COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Modelling of pathogen-host systems using deeper ORF annotations and transcriptomics to inform proteomics analyses

Sebastien Leblanc [a,b], Marie A. Brunet [a,b,*]

[a] Department of Biochemistry and Functional Genomics, Université de Sherbrooke, Sherbrooke, Québec, Canada
[b] PROTEO, Quebec Network for Research on Protein Function, Structure, and Engineering, Canada

ARTICLE INFO

Article history:
Received 29 April 2020
Received in revised form 7 October 2020
Accepted 8 October 2020
Available online 14 October 2020

Keywords:
Flavivirus
Zika
Alternative ORFs
Proteogenomics
Protein network
Proteome profiling

ABSTRACT

The Zika virus is a flavivirus that can cause fulminant outbreaks and lead to Guillain-Barré syndrome, microcephaly and fetal demise. Like other flaviviruses, the Zika virus is transmitted by mosquitoes and provokes neurological disorders. Despite its risk to public health, no antiviral nor vaccine are currently available. In the recent years, several studies have set to identify human host proteins interacting with Zika viral proteins to better understand its pathogenicity. Yet these studies used standard human protein sequence databases. Such databases rely on genome annotations, which enforce a minimal open reading frame (ORF) length criterion. An ever-increasing number of studies have demonstrated the shortcomings of such annotation, which overlooks thousands of functional ORFs. Here we show that the use of a customized database including currently non-annotated proteins led to the identification of 4 alternative proteins as interactors of the viral capsid and NS4A proteins. Furthermore, 12 alternative proteins were identified in the proteome profiling of Zika infected monocytes, one of which was significantly upregulated. This study presents a computational framework for the re-analysis of proteomics datasets to better investigate the viral-host protein interplays upon infection with the Zika virus.

1. Introduction

Flaviviruses are the most common cause of arthropod-borne human diseases, such as dengue and yellow fevers, Japanese encephalitis, West Nile and Zika virus infections [1]. Among these, the Zika virus (ZIKV) has known a surge of interest from the scientific community following recent outbreaks [2,3]. The first known outbreak took place on Yap Island in 2007, where 73% of the residents were infected and developed mild and short-lived symptoms [4]. The first association with more severe symptoms, such as Guillain-Barré syndrome, a neuro-inflammation of the peripheral nervous system, was discovered in the 2013–2014 outbreak in French Polynesia [5]. In late 2014, the virus spread through Brazil with a concomitant rise in Guillain-Barré syndrome and microcephaly in newborns [6]. Following these symptoms, the World Health Organization declared the Zika virus infection a Public Health Emergency of International Concern in 2016 [2].

ZIKV is an enveloped, positive-strand RNA virus. The primary source of infection is by mosquito bite, sexual and perinatal transmission [2]. ZIKV is blood-borne and can be detected 10 to 24 days post-infection in blood, and up to 60 days in semen [2]. However, the cellular and molecular intricacies of ZIKV pathogenesis, especially how the virus rewires the cellular pathways in favor of its replication, are not completely understood. Recently, several groups have used transcriptomics and proteomics methods to identify up- or down-regulated proteins during Zika virus infection, and to decipher its interaction with host proteins [3,7–12]. These studies are pivotal as they identify regulated pathways and protein–protein interactions between the host cell and the virus. However, and as reviewed in [3], such datasets are complex and findings are for the most part not reproduced across studies. First, comparisons across networks of viral-host protein interactions revealed a poor overlap [3]. These discrepancies certainly reflect

Abbreviations: ORF, open reading frame; ncRNA, non-coding RNA; sORF, small open reading frame; ZIKV, Zika virus; HCMV, human cytomegalovirus; MS, mass spectrometry; AP-MS, affinity-purification mass spectrometry; PSM, peptide spectrum match; altProt, alternative protein; HCIP, highly confident interacting proteins; DEP, differentially expressed proteins; LFQ, label free quantification; FPKM, fragments per kilobase of exon model per million reads mapped; FDR, false discovery rate.

[*] Corresponding author at: Department of Biochemistry, Université de Sherbrooke, 3201 Jean Mignault, Sherbrooke, Québec J1E 4K8, Canada.

E-mail address: marie.brunet@usherbrooke.ca (M.A. Brunet).

https://doi.org/10.1016/j.csbj.2020.10.010

experimental specificities such as the viral strain, the cellular background, the gene-delivery method and the experimental design. However, an often-overlooked source of variabilities is the analytical method used, both for the analysis and for the comparison. Furthermore, comparisons of RNA-seq and proteomics-based data in Zika-infected cells also revealed large differences. These indicate that the protein abundances cannot be confidently inferred from mRNA expression levels [13]. Although this is also true for physiological states, it is particularly relevant during infections as viruses modulate gene expression in part by altering mRNA processing, transport and translation rate [14–16].

Interestingly, viral infection changes the translational landscape of the host cell [15,16]. These translational changes can lead to the production of novel unannotated proteins. For example, in cells infected with the human cytomegalovirus (HCMV) several small open reading frames (sORFs) from the β2.7 allegedly non-coding RNA (ncRNA) were translated [17]. Sera samples from HCMV-positive blood donors revealed a strong response to these sORFs-encoded peptides, suggesting expression of these proteins and presentation on MHC molecules as functional antigens [18]. Additionally, a genetic screen study identified a novel protein, CYREN, encoded in a previously annotated ncRNA, among peptides important for resistance to retroviral infection [19]. More recently, a large-scale study identified 19 novel proteins, encoded by non-annotated ORFs in the human genome, as differentially regulated upon infection with flu lysates [20]. These findings highlight how current genome annotations shape the investigations of viral-host protein interplays.

The shortcomings of current annotations have been increasingly demonstrated throughout the last decade [21–23]. In order to minimize the identification of random non-functional ORFs, genome annotations enforce two arbitrary criteria : a minimal length of 100 codons and a single ORF per transcript, except for previously characterized examples [21]. However, these criteria are not supported by experimental evidence and led to an oversight of small and alternative ORFs [24–26]. Alternative ORFs are found in regions currently annotated as non-coding (ncRNAs, 5′ and 3′ UTRs of mRNAs) or overlapping an annotated ORF in a different reading frame [21,27]. To foster a more systematic exploration of alternative ORFs, several repositories have been published [28–30]. Such resources are pivotal for a deeper exploration of cellular events [21,26,31]. For example, in mass spectrometry (MS)-based proteomics, mass spectra must be matched to theoretical spectra generated from a database of possible proteins during the analysis. Thus, if a protein is absent from the database, it cannot be identified [21]. Differences in the database used for the analysis of MS data to elucidate the viral-host protein interactions can also be a source of discrepancies across studies. More importantly, it could prevent the identification of key interactors that are currently not annotated [17,19,20,32].

In this study, we propose a computational framework to investigate the importance of non-annotated proteins in the Zika infection through the re-analysis of published MS and RNA-seq datasets. We first address the comparison between such complex datasets and validate our MS analysis pipeline to identify highly confident interacting proteins (HCIPs). We then take advantage of published RNA-seq data to build a custom database using annotated proteins (UniProtKB, Ensembl and NCBI RefSeq) [33–35] and alternative proteins (OpenProt) [30]. Using this custom database, we queried an affinity-purification MS (AP-MS) dataset and a proteome profiling dataset for novel proteins important in the Zika virus pathogenicity. This computational framework demonstrates the importance of transcriptomic-informed proteomic analyses to identify changes in transcriptomic, translational and proteomic landscapes upon ZIKV infection.

