood of the refProt ELP6 extracted from the network assembled without altProts (i) and with altProts (ii). Inclusion of altProts in the network revealed that ELP6 connects to 6 additional proteins through its interaction with altProt IP_688853. Larger nodes represent the proteins for which the second neighborhood was extracted.

C Detailed second neighborhood of two pseudogene-encoded altProts. (i) GAPDH refProt shows 9 altProt interactors encoded by pseudogenes of GAPDH. (ii) AltProt encoded by *PHBP19* seen in the neighborhood of the PHB refProt. Larger nodes represent the proteins for which the second neighborhood was extracted.

D (i) AltProt found in the direct interactome of corresponding refProt from parent genes display a wide array of sequence similarity to the refProt. Pairs of altProt-refProt from pairs of pseudogene-parental genes are slightly closer in the network if their Needleman-Wunch (NW) protein sequence global alignment score is higher.

(ii) The distribution of degrees of separation between altProt-refProt pairs of the same gene is bimodal with a sub-population (75 %) following a distribution similar to the full network (see Figure 2D), and the other placing altProts in the direct neighborhood of refProts from the same gene.

Figure 4 - Protein communities obtained via unsupervised community detection reveal new members

A Protein communities identified via the Markov clustering algorithm (Enright *et al*, 2002). A total of 1045 clusters and 266 connections between them were identified; however, here are shown only components of 3 clusters or more for brevity. Nodes represent protein clusters sized relative to the number of proteins. Connections between clusters were determined by calculating enrichment of links between proteins in pairs of clusters using a hypergeometric test with maximal alpha value of 0.05 and correction for multiple testing was applied with 1 % FDR. **B** Focus on selected clusters showing significant enrichment of gene ontology terms. Enrichment was computed against background of whole genome with alpha value set to <0.05 Benjamini-Hochberg corrected FDR of 1 %. BP: biological process, MF: molecular function, CC: cellular compartment.

Figure 5 - Communities of proteins with altProt members are associated to disease phenotypes

A Network of association between protein clusters (blue and red nodes) and diseases (yellow nodes) from DisGenNet. Gene-disease enrichment was computed for each pair of disease-cluster, and associations were deemed significant after hypergeometric test with alpha set to 0.01 and multiple testing correction set at maximum 1 % FDR.

B Disease-cluster associations counted by disease classification (altProt containing clusters as red bars, and refProt only clusters as blue bars) and sorted by portion of association involving a cluster with altProts (dark red bars).

C Focus on clusters with significant disease associations showing involvement of altProts. *ADAM10* is a gene associated with tumorigenesis and produces an altProt here detected as part of a cluster associated to neoplastic processes (i). Other cluster-disease associations include genetic connective tissue diseases involving a pair of proteins encoded by the same gene (ii) and a cluster comprising pseudogene derived altProts and parental gene refProt in association with another oncological pathology (iii). Cluster #133 (iv) highlights associations of a cluster to both rare and common diseases with a community of proteins located at the membrane.

Figure 6 – Experimental validation of refProt-altProt interactions.

A Validation of FADD and IP_198808 protein interaction encoded by a bicistronic gene. Left panel: Immunoblot of co-immunoprecipitation with GFP-trap sepharose beads performed on HEK293 lysates co-expressing Flag-FADD and IP_198808-GFP or GFP only. Right panel: confocal microscopy of HeLa cells co-transfected with IP_198808-GFP (green channel) and Flag-FADD construct immunostained with anti-Flag (red channel). r = Pearson's correlation. The associated Manders' Overlap Coefficients are respectively M1= 0.639 and M2 = 0.931.

B Validation of eEF1A1 and IP_624363 protein interaction encoded from a pseudogene/parental gene couple. Left panel: immunoblot of co-immunoprecipitation with Anti-FLAG magnetic beads performed on HEK293 lysates co-expressing GFP-eEF1A1 and IP_624363-Flag or pcDNA3.1 empty vector with IP_624363-Flag constructs. Right panel: confocal microscopy of HeLa cells co-transfected with GFP-eEF1A1 (green channel) and IP_624363-Flag constructs immunostained with anti-Flag (red channel). r = Pearson's correlation. The associated Manders' Overlap Coefficients are respectively M1= 0.814 and M2 = 0.954.

C Validation of PHB1 and IP_762813 protein interaction encoded by a pseudogene/parental gene couple. Left panel: immunoblot of co-immunoprecipitation with Anti-FLAG magnetic beads performed on HEK293 lysates co-expressing PHB1-GFP and IP_762813-Flag or pcDNA3.1 empty vector with IP_762813-Flag constructs. Right panel: Comparison of the interaction network of IP_762813-Flag (purple) and PHB1-GFP (blue) from independent affinity purification mass spectrometry (AP-MS) of both proteins. 3 independent AP-MS for each protein.

D Validation of RPL18 and IP_117582 protein interaction. Left panel: immunoblot of co-immunoprecipitation with Anti-FLAG magnetic beads performed on HEK293 lysates co-expressing RPL18-GFP and IP_117582-Flag or pcDNA3.1 empty vector with IP_117582-Flag constructs. Right panel: confocal microscopy of HeLa cells co-transfected with RPL18-GFP (green channel) and IP_117582-Flag constructs immunostained with anti-Flag (red channel). r = Pearson's correlation. The associated Manders' Overlap Coefficients are respectively M1= 0.993 and M2 = 0.972.

All western blots and confocal images are representative of at least 3 independent experiments.

Tables and their legends

Table 1 - Terminology definitions

ORF	Open Reading Frame: sequence of nucleotides bounded by start and stop codons potentially translated into protein by ribosomes.
refORF	Annotated ORF producing a known protein.
altORF	Unannotated ORF producing an unknown/unannotated protein. AltORFs can be found on messenger RNAs overlapping refORFs or in untranslated regions, or on non-coding RNAs.
refProt	Annotated protein product resulting from the translation of a refORF.
altProt	Unannotated protein product resulting from the translation of an altORF with no significant homology with any refProt from the same gene.
Novel isoform	Unannotated protein product resulting from the translation of an altORF with high homology to a refProt from the same gene.

Extended View Tables Footnotes

Table extended view 1 - ***Transcripts and detected altProts for which at least one peptide spectrum match was validated via PepQuery.***

[1] Transcript accessions in bold indicate the longest transcript (used downstream for refProt relative localization).

[2]Biotype that should be assigned given the evidence from the current re-analysis.

[3]If multiple ORFs are present on the transcript and overlap, the transcript is dual coding; if they are sequential the transcript is called bicistronic.

[4]Colored rows indicate pseudogene transcripts that are assigned a multi-coding type.

Table extended view 2 - Bait-prey pairs involving detected altProts

[1]A score of 1 indicates that the bait-prey pair constitutes an altProt interacting with the refProt of the same gene, with a shortest path lenght of 1.