2. Materials & Methods

2.1. Mass spectrometry-based proteomics

2.1.1. Datasets and databases

Affinity-purification mass spectrometry (AP-MS) data originated from a study by Shah and colleagues [8], and was retrieved from the Chorus repository (accession Project ID 1438). This dataset was generated using HEK293 cells infected with 2 strains of Zika virus (ZIKV French Polynesia 2013H/FP/2013 and ZIKV Uganda 1947 MR766). The experimental procedures have been described in the original study [8]. Briefly, Zika virus ORFs were tagged with a C-terminal 2xStrep II affinity tag and inserted into a pCDNA4_TO plasmid for expression in HEK293 cells. The authors used Strep-tactin beads to purify viral proteins and their interactors 40 h post-infection. The samples were digested with trypsin overnight before analysis on a Q-Exactive Plus Orbitrap (ThermoFisher) mass spectrometer.

Whole proteome label-free quantification data originated from a study by Ayala-Nunez and colleagues [9], and was retrieved from the PRIDE repository (accession PXD014002). This dataset was generated using human monocytes purified from 2 healthy blood donors. The experimental procedures have been described in the original study [9]. Briefly, cells were either infected with the Zika virus or not infected and quantitative proteomic profiling was performed 48 h after infection using a Q-Exactive Plus Orbitrap (ThermoFisher) mass spectrometer.

For annotated proteins, sequences were retrieved from the UniProtKB resource (*Homo sapiens* SwissProt, 2020–03). This fasta file contained 20,352 proteins. Protein sequences from the Zika virus were appended as described in [8]. The custom database was built using RNA-seq data to filter the whole OpenProt database (version 1.5: a non-redundant list of proteins from the full UniProtKB (03–2019), Ensembl (Jan 2019) and NCBI RefSeq (Jan 2019) databases as well as novel predicted proteins). Thus, our custom database contained 98,508 protein sequences. These included 13,048 proteins from the full UniProtKB database, 83 proteins from Ensembl 95 and 311 proteins from NCBI RefSeq not contained in the UniProtKB database, 4,469 novel isoforms from the OpenProt database (version 1.5, 2020–06), 80,573 alternative proteins from the OpenProt database, and 19 protein sequences derived from the RNA-seq data (see Section 2.3.2).

2.1.2. Mass spectrometry analysis pipeline

Raw AP-MS files were first converted to mgf files using the ThermoRawFileParser (version 1.2.0) [36]. The files were analysed using PeptideShaker software (version 1.16.42) [37] configured to use three search engines (X!Tandem, MS-GF+ and Comet) via SearchGUI (version 3.3.17) [38]. The decoy database was generated using reversed sequences. SearchGUI general parameters were set as follows: the fragment mass tolerance was set to 20 ppm and the precursor ion tolerance to 4.5 ppm; the enzyme was set to trypsin with a maximum of 2 missed cleavages; oxidation of methionine and acetylation of protein N-terminus were set as variable modifications, and carbamidomethylation of cysteine was set as fixed modification; a maximum of 5 modifications were allowed per peptide, a maximum charge of 7 + and minimal length of 7 amino acids. False discovery rates (FDR) were set to 1% at the peptide and protein level. Additionally, a novel protein was deemed confidently identified only if supported by at least one unique peptide. Thus, similarly to the OpenProt pipeline, the following peptide assignation rules were enforced: if a peptide was shared between 2 known proteins the spectrum was assigned to both and a protein group was created; if a peptide was shared between 2 novel proteins the spectrum was assigned to both, and a protein group was created; if the peptide is shared between known and novel proteins, the spectrum was only be assigned to the previously known proteins. Although biased against novel proteins, this approach ensured a certain robustness in their identification.

For quantitative proteome profiling, raw files were analysed with MaxQuant (version 1.6.0.16) [39]. The decoy database was generated using reversed sequences. The search parameters were set as follows: the fragment mass tolerance was set to 20 ppm for the first search and 5 ppm for the main search; MS/MS tolerance was set to 40 ppm; the enzyme was set to trypsin with a maximum of 1 missed cleavages; oxidation of methionine was set as variable modification, and carbamidomethylation of cysteine as fixed modification; label-free quantification (LFQ) was implemented using a minimal ratio count of 1 [40]; and the match between runs was enabled with a 2-min time window after retention-time alignment. The protein FDR was set to 1%, with the same peptide assignation rules as the AP-MS data analysis.

2.2. Protein interaction and quantification analyses

2.2.1. Protein interaction scoring with MiST and CompPASS

Protein-protein interaction candidates from the AP-MS analysis were scored using both MiST [41] and CompPASS [42] scoring algorithms. First, to control for data quality and reproducibility across replicates, peptide and protein counts per sample, as well as bait spectral counts were used to screen unreliable samples. At least three replicates for each bait were kept.

For MiST scoring, peptide spectrum match (PSM) counts were used as quantifying feature. The recommended weights relating to reproducibility (R), bait specificity (S) and abundance (A) of the interaction were used: R = 0.309, S = 0.686 and A = 0.006 [41,43]. Only proteins with a MiST score above 0.75 were considered when using MiST alone to score highly confident interacting proteins (HCIPs). For CompPASS scoring, PSM counts were also used as quantifying feature. When using CompPASS alone to score HCIPs, proteins with a weighted D-score in the top 5% were considered. To evaluate the performance of each HCIP scoring protocol, the proteins identified in the original study were used as target true positives [8]. Thus, the precision was calculated as: $\frac{Truepositive}{Truepositive+Falsepositive}$. The recall was calculated as: $\frac{Truepositive}{Truepositive+Falsenegative}$. The F1 score was calculated as: $2 \times \frac{Precision \times Recall}{Precision+Recall}$. The precision, recall and F1 score calculated for each approach is shown in Supplementary Fig. 1.

2.2.2. Naive Bayes classification

To further optimize prey hits filtration (optimize the F1 score) and to capitalize on the optimization work from the original study [8], a Naive Bayes model was trained. Such classifiers have previously been published to identify HCIP in large datasets [44]. The model was trained to distinguish HCIP from background proteins using interactions from the original study and another AP-MS experiment with ZIKV viral proteins in HEK293 cells [11] as positive labels. MiST score and the CompPASS Z-score were used as features. We used the CompPASS Z-score as it complemented the MiST score whereas the weighted D-score correlated with it. To avoid over-fitting, the set of candidate interactions was randomly divided into 10 sets and each was scored with a classifier trained on data from the other nine sets in a 10-fold cross-validation manner. Thus, 10 models were trained, each blinded to the set of interactions it was aimed to evaluate. The model assigned a score reflecting the likelihood of a specific interaction for each bait-prey pair. The precision, recall and F1 score were calculated as described above using interactions from the original study as labels for positive HCIPs. The threshold for the Naïve Bayes was selected by observing the behavior of the three metrics across a range of thresholds. The threshold that optimized F1 score while prioritizing recall over precision (Supplementary Fig. 1). The inherent lack of true negative examples in protein interaction datasets requires prioritizing recall over precision when selecting the classifier threshold. It is generally expected that the rate of false positives be over estimated due to this bias. Thus, a threshold of 0.45 was chosen for the analysis with the UniProt database, and 0.42 for the analysis with the custom database (Supplementary Fig. 2).

2.2.3. Protein quantification and differential expression analysis

LFQ intensities were reported by MaxQuant and used for quantification and differential expression analysis. To be included in the analysis, proteins needed be identified in all replicates (4 replicates) in at least one condition (4 conditions: 2 non-infected and 2 Zika-infected). For known proteins, a requirement for a minimum of two unique peptides was also enforced. For novel proteins, considering their smaller size (Fig. 3) and the bias inherent to the peptide assignation rules (see Section 2.1.2), a minimum of one unique peptide was used. Missing values were imputed using the ProStaR software (version 1.18.6) [45] as done in the original study. Briefly, the imputation was done differentially based on the nature of the missing values. We used a structured least square adaptative regression (SLSA mode) for partially observed values (POV), while a deterministic value (DetQuantile method – 1% quantile, multiplying factor of 1) was applied for values missing in the entire condition (MEC). The DetQuantile method was chosen for MEC as these values corresponded to proteins below the limit of detection in one condition, thus they could not be imputed based on values observed in other conditions [45]. No data distortion could be observed after imputation of missing values (Supplementary Fig. 6). Differential expression analysis was performed using a Limma-moderated *t*-test with a Benjamini-Hochberg correction for multiple comparisons. The FDR was set to 1% for downstream analyses.

2.3. Transcriptomics analysis

2.3.1. RNA-seq dataset

The RNA-seq data originates from a study by Tang and colleagues [10], and was retrieved from the NCBI Gene Expression Omnibus (GEO) repository (accession GSE78711). This dataset was generated using human cortical neural progenitor cells (hNPCs) that were either non-infected or infected with Zika virus (ZIKV Uganda 1947 MR766 strain) at low multiplicity of infection (MOI) (<0.1). The experimental procedures have been detailed in the original publication [10]. Cells were collected 56 h after infection. The dataset includes 2 replicates for each condition, and each was paired-end sequenced.

2.3.2. RNA-seq data analysis

We used the FastQC toolkit (version 0.11.8) with default parameters to filter low quality reads. The reads were trimmed using TrimGalore (version 0.6.4) set for paired reads with the default parameters except for the maximal number of N set to 5, and end clips set to 3. The reads were subsequently mapped to the human genome (GRCh38, Gencode v32, primary assembly) using the STAR software (version 2.7.3a) [46] with the default parameters except for the maximal number of mismatch set to 5, the maximal number of multi-mapping locations set to 10 and SAM primary flag set to all best scores. The outputs were ordered by genomic coordinates. Cufflinks (version 2.2.1) [47] was used for transcript assembly and evaluation of the transcript expression level (FPKM - fragments per kilobase of exon model per million reads mapped). For inclusion in the custom database, transcripts must be identified in both replicates of the Zika-infected condition and have a FPKM above 2.8 (see Fig. 3).

2.4. Novel protein predictions

2.4.1. OpenProt resource mining

Novel proteins were retrieved from the OpenProt resource [30]. We used version 1.5 of the OpenProt database for *Homo sapiens* (06–2020). These predictions were based on the GRCh38.p12 genome assembly and contain all ORFs, currently not present in annotations, longer than 30 codons and starting with an ATG. Subsequently, the corresponding proteins were classified as novel isoforms or alternative proteins based on their homology (or lack thereof) with the canonical protein from the same gene. The Fasta files containing alternative proteins and novel isoforms were downloaded using the downloads interface (www.openprot.org). For downstream analyses, the search interface was used to gather previous experimental evidence, conservation and prediction of functional domains for novel proteins.

2.4.2. ORF prediction on novel transcripts

For novel transcripts predicted by Cufflinks and with enough supporting evidence (detected in both replicates with a FPKM above 1.499), ORFs were predicted by *in silico* translation using the same criteria as those implemented by OpenProt: a minimal length of 30 codons and an ATG start. Each transcript was assigned an accession number (tx0000) preceded by the string "CUFF_zika" and each ORF from a given transcript was assigned a unique number. Each ORF was then searched within the OpenProt database. If the exact same ORF exists in the OpenProt database from another transcript from the same gene, the encoded protein was given the accession listed by OpenProt. When the ORF was not present in the OpenProt database (novel ORF), the resulting protein accession corresponded to the transcript accession concatenated to the ORF number to provide a unique identifier for each predicted ORF.

2.5. Network analyses

2.5.1. Network similarity measures

Similarity of networks can be measured at two levels: similarity of identified interacting proteins, and similarity of protein complexes identified. For the overlap of identified interacting proteins, the list of HCIPs for each network is compared using the Jaccard similarity index [48] which considers the ratio of size of the intersection of both sets to the size of the union. The significance of this overlap is then evaluated with a Fisher's exact test. A p-value below 0.05 was considered significant. To evaluate the similarities between protein–protein interaction networks from different analyses, we measured network characteristics: degree, local clustering coefficient, and shortest path length distributions. In a protein–protein interaction network, the nodes correspond to proteins and edges represent interactions. The degree of a node is defined as its number of connections. The local clustering coefficient relates to the interconnectivity of the neighborhood of a node with higher values indicating denser connectivity. The shortest path length is defined as the minimal number of edges required to connect two nodes on the network. These metrics were calculated for each protein (degree and clustering coefficient) or pair of proteins (shortest path).

2.5.2. Biological processes enrichment

Enrichment of biological processes was measured using the GOATOOLS Python package (version 1.0.2) [49], corresponding to the GO-term annotation version 1.2. The GOATOOLS package was run with the following parameters: alpha = 0.05, and count propagation to the parental terms set to true. The enrichment was calculated against the human proteome for each bait with a Benjamini-Hochberg correction to adjust for multiple comparisons. The FDR was set to 5%.

2.5.3. Edge mapping from STRING database

Known interactions between proteins within the network were retrieved from the STRING database (version 11.0). Only interactions with a combined score of 0.75 (considered highly confident) were retrieved.

2.5.4. CORUM complexes overlap

The similarity of networks from different analyses was also considered via the retrieval of CORUM complexes. Briefly, three sets of complexes with at least one protein identified in the original study [8], our analysis, or both, were retrieved from the CORUM database (version 3.0). The number of subunits identified was then calculated for each identified complex.

2.5.5. Code and data availability

The AP-MS data was retrieved from the Chorus repository (www.chorusproject.org, Project ID 1438). The quantitative proteome profiling was retrieved from the PRIDE repository (https://www.ebi.ac.uk/pride/archive/, PXD014002). The RNA-seq data was retrieved from the GEO database (https://www.ncbi.nlm.nih.gov/geo/, GSE78711). The human genome assembly was retrieved from the Gencode server (https://www.gencodegenes.org/, v32 primary assembly). Alternative protein and novel isoform sequences were retrieved from the OpenProt resource (https://openprot.org/, version 1.5 06–2020). All scripts were written with Python 3.7 and Networkx 2.4, and are available in GitHub (https://github.com/MAB-Lab/Zika_Project).

3. Results

3.1. Assessing reproducibility from AP-MS data re-analysis

The most-used technique to build a network of protein interactions is the AP-MS technique [50,51]. "Bait" proteins are expressed with a tag which allows purification and subsequent identification of interacting proteins ("preys") by MS [3]. Here, we retrieved a publicly available AP-MS dataset of Zika viral proteins in human cells [8]. As in the original study, we used the SwissProt database concatenated with sequences of the Zika proteins as a reference protein database. However, our analytical pipeline used a different strategy for the protein identification and interaction scoring strategies (Fig. 1A). For the protein identification step, we used the SearchGUI and PetideShaker softwares to take advantage of multiple search engines [52]. We identified 2,490 unique potential interactors, from which 1,762 (70.8%) were also identified in the original study (Suppl. Fig. 1). Out of the 277 proteins found in the original study but not in our analysis, 4 were not present in the SwissProt database we used (03–2020) and 196 were found within protein groups but were not selected as they were not supported by unique peptides in our hands (Suppl. Fig. 1). These numbers demonstrate the importance of databases and protein group handling in AP-MS analyses.