[2]A score of 1 indicates that the bait-prey pair constitutes a pseudogene-encoded altProt interacting with the refProt of the corresponding parent gene, with a shortest path lenght of 1.

[3]Set of non-nested (2 aa margin) peptides uniquely mapping to the corresponding altProt.

Table extended view 3 – CORUM complexes

[1]Fraction of subunits recovered in the complex.

Table extended view 4 – altProts coded by pseudogenes for which corresponding parent genes are annotated in psiCUBE (see Materials and Methods)

[1] No path indicates that (1) for the pseudogene-encoded altProt, the parent gene-encoded refProt was not identified; or (2) that the altProt and the refProt are not part of the same component in the network.

Expanded View Figure legends

Expanded View 1 - Network assembly details

A Overlap of total proteins (nodes) in BioPlex 2.0 and OpenProt derived networks.

B Classifier performance across thresholds. Scores were computed using the BioPlex 2.0 network as ground truth.

C The overlap of unfiltered interactions between BioPlex 2.0 and the result of OpenProt 1.6 derived re-analysis was considerable (92 % of re-analysis candidate PPIs) (i). Upon filtration the overlap is still significant despite the marked smaller size of the OpenProt derived network (59 % of re-analysis PPIs).

D Detailed counts of protein and interaction identifications.

Expanded View 2 - Community detection details

A Full network of protein clusters. Connections between clusters are drawn if the count of links between their constituent proteins is deemed enriched via a hypergeometric test with alpha set to 0.01 and multiple testing correction set at maximum 1 % FDR.

B All proteins in the network were either part of a cluster or not and either an altProt or a refProt.

C Distribution of cluster sizes (count of proteins in clusters).

D Distribution of cluster connectivity (cluster degree i.e. number of connections a cluster has with other clusters).

Expanded View 3 - Validation details

A Validation of interaction between proteins FADD and IP_198808 encoded by the same mRNA. IP_198808 peptides iii, iv, and v were detected in re-analyses of both ViroTrap and BioPlex 2.0 AP-MS of FADD. Peptides i and ii were exclusively identified in ViroTrap and BioPlex 2.0 re-analyses respectively. Peptides spectra matches (PSMs) for i and v from the ViroTrap dataset were validated against unrestricted modifications of reference proteins using PepQuery.

B FADD network after re-analysis of ViroTrap mass spectrometry data including IP_198808 sequence in the database.

C Detailed view of the combined network from AP-MS experiments of PHB refProt and PHBP19 altProt.

D Alignment of IP_762813 altProt encoded by pseudogene PHBP19 and PHB1 refProt sequences based on amino acids using Clustalω with default settings. Blue shading indicates amino acid similarity. Unique peptides detected are underlined red.

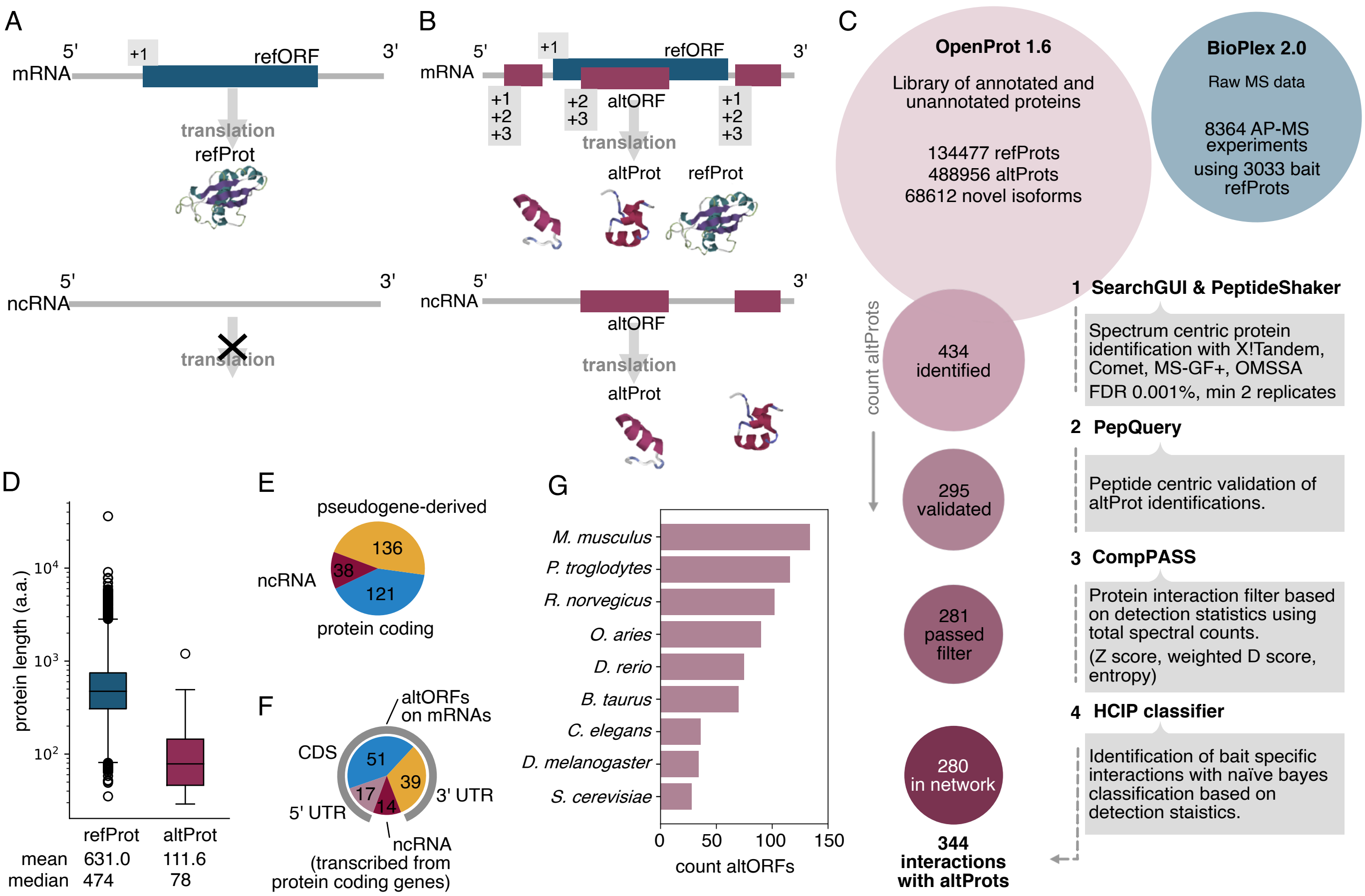


Figure 1 - Analysis overview and identification of alternative proteins in the human interactome.