To identify highly confident interacting proteins (HCIPs) from background interactions, we tested 4 filtering protocols: the MiST algorithm alone [51], the CompPASS algorithm alone [42], the intersection of MiST and CompPASS calls, and the use of a Naive Bayes classifier building on MiST and CompPASS scores. Using MiST alone did not produce enough filtering of our dataset (1,501 HCIPs called, corresponding to the top 12.1% of interactions kept), where CompPASS alone sufficiently filtered but yielded a poor recall (28%) and F1 score (25%) (Suppl. Fig. 1). The intersection

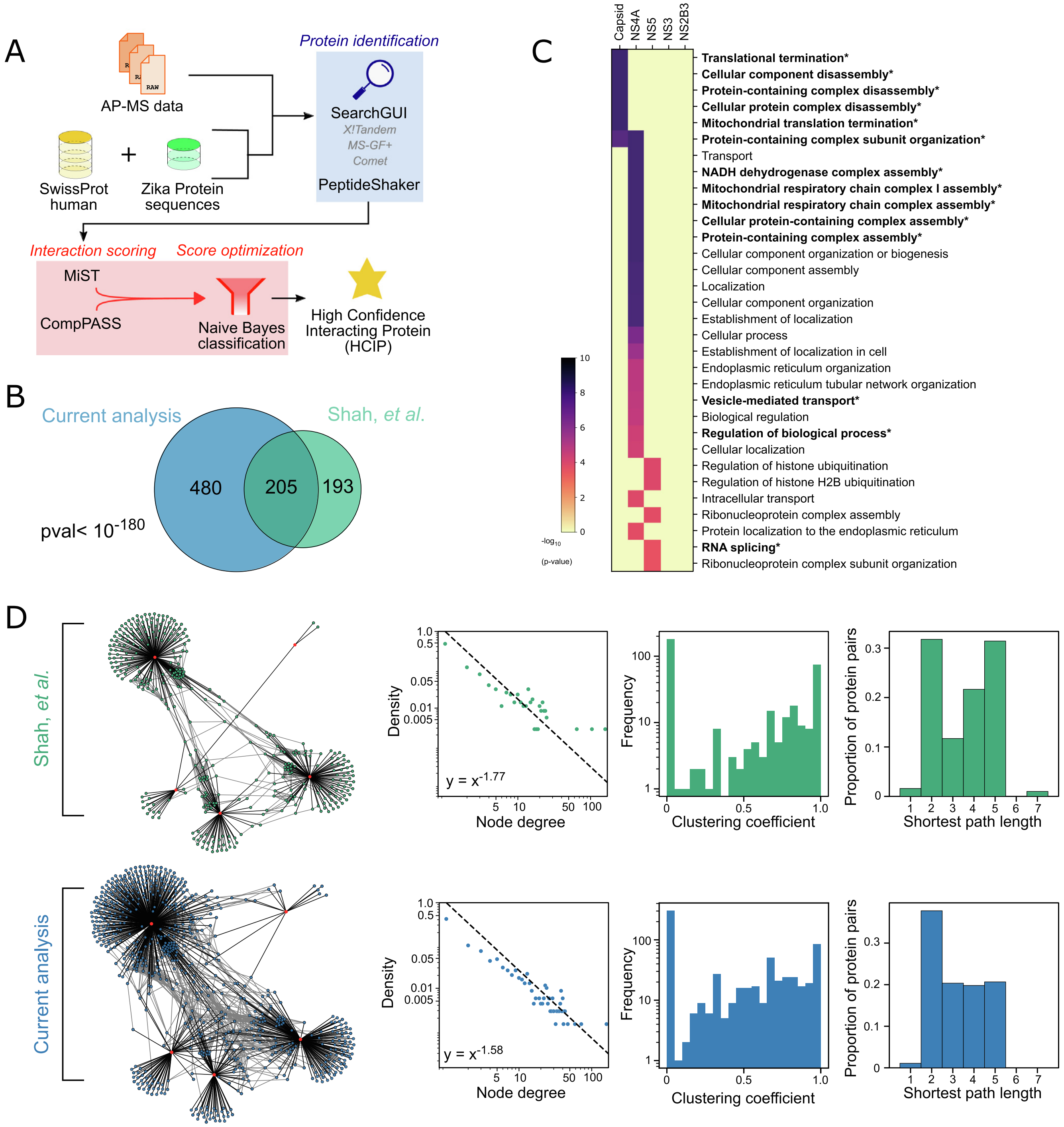

Fig. 1. Evaluating reproducibility across AP-MS data analysis A. Graphical representation of the pipeline used for affinity-purification mass spectrometry (AP-MS) data analysis. AP-MS data is interrogated using a combined database of human protein sequences (yellow) and Zika viral protein sequences (green). The pipeline contains a step of protein identification (blue) using SearchGUI and PeptideShaker, followed by interaction scoring (red) using MiST, CompPASS and a Naïve Bayes classification algorithm, to produce a list of highly confident interacting proteins (HCIPs). B. Overlap of proteins identified in this study (current analysis) with that of the original study from Shah and colleagues [8]. Significance of the overlap was evaluated using a Fisher's exact test. C. Gene ontology (GO) enrichment of interacting proteins (preys) for each viral proteins (baits – top). The heatmap displays all significant enrichments (FDR < 1%, Benjamini-Hochberg correction). The GO terms highlighted in bold and a star correspond to GO terms also identified in the original study. D. Network characteristics between the current analysis and that of the original study. Networks are represented with bait proteins in red (Zika proteins) and interacting proteins in green for the original study and in blue for the current analysis. The plots from left to right correspond to the node degree distribution, the distribution of local clustering coefficients and the distribution of shortest pair length. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of both scoring methods improved the F1 score (31%) but still yielded a poor recall with a stringent filter (274 HCIPs called, corresponding to the top 2.2% of interactions kept). The Naïve Bayes built on the strength of both MiST and CompPASS reaching the highest F1 score (40%) with a reasonable recall (57%) and calling a total of 665 HCIPs for a threshold of 0.45 (Suppl. Fig. 1).

Out of these 665 HCIPs, 199 were also identified in the original study [8] (Suppl. Table 1). Thus, our pipeline was able to identify

56.7% of the protein–protein interactions identified in the original study (Fig. 1B). Although this overlap might not seem large, it is highly significant (Fisher's exact test, p-value < 10^{-180}). As mentioned in the introduction, discrepancies between AP-MS datasets have been reported previously [3]. We show here that such discrepancies can originate from the computational analysis of the raw data. That is in part because AP-MS techniques pull on protein complexes, which often share peptides across subunits or elements of a complex. The method by which the search algorithm and downstream analyses handle the protein groups skews the identification towards particular subunits or members of a protein complex. Hence, to verify the validity of our results we calculated the enrichment of biological processes from our 665 HCIPs (Fig. 1C). All GO terms identified have been reported previously in association with viral infection, and 14 terms were also reported in the original study. Interestingly, we noticed that both analyses agreed on the most significant GO terms (11 out of 16 GO terms with a p-value below 0.00005, Suppl. Table 2). We hypothesized that our analysis identified the same biological complexes than those in the original study, but different proteins within these complexes. To test this hypothesis, we retrieved the network published by Shah and colleagues and compared it with the network generated from our analysis (Fig. 1D). Baits are positioned identically to ease the visual comparison. We observe that the network from the current analysis is more interconnected than that of the original study. Relatively densely interconnected networks are expected when looking at protein–protein interactions [44,53,54]. As expected of naturally occurring networks node degree distributions fitted a power law. The distribution from the original study however seemed to deviate more, especially towards higher degrees (Fig. 1D). Local clustering coefficient distribution from the current analysis indicated a more coherent subnetwork region of the human whole interactome likely involved in ZIKV activity. Perhaps because of excessive stringency, the network from the original publication may have overlooked peripherally located yet important proteins in ZIKV's effect on system shift towards a regime geared for its replication. Indeed, the original study reports only the most densely connected regions of the network and likely missed protein interactors more relevant to rewire connections between complexes necessary for viral hijacking of cellular processes [54]. The sparsity of the network from the original study is also visible from the bimodal distribution of the shortest path between all protein pairs. While the network from the current analysis still shares a significant homology (i.e. relative number of preys for each baits) with that of the original study (Fig. 1D), the more continuous distribution in shortest path lengths shows structure closer to that expected of typically interconnected protein networks and may provide a more holistic model of ZIKV infection. Furthermore, protein complexes from the CORUM database with subunits found in the original study and the current analysis were recovered more reliably in the current study. 132 complexes were found in both analyses (72.7%), with the current analysis identifying more subunits within these complexes (Suppl. Fig. 1).

3.2. Zika viral protein interactions reveal viral host-mRNA translational control

We built a network incorporating our HCIPs and highlighted the proteins also identified in the original study (for brevity Fig. 2 represents only proteins with a degree above 1, see Suppl. Fig. 3 for whole network). For the capsid, NS5 and NS4A viral proteins, both networks agreed on protein complexes. For some the same proteins have been identified (dashed black edges, Fig. 2, Suppl. Fig. 3), as it is for the translational termination complex (MRPL9 (Q9BYD2), MRPL20 (Q9BYC9), MRPL23 (Q16540) and MRPL47 (Q9HD33)) or the spliceosome (AQR (O60306), CWC25 (Q9NXE8), PRPF38A (Q8NAV1) and SRRM1 (Q8IYB3)). For others different members of a same complex were identified (solid lines), as it was for the mitochondrial translational termination complex (in green on Fig. 2) or the mitochondrial respiratory chain complex assembly (in pink on Fig. 2). Both networks diverge mostly on protein identifications for the NS3 and NS2B3 (Fig. 2). In both analyses, no confident GO term enrichment was found for NS2B3. In the original study, the NS3 interactome was found enriched for the GO term "Spindle formation" driven by a single protein (HSPA2 (P54652)). That protein is also identified in our analysis, however we filtered out enrichment driven by a single protein. In total, we identified 14 proteins interacting with NS2B3 and 67 proteins interacting with NS3 (Suppl. Table 1).

Interestingly, in agreement with the published literature on flaviviruses, 35% of GO terms identified in our analysis relate to mRNA processing, splicing, transport and translational control (Fig. 2). These associations were mostly linked to the NS5 and capsid viral proteins, as previously reported [7,8,11]. With previous studies having highlighted pervasive translation and the biological importance of some alternative proteins upon viral infection [17,18,20], we hypothesized that Zika viral proteins may also interact with proteins currently missing from protein databases.

3.3. Building a custom database for deeper proteome exploration

One caveat of inclusion of alternative proteins in databases for proteomics experiments is the consequent increase in the size of the resulting database. Large databases lead to a decrease in the specificity and sensitivity of the analysis, resulting in a low number of confidently identified proteins and a higher rate of false positives [55,56]. Thus, it is recommended to use adapted pipelines, such as a stringent 0.001% false discovery rate (FDR), or to limit the size of the database to a maximum of 100,000 entries [30,32,57]. Here, we retrieved RNA-seq data to identify transcripts confidently expressed in Zika-infected cells and build a custom database tailored to a Zika infection context. Since the Zika virus leads to alternative splicing of transcripts [58], as supported by the AP-MS data analysis (in orange on Fig. 2), we used a pipeline to identify both canonical transcripts and novel splice variants (Fig. 3A). Most transcripts displayed an expression level below 200 FPKM, with a few highly expressed transcripts (Suppl. Table 3). The custom database is designed to include any protein (canonical, alternative protein, novel isoform, and Cuff-Prot – defined in Table 1) from transcripts detected by RNA-seq above a certain expression (FPKM) threshold. To select the optimal threshold, we evaluated the size of the resulting custom database and the proportion of UniProt proteins identified in our previous analysis (Suppl. Table 1) included in the database (Fig. 3B). We identified 1.5 as the ideal FPKM threshold at which 76% of proteins identified in our first analysis were included in the resulting database, while maintaining an overall database size below 100,000 proteins (Suppl. Data 1).

The resulting custom database contained 98,508 proteins (Fig. 3C). These included 13,442 canonical proteins, detailed as 13,048 proteins from the full UniProtKB resource [33], 83 proteins from the Ensembl annotation [34] and 311 from the NCBI RefSeq annotation [35] not present in the UniProtKB database. The database also contained 80,573 alternative proteins and 4,469 novel isoforms of canonical proteins from the OpenProt resource [30]. Finally, 19 protein sequences were derived from novel splice transcripts identified in the RNA-seq analysis. Since genome annotations enforce a minimal length criteria of 100 codons, except for previously characterized examples, this introduced a protein length bias across the different protein categories (Fig. 3D), as previously observed [26]. AltProts displayed the lowest median length