A-B The classical model of RNA transcript coding sequence annotation includes only one reference open reading frame (ORF) on mRNAs encoding a reference protein (refProt) and no functional ORF within ncRNAs (A), while the alternative translation model considers multiple proteins encoded in different reading frames in the same transcript including refProts and alternative proteins (altProt)(B).

C Our re-analysis pipeline of high throughput AP-MS experiments from BioPlex 2.0 employs stringent criteria to ensure confident identification of both protein detection and interaction detection. Of the 434 altProts initially identified in the dataset, 280 joined the network of protein interactions after filtration.

D AltProts are in general shorter than reference proteins. Boxes represent the inter quartile range (IQR) marked at the median and the whiskers are set at 1.5*IQR over and under the 25th and 75th percentiles.

E Identified altProts (295) were encoded by transcripts (455) of a variety of biotypes. 110 of identified altProts are encoded by transcripts of protein coding biotype, 136 by transcripts of pseudogenes, and 58 exclusively by transcripts of non-coding biotype (ncRNA).

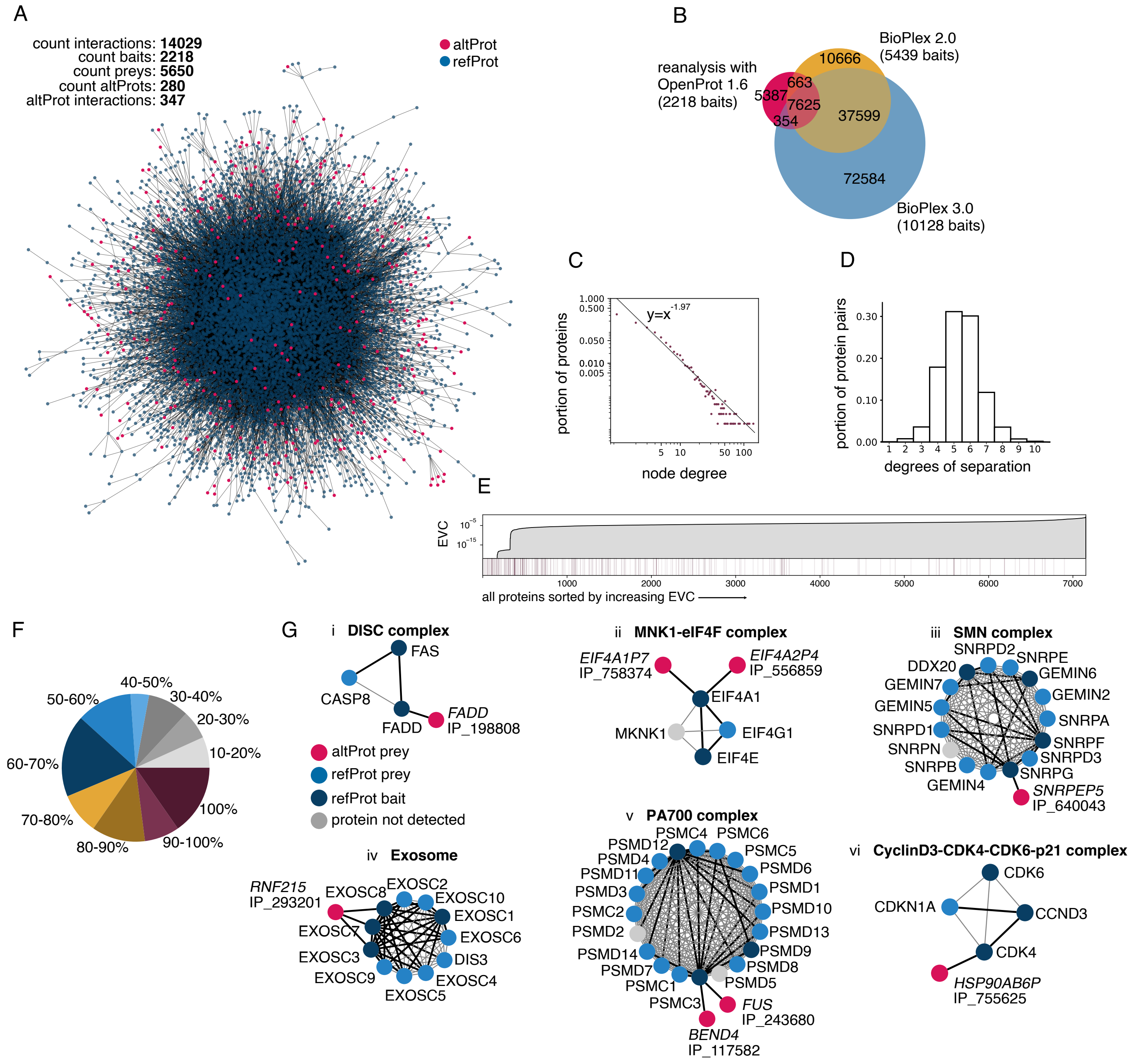


Figure 2 - Interaction mapping and network features of protein-protein interactions.

A The largest component of the network assembled from the OpenProt based re-analysis of high throughput affinity purification mass spectrometry data from BioPlex 2.0.

B A venn diagram of bait-prey interactions identified with the OpenProt derived re-analysis, BioPlex 2.0, and BioPlex 3.0 shows a significant overlap despite the smaller overall size of the re-analysis results (due to stringent filtration). It should also be noted that alternative proteins were not present in the BioPlex 2.0 analytical pipeline which accounts for part of the gap in overlap.

C The degree distribution (distribution of node connectivity) follows a power law as demonstrated by a discrete maximum likelihood estimator fit. The great majority of proteins have a small number of connections while a few are highly connected (often called hubs).

D The distribution of degrees of separation between all protein pairs (i.e. the length of the shortest path between all pairs of proteins) indicates that the network fits small-world characteristics.

E Alternative proteins were found diffusely throughout the network and across the spectrum of eigenvector centrality (EVC) (dark lines). EVC is a relative score that indicates the degree of influence of nodes on the network; here, altProts display involvement in both influential and peripheral regions.

F Known protein complexes from the CORUM 3.0 resource (Giurgiu et al, 2019) were mapped onto the network. Subunit recovery rate confirms the overall validity of the interactions confidently identified by the pipeline. All CORUM core complexes for which at least two subunits appear as baits in the network were considered.

G Selected CORUM complexes are shown with the addition of altProts found in the interaction network of baited subunits. Black edges indicate detection in the re-analysis, grey edges indicate those only reported by CORUM.