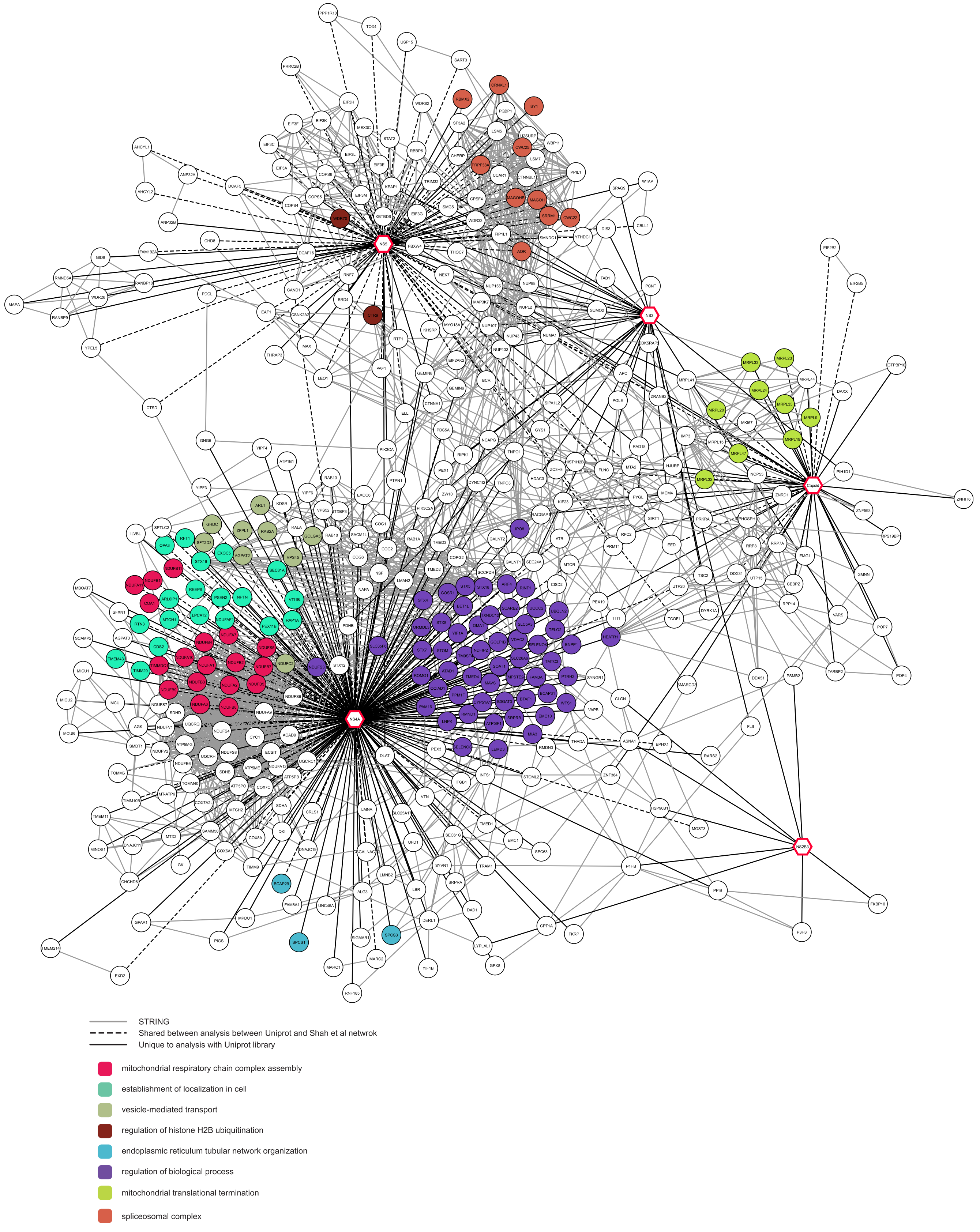
NS5
NS3
Capsid
NS4A
NS2B3
STRING
Shared between analysis between Uniprot and Shah et al netwrok
Unique to analysis with Uniprot library
mitochondrial respiratory chain complex assembly
establishment of localization in cell
vesicle-mediated transport
regulation of histone H2B ubiquitination
endoplasmic reticulum tubular network organization
regulation of biological process
mitochondrial translational termination
spliceosomal complex

Fig. 2. Network of viral-host protein–protein interactions Network of protein interactions for proteins of the Zika virus (indicated as octogonal nodes with thick red borders) with proteins with a degree higher than 1. The nodes (circle with black borders) correspond to human proteins identified by AP-MS data analysis. Nodes are coloured based on their associated gene ontology as indicated on the bottom left corner. Edges represent confident interactions between two proteins as follows: solid black lines are interactions only identified in the current analysis; dashed black lines are interactions shared with the original study (Shah, *et al.*); and full grey lines are host–host protein interactions retrieved from the STRING database. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

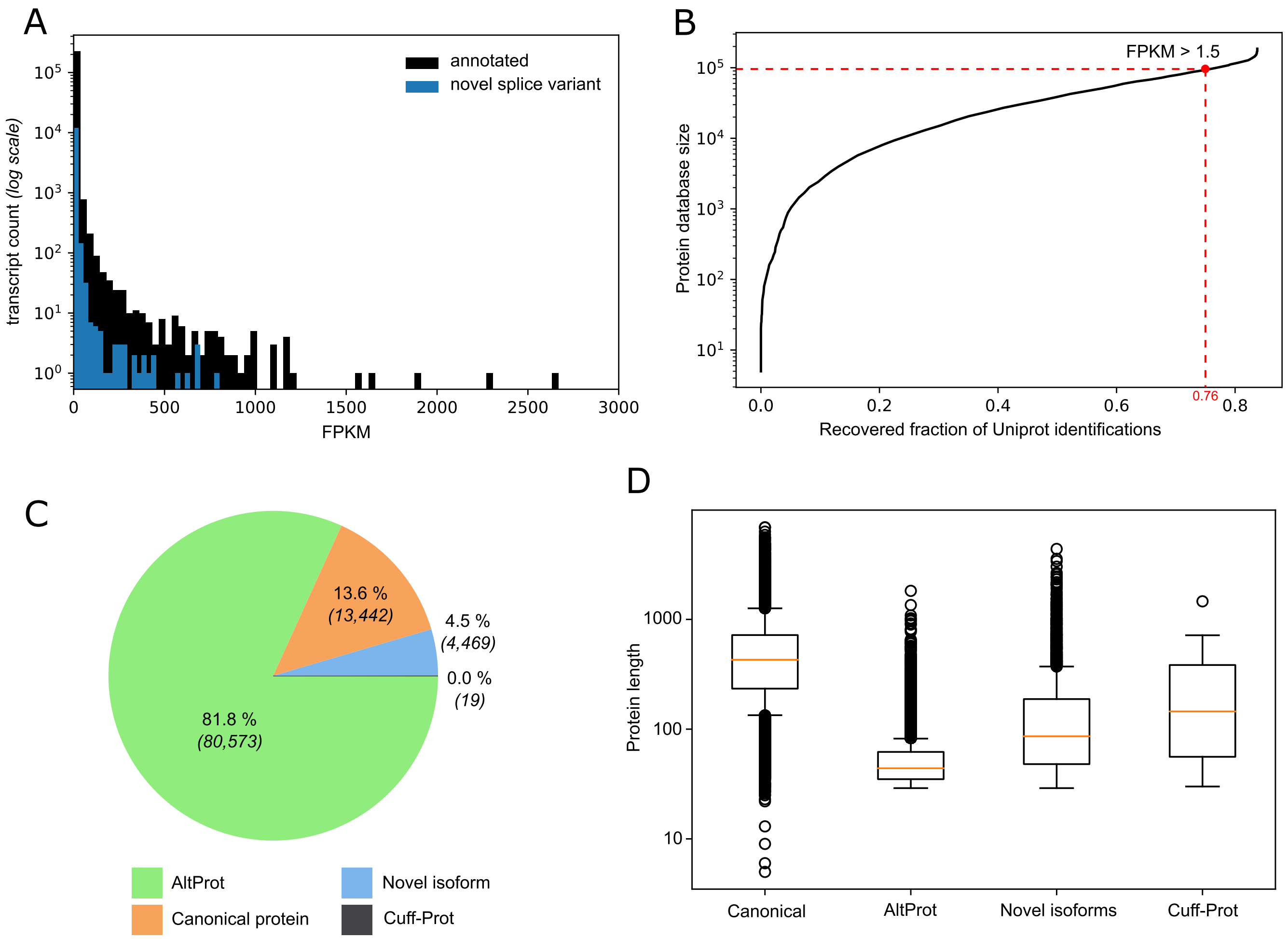

Fig. 3. Construction and specifics of the custom database A. Distribution of transcripts identified in the RNA-seq analysis by their estimated level of expression (FPKM). Annotated transcripts are indicated in black, while novel transcripts are indicated in blue. B. Relation curve between the size of the custom database and the fraction of UniProt identifications included in the custom database, based on the FPKM threshold from the RNA-seq analysis. The chosen FPKM threshold (1.5) is indicated by a red dot. C. Composition of the custom database. Proteins are classified as canonical proteins if annotated in UniProt, NCBI RefSeq and/or Ensembl (orange). Proteins are otherwise called novel isoforms (blue), alternative protein (altProt - green) of CUFF-transcript derived protein (Cuff-Prot - grey). D. Distribution of the protein length for each protein category defined in panel C. Boxes represent the inter quartile range (IQR) notched at the median and the whiskers are set at 0.5*IQR over and under the 25th and 75th percentiles. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Definition of categories of proteins in the custom database.

Name	Full name	Description	Source
Canonical	Canonical protein	Protein annotated in current annotations	UniProtKB, Ensembl and/or NCBI RefSeq
AltProt	Alternative protein	Protein currently not annotated with no significant homology with the canonical protein of the same gene	OpenProt
Novel isoform	Novel protein isoform	Protein currently not annotated with a significant homology with the canonical protein of the same gene	OpenProt
Cuff-Prot	Cuff-transcript derived protein	Predicted protein from an ORF within a novel splice variant transcript detected by RNA-seq	*In silico* translation

(44 amino acids), while novel isoforms and Cuff-Prots displayed a slightly increased median length (86 and 145 amino acids respectively) although still much shorter than that of the canonical proteins (429 amino acids).

3.4. AP-MS data analysis with a custom database retrieves confident canonical interactions

We re-analysed the AP-MS dataset from Shah and colleagues [8] with our custom database concatenated with the Zika viral protein sequences. Prior to HCIP filtering, we identified across all replicates for all baits 18,258 interactions with canonical proteins, 81 with altProts, 472 with novel isoforms and 25 Cuff-Prots (Fig. 4A). However, as usual in standard MS-based proteomics, the concept of peptide unicity is inherent to the database used. Here, we want to report novel proteins only if the peptide cannot be explained by a canonical protein, including those not present in the custom database (i.e. due to low transcript abundance). Thus, we enforced a second peptide unicity check against the whole OpenProt database (version 1.5). This left us with 18,258 interactions with a canonical protein and 16 with an altProt (Fig. 4A). It is to note that such peptide assignation rules create a bias against the detection of

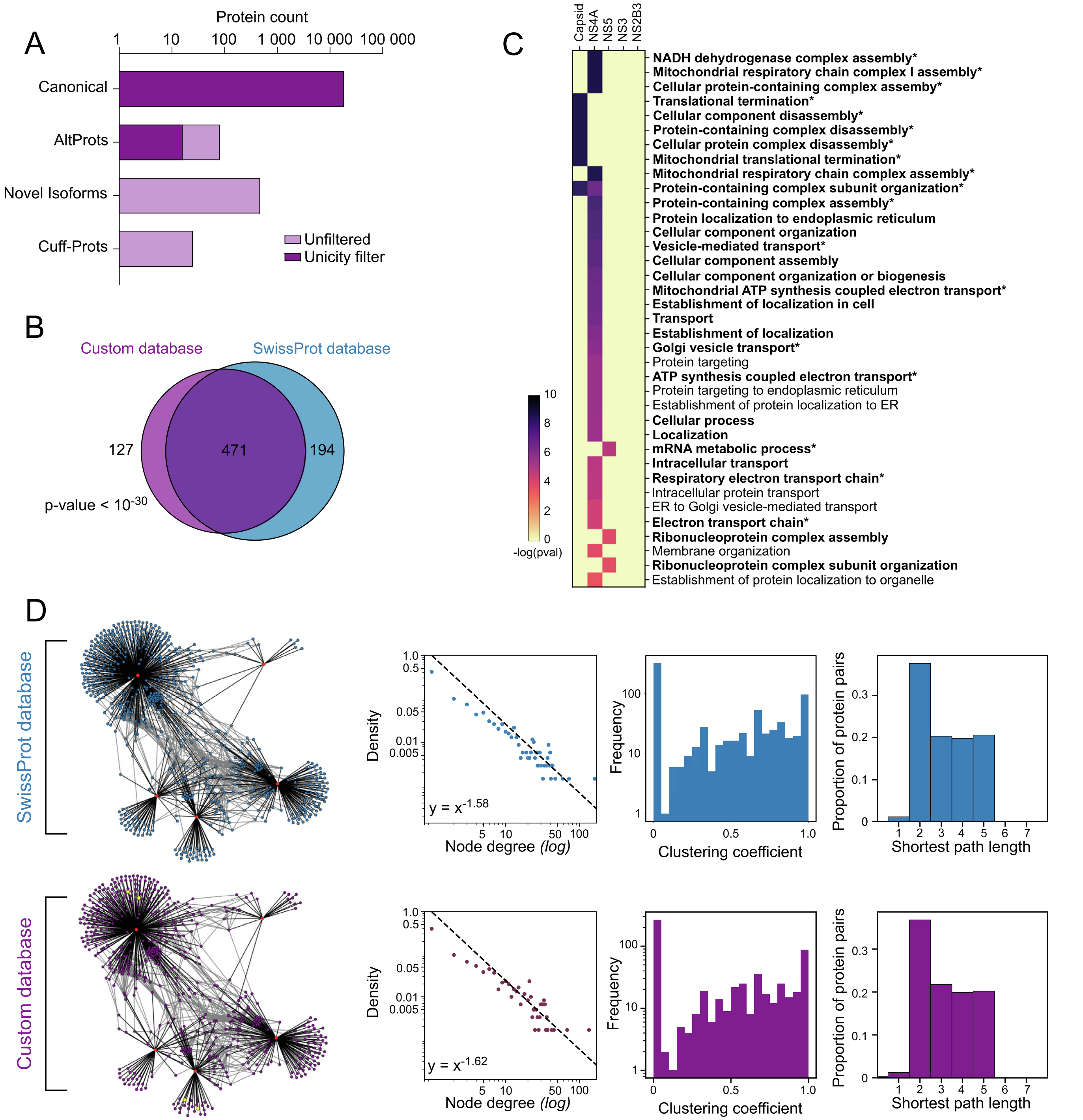

Fig. 4. AP-MS data analysis features with the custom database A. Count of proteins identified in the protein identification step of the analysis for each protein category (canonical, altProt, novel isoforms and Cuff-Prots). The light purple corresponds to the counts prior to the peptide unicity additional filter. The dark purple corresponds to the counts after filtering for peptide unicity across the whole proteome. B. Overlap of proteins identified with the custom database or the SwissProt database (see Fig. 1). Significance of the overlap was evaluated using a Fisher's exact test. C. Gene ontology (GO) enrichment of interacting proteins (preys) for each viral proteins (baits – top) identified with the custom database. The heatmap displays all significant enrichments (FDR < 1%, Benjamini-Hochberg correction). The GO terms highlighted in bold and a star correspond to GO terms also identified with the SwissProt database. D. Network characteristics between the current analysis and that of the original study. Networks are represented with bait proteins in red (Zika proteins) and interacting proteins in blue for the analysis with SwissProt and in purple for the current analysis. Alternative proteins are indicated in yellow. The plots from left to right correspond to the node degree distribution, the distribution of local clustering coefficients and the distribution of shortest pair length. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

novel proteins, although the detections are more confident. After parsing for detection in at least two replicates, we obtained a list of 1,954 identified proteins, corresponding to 9,770 potential interactions, which included 7 altProts.