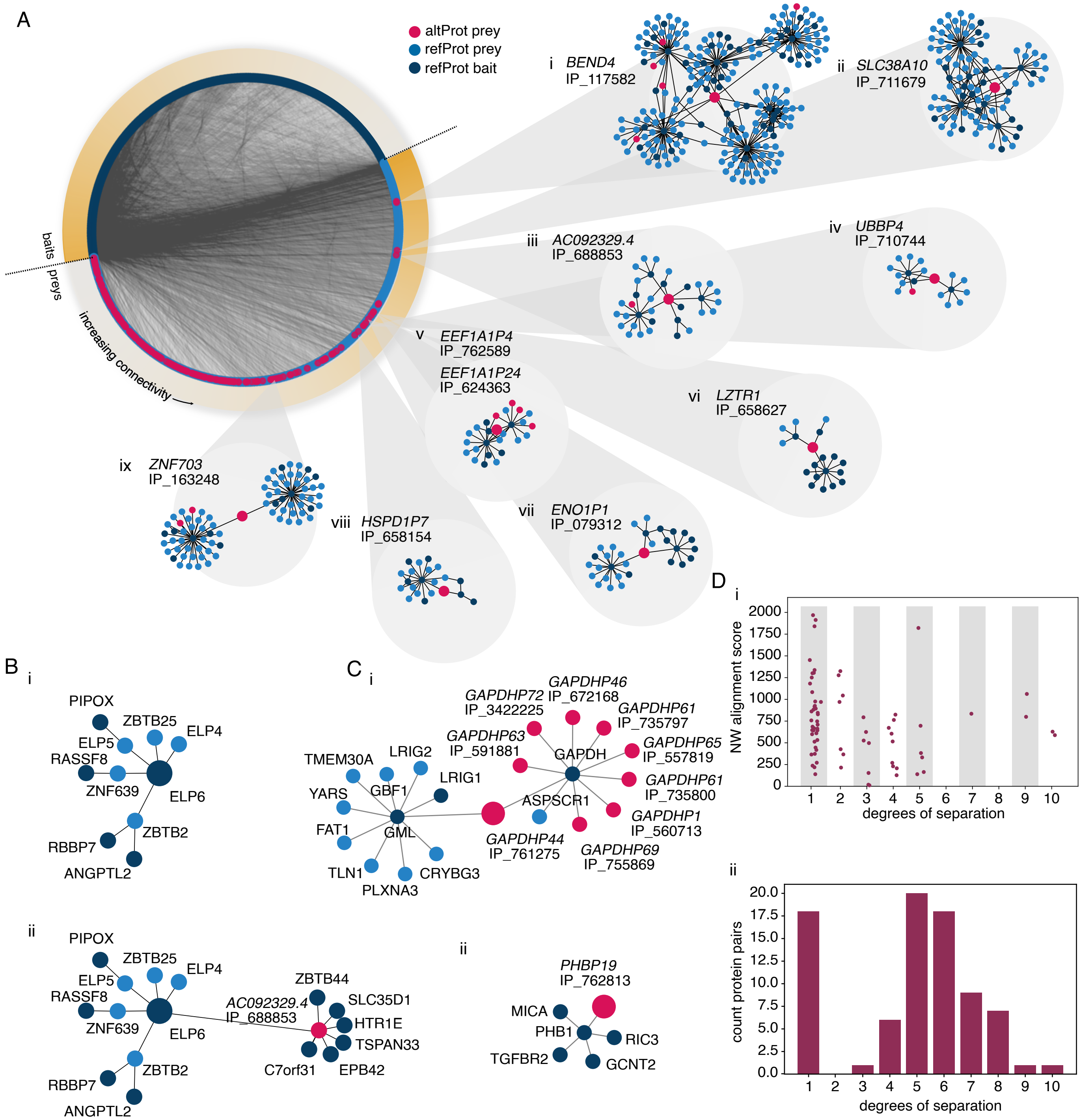


Figure 3 - Specific features of protein-protein interactions involving preyed alternative proteins.

A Degree-sorted circular layout of the OpenProt derived full network separated by bait and preys. Direct neighbors and neighbors of neighbors (here called second neighborhood) were extracted for each altProt. Second neighborhoods of alternative proteins display a variety of topologies with some acting as bridges (iv, vi,vi11,ix) and others embedded in interconnected regions (i-iii, v). Larger nodes represent the proteins for which the second neighborhood was extracted.

B Second neighborhood of the refProt ELP6 extracted from the network assembled without altProts (i) and with altProts (ii). Inclusion of altProts in the network revealed that ELP6 connects to 6 additional proteins through its interaction with altProt IP_688853. Larger nodes represent the proteins for which the second neighborhood was extracted.

C Detailed second neighbourhood of two pseudogene encoded altProts. (i) GAPDH refProt shows 9 altProt interactors encoded by pseudogenes of GAPDH. (ii) altProt encoded by PHBP19 seen in the neighborhood of the PHB refProt. Larger nodes represent the proteins for which the second neighborhood was extracted.

D (i) AltProt found in the direct interactome of corresponding refProt from parent genes display a wide array of sequence similarity to the refProt. Pairs of altProt-refProt from pairs of pseudogene-parental genes are slightly closer in the network if their Needleman-Wunch (NW) protein sequence global alignment score is higher. (ii) The distribution of degrees of separation between altProt-refProt pairs of the same gene is bimodal with a sub-population (75 %) following a distribution similar to the full network (see Figure 2D), and the other placing altProts in the direct neighborhood of refProts from the same gene.

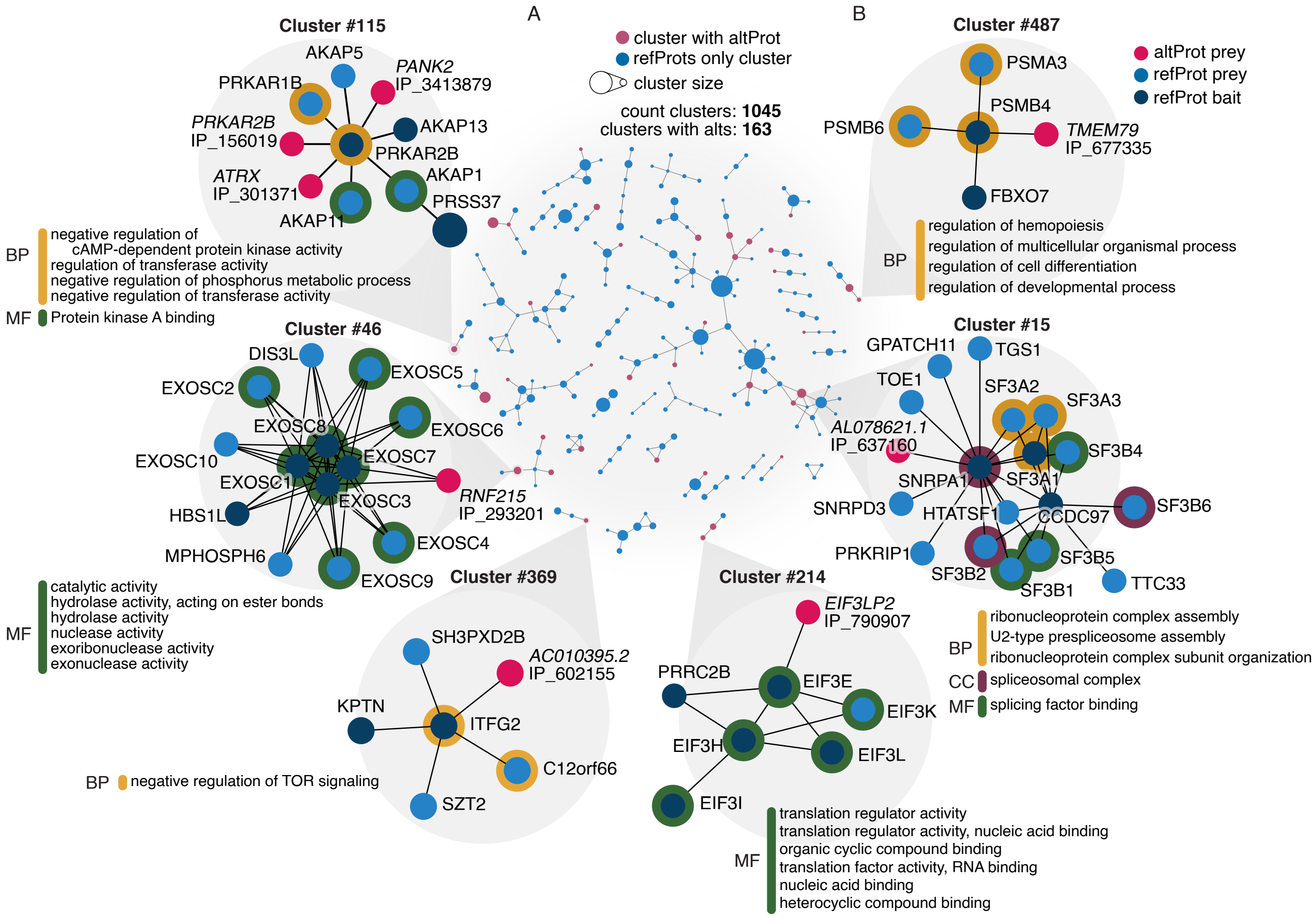


Figure 4 - Protein communities obtained via unsupervised community detection reveal new members

A Protein communities identified via the Markov clustering algorithm (Enright et al, 2002). A total of 1045 clusters and 266 connections between them were identified; however, here are shown only components of 3 clusters or more for brevity. Nodes represent protein clusters sized relative to the number of proteins. Connections between clusters were determined by calculating enrichment of links between proteins in pairs of clusters using a hypergeometric test with maximal alpha value of 0.05 and correction for multiple testing was applied with 1 % FDR.

B Focus on selected clusters showing significant enrichment of gene ontology terms. Enrichment was computed against background of whole genome with alpha value set to <0.05 Benjamini-Hochberg corrected FDR of 1 %. BP: biological process, MF: molecular function, CC: cellular compartment.

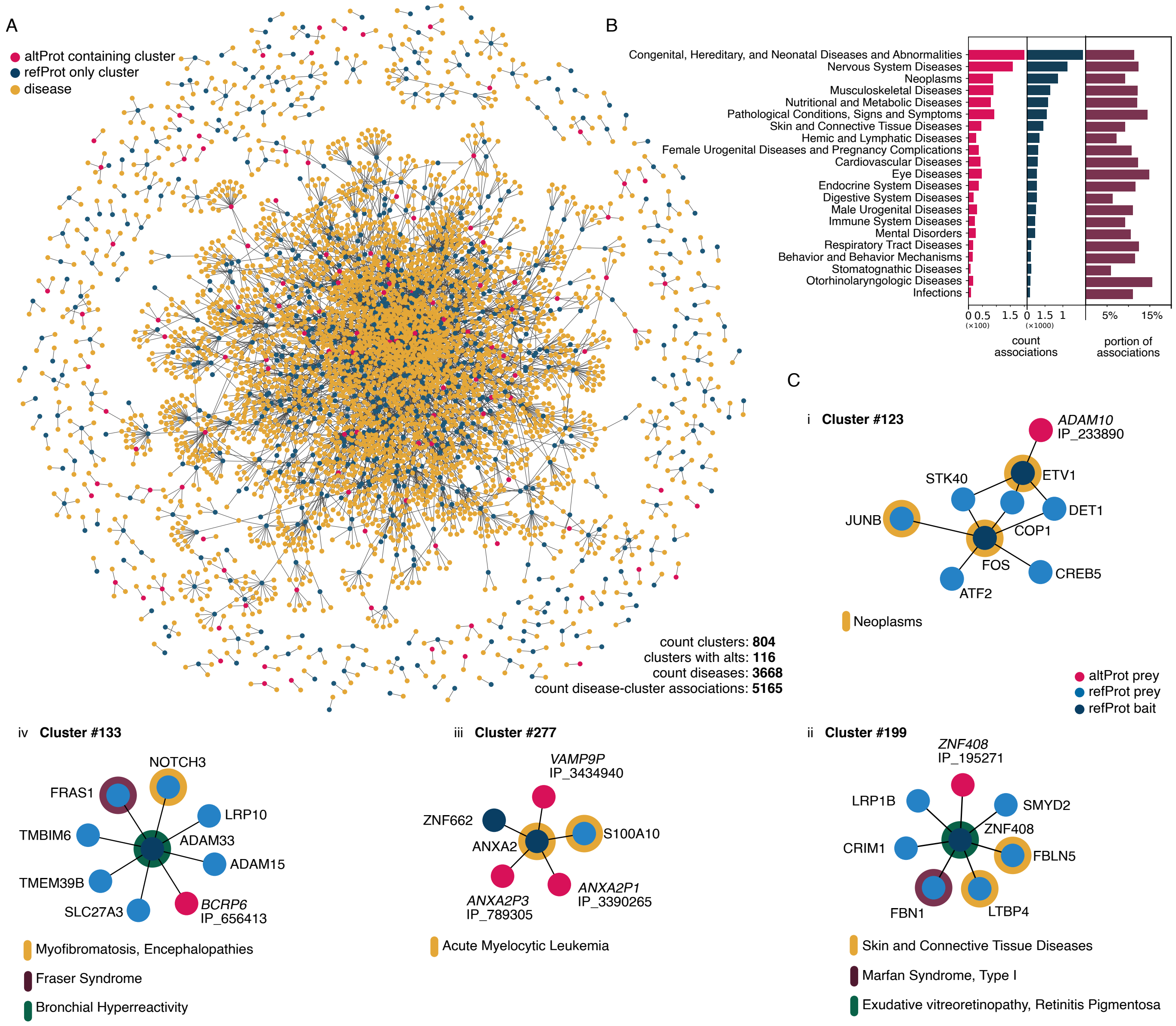


Figure 5 - Communities of proteins with altProt members are associated to disease phenotypes

A Network of association between protein clusters (blue and red nodes) and diseases (yellow nodes) from DisGenNet. Gene-disease enrichment was computed for each pair of disease-cluster, and associations were deemed significant after hypergeometric test with alpha set to 0.01 and multiple testing correction set at maximum 1 % FDR.

B Disease-cluster associations counted by disease classification (altProt containing clusters as red bars, and refProt only clusters as blue bars) and sorted by portion of association involving a cluster with altProts (dark red bars).

C Focus on clusters with significant disease associations showing involvement of altProts. ADAM10 is a gene associated with tumorigenesis and produces an altProt here detected as part of a cluster associated to neoplastic processes (i). Other cluster-disease associations include genetic connective tissue diseases involving a pair of proteins encoded by the same gene (ii) and a cluster comprising pseudogene derived altProts and parental gene refProt in association with another oncological pathology (iii). Cluster #133 (iv) highlights associations of a cluster to both rare and common diseases with a community of proteins located at the membrane.

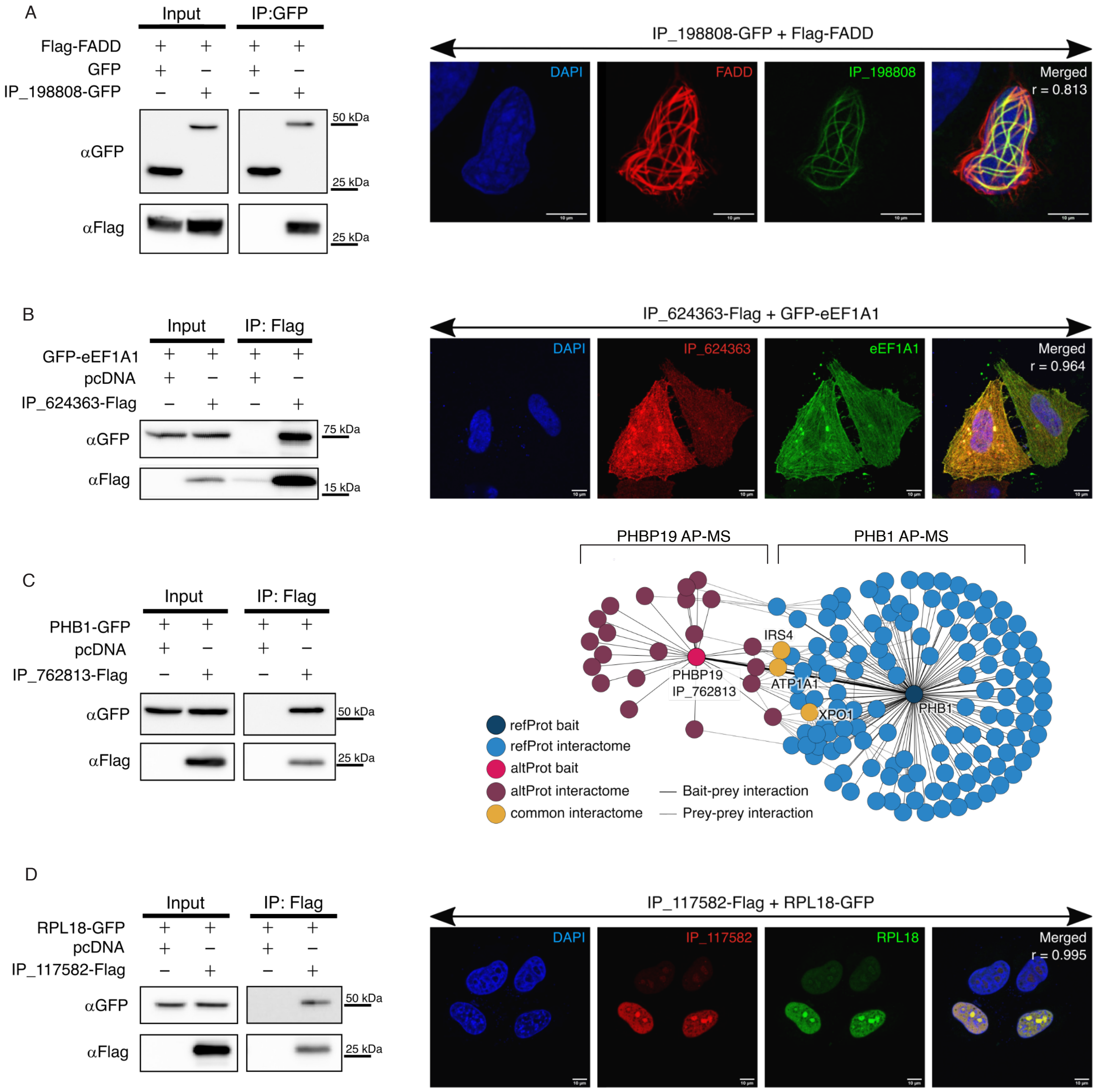


Figure 6 – Experimental validation of refProt-altProt interactions.

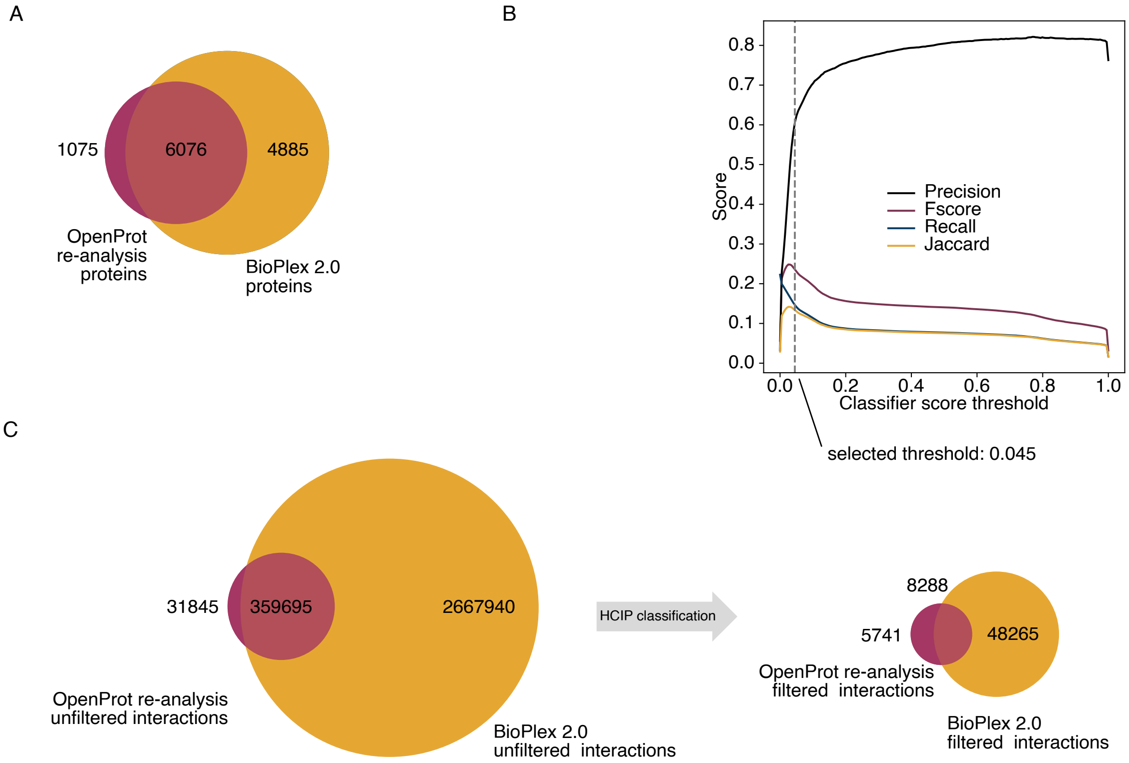
A Validation of FADD and IP_198808 protein interaction encoded by a bicistronic gene. Left panel: Immunoblot of co-immunoprecipitation with GFP-trap sepharose beads performed on HEK293 lysates co-expressing Flag-FADD and IP_198808-GFP or GFP only. Right panel: confocal microscopy of HeLa cells co-transfected with IP_198808-GFP and Flag-FADD construct immunostained with anti-GFP (green channel), anti-Flag (red channel). r = Pearson's correlation. The associated Manders Correlation Coefficients are respectively M1= 0.639 and M2 = 0.931.

B Validation of eEF1A1 and IP_624363 protein interaction encoded from a pseudogene/parental gene couple. Left panel: immunoblot of co-immunoprecipitation with Anti-FLAG magnetic beads performed on HEK293 lysates co-expressing GFP-eEF1A1 and IP_624363-Flag or pcDNA3.1 empty vector with IP_624363-Flag constructs. Right panel: confocal microscopy of HeLa cells co-transfected with GFP-eEF1A1 and IP_624363-Flag constructs immunostained with anti-GFP (green channel), anti-Flag (red channel). r = Pearson's correlation. The associated Manders Correlation Coefficients are respectively M1= 0.814 and M2 = 0.954.

C Validation of PHB1 and IP_762813 protein interaction encoded by a pseudogene/parental gene couple. Left panel: immunoblot of co-immunoprecipitation with Anti-FLAG magnetic beads performed on HEK293 lysates co-expressing PHB1-GFP and IP_762813-Flag or pcDNA3.1 empty vector with IP_762813-Flag constructs. Right panel: Comparison of the interaction network of IP_762813-Flag (purple) and PHB1-GFP (blue) from independent affinity purification mass spectrometry (AP-MS) of both proteins. 3 independent AP-MS for each protein.

D Validation of RPL18 and IP_117582 protein interaction. Left panel: immunoblot of co-immunoprecipitation with Anti-FLAG magnetic beads performed on HEK293 lysates co-expressing RPL18-GFP and IP_117582-Flag or pcDNA3.1 empty vector with IP_117582-Flag constructs. Right panel: confocal microscopy of HeLa cells co-transfected with RPL18-GFP and IP_117582-Flag constructs immunostained with anti-GFP (green channel), anti-Flag (red channel). r = Pearson's correlation. The associated Manders Correlation Coefficients are respectively M1= 0.993 and M2 = 0.972.

All western blots and confocal images are representative of at least 3 independent experiments.



D

Count unfiltered interactions	391540
Count baits in unfiltered interactions	3033
Count interactions filtered out by naïve bayes classifier	378151
Count HCIP baits	2218
Count unique ref preys	5369
Count unique alt preys	280
Count unique baits interacting with alts	292
Count interactions with alts	347
Count total refProt only interactions	13682

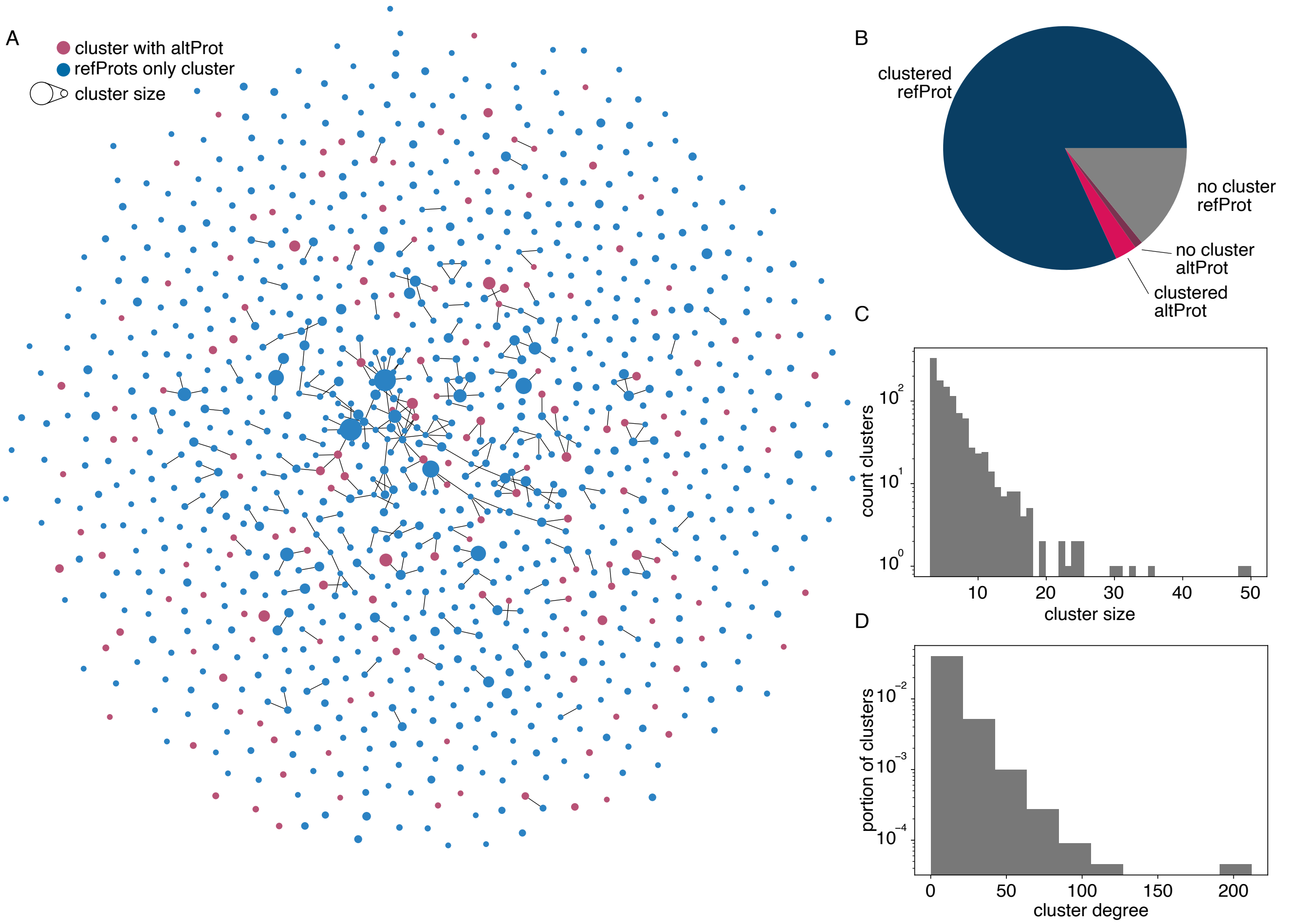
Expanded View 1 - Network assembly details

A Overlap of total proteins (nodes) in BioPlex 2.0 and OpenProt derived networks.

B Classifier performance across thresholds. Scores were computed using the BioPlex 2.0 network as ground truth.

C The overlap of unfiltered interactions between BioPlex 2.0 and the result of OpenProt 1.6 derived re-analysis was considerable (92 % of re-analysis candidate PPIs) (i). Upon filtration the overlap is still significant despite the marked smaller size of the OpenProt derived network (59 % of re-analysis PPIs).

D Detailed counts of protein and interaction identifications.



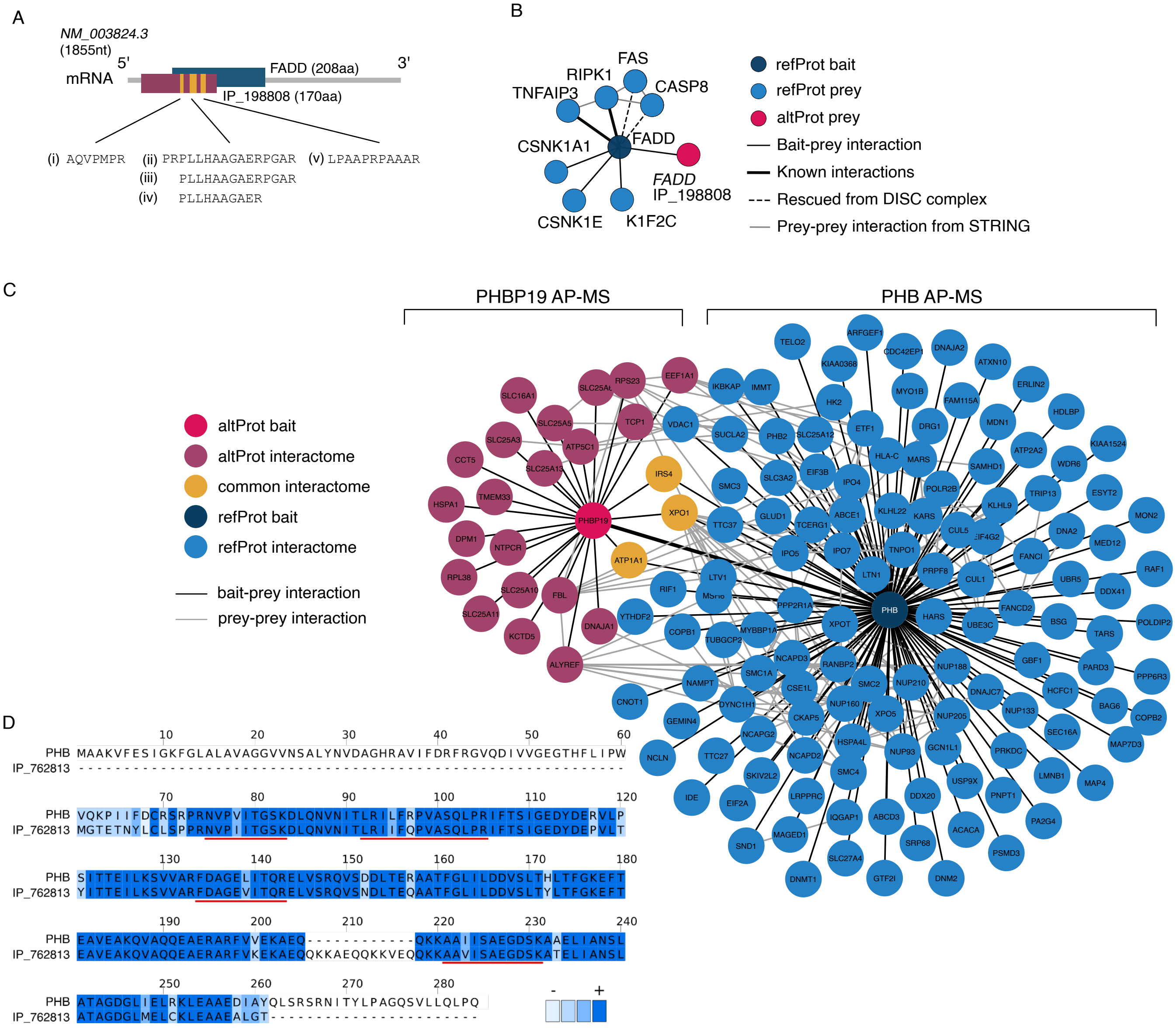
Expanded View 2 - Community detection details

A Full network of protein clusters. Connections between clusters are drawn if the count of links between their constituent proteins is deemed enriched via a hypergeometric test with alpha set to 0.01 and multiple testing correction set at maximum 1 % FDR.

B All proteins in the network were either part of a cluster or not and either an altProt or a refProt.

C Distribution of cluster sizes (count of proteins in clusters).

D Distribution of cluster connectivity (cluster degree i.e. number of connections a cluster has with other clusters).



Expanded View 3 - Validation details

A Validation of interaction between proteins FADD and IP_198808 encoded by the same mRNA. IP_198808 peptides iii, iv, and v were detected in re-analyses of both ViroTrap and BioPlex 2.0 AP-MS of FADD. Peptides i and ii were exclusively identified in ViroTrap and BioPlex 2.0 re-analyses respectively. Peptides spectra matches (PSMs) for i and v from the ViroTrap dataset were validated against unrestricted modifications of reference proteins using PepQuery.

B FADD network after re-analysis of ViroTrap mass spectrometry data including IP_198808 sequence in the database.

C Detailed view of the combined network from AP-MS experiments of PHB refProt and PHBP19 altProt.

D Alignment of IP_762813 altProt encoded by pseudogene PHBP19 and PHB1 refProt sequences based on amino acids using Clustalω with default settings. Blue shading indicates amino acid similarity. Unique peptides detected are underlined red.