598 HCIPs were ultimately identified using the custom database (Fig. 4B, Suppl. Table 4). Out of these 598 HCIPs, 471 were also identified in the analysis with the SwissProt database. Thus, the pipeline was able to identify 70.8% of the proteins identified with

the SwissProt database. This overlap is both highly significant (Fisher's exact test, p-value < 10^{-30}) and underestimated: since only 76% of the SwissProt protein entries identified in the first analysis were included in the custom database based on the RNA-seq analysis, the overlap is of 93.1%. Furthermore, the GO enrichment analysis revealed a strong overlap with the analysis using the SwissProt database (30 of the 37 terms – Fig. 4C, Suppl. Table 5). The new terms found with the custom database are particularly relevant to viral infection. Notably, the custom database identified proteins involved in SNAREs-mediated membrane fusion events (USE1 (Q9NZ43), BET1 (O15155), COG3 (Q96JB2) and GOSR2 (O14653)) as interactors of the viral NS4A protein. These proteins were identified in the analysis with the SwissProt database but scored just under the threshold of the Naïve Bayes (Suppl. Table 1). This relates to studies in other flaviviruses that found NS4A to be implicated in the formation of replication factories, organelle-like membranous structures resulting from drastic reorganization of ER membranes [14,59]. Furthermore, proteins associated with the nuclear chromatin were enriched amongst Capsid interactors (YY1 (P25490), HIST2H2BE (Q16778), BAZ2A (Q9UIF9), and SMARCD2 (Q92925)). Interestingly, all of these have been found interacting with the sirtuin protein (SIRT1), and sirtuin inhibitors were reported to block Zika virus infection downstream of viral entry [60]. These proteins were not identified in the analysis with the SwissProt database (apart from BAZ2A with a Naïve Bayes score of 0.44). Overall, the HCIPs identified with our custom database included the most confident HCIPs from the analysis done with the SwissProt database, but also highlighted novel interactions, creating a filtered yet interconnected network (Fig. 4D). Comparing the networks generated with the SwissProt or our custom database, this latest network displayed very similar characteristics with a degree distribution consistent with that of known protein–protein networks and retained interconnectivity (Fig. 4D). The network displayed a similar distribution of shortest path lengths (average of 3.21 as for the network with the SwissProt database), and maintained an even distribution of clustering coefficients. This suggests that using targeted custom protein databases allows for specific enrichment of protein complexes.

3.5. Deeper ORF annotation highlights novel interactors of viral proteins

We reconstructed the viral-host protein interaction network from the analysis using our custom database (Fig. 5, Suppl. Fig. 4). For the capsid, NS5 and NS4A viral proteins, our tailored analysis identified proteins for the vast majority also identified with the SwissProt database (67.2%, 81.5% and 82.4% respectively). Interestingly, we identified four novel proteins (purple nodes, Fig. 5) confidently interacting with viral proteins. Two alternative proteins (IP_209094 and IP_148668, accessions from the OpenProt resource) were found interacting with capsid, and two others (IP_086141 and IP_058843) were found interacting with NS4A.

The alternative protein IP_209094 originates from messenger RNAs (mRNA) from the *COL2A1* gene. According to the OpenProt resource, this protein is 54 amino acids long and has been detected in two proteomics studies on NCI-60 cancer cell line panel. One study looked at the global proteome (PRDB000035), while the other looked at extracellular vesicles (PXD005479). The protein is conserved across 3 species (IP_209094). We confidently identified this alternative protein with the unique peptide MVKLENLEKLVK (Suppl. Fig. 5A), not shared with any other known protein. Although this protein does not contain any predicted functional domain on OpenProt, the DeepGOPlus tool found similarities with proteins involved in viral processes and ribosome biogenesis [61]. This is in line with other proteins interacting with the capsid in our analysis and others who reported known ribosomal proteins interacting with the capsid [7,8,11].

The IP_148668 alternative protein also interacts with the capsid. The protein originates from a ncRNA associated with *AL109918.1* (or *LOC730101*), a gene annotated as non-coding and associated with cancer [62]. This alternative protein displays both MS-based and ribosome profiling-based evidence on the OpenProt resource (MS score of 2 and translation score of 3). We confidently identified this alternative protein with the unique peptide CLTLPFVSPMNQSWDTSKK, not shared with any other known protein. Although OpenProt does not list any identified predicted functional domain, the latest CDD annotation identifies a RNA recognition motif (accession cd12757) [63].

The alternative protein IP_058843, interacts with the NS4A viral protein. This alternative protein was confidently identified with a unique peptide, QSVVLLSSSRR (Suppl. Fig. 5B). Interestingly, this protein is encoded in the 3′ untranslated region (UTR) of the *UBIAD1* gene. The protein is 80 amino acids long, is conserved in the chimpanzee (*Pan troglodytes*), and has already been detected once according to the OpenProt resource.

The alternative protein IP_086141 interacts with the NS4A viral protein. This alternative protein was confidently identified with a unique peptide (Suppl. Fig. 5C). Interestingly, this protein originates from the *ANTXR1* gene and its ORF overlaps the ANTXR1 canonical ORF in a different reading frame (Suppl. Fig. 5D). The protein is 78 amino acids long and is conserved between chimpanzee (*Pan troglodytes*), mouse (*Mus musculus*), cow (*Bos Taurus*), and sheep (*Ovis aries*) (Suppl. Fig. 5E). It has been detected once by MS according to the OpenProt resource (PXD011929) and was also detected in a study in HeLa cells nuclei [64].

As demonstrated here, deeper ORF annotation and the use of transcriptomic-informed database allow for the identification of novel proteins interacting with Zika viral proteins.

3.6. Deeper ORF annotations highlight novel differentially expressed proteins

Additionally, as it has been demonstrated before with other viruses, the translational changes upon Zika virus infection could lead to differentially expressed novel proteins [17,18,20]. We retrieved a publicly available dataset of quantitative proteome profiling upon Zika virus infection in monocytes [9]. We used the same pipeline as in the original study but queried the MS data with our custom database (Fig. 6A). When compensating for multiple comparisons, the original study adjusted the p-value for a FDR between 1% and 5%, while we fixed the FDR at 5% or 1%. Using the FDR set to 5%, we identified 793 differentially expressed proteins between the control and the Zika-infected samples (Fig. 6B). Out of these, 322 proteins were also identified as differentially expressed in the original study (Suppl. Table 6). This corresponds to a highly significant overlap of 59.1% (Fisher's exact test, p-value < 10^{-287}). With the FDR set to 1%, we identified 380 differentially expressed proteins (DEPs), from which 235 were also identified in the original study. This 1% FDR filtering thus corresponded to a highly significant overlap of 43.2% with the original dataset (Fisher's exact test, p-value < 10^{-287}, Suppl. Table 6). For downstream analyses, we used this more confident set of DEPs filtered at a 1% FDR. We identified a total of 138 up-regulated proteins and 242 down-regulated proteins (Fig. 6C).

We confidently identified 12 alternative proteins (IP_068551, IP_070279, IP_074954, IP_139806, IP_157897, IP_195829, IP_216771, IP_232994, IP_233268, IP_240469, IP_265139, and IP_276654 - https://openprot.org/p/savedSearch/lCa). One of these proteins, IP_265139, encoded in an ORF overlapping that of the ATP5F1A protein, was found significantly up-regulated in Zika-infected samples (Fig. 6C).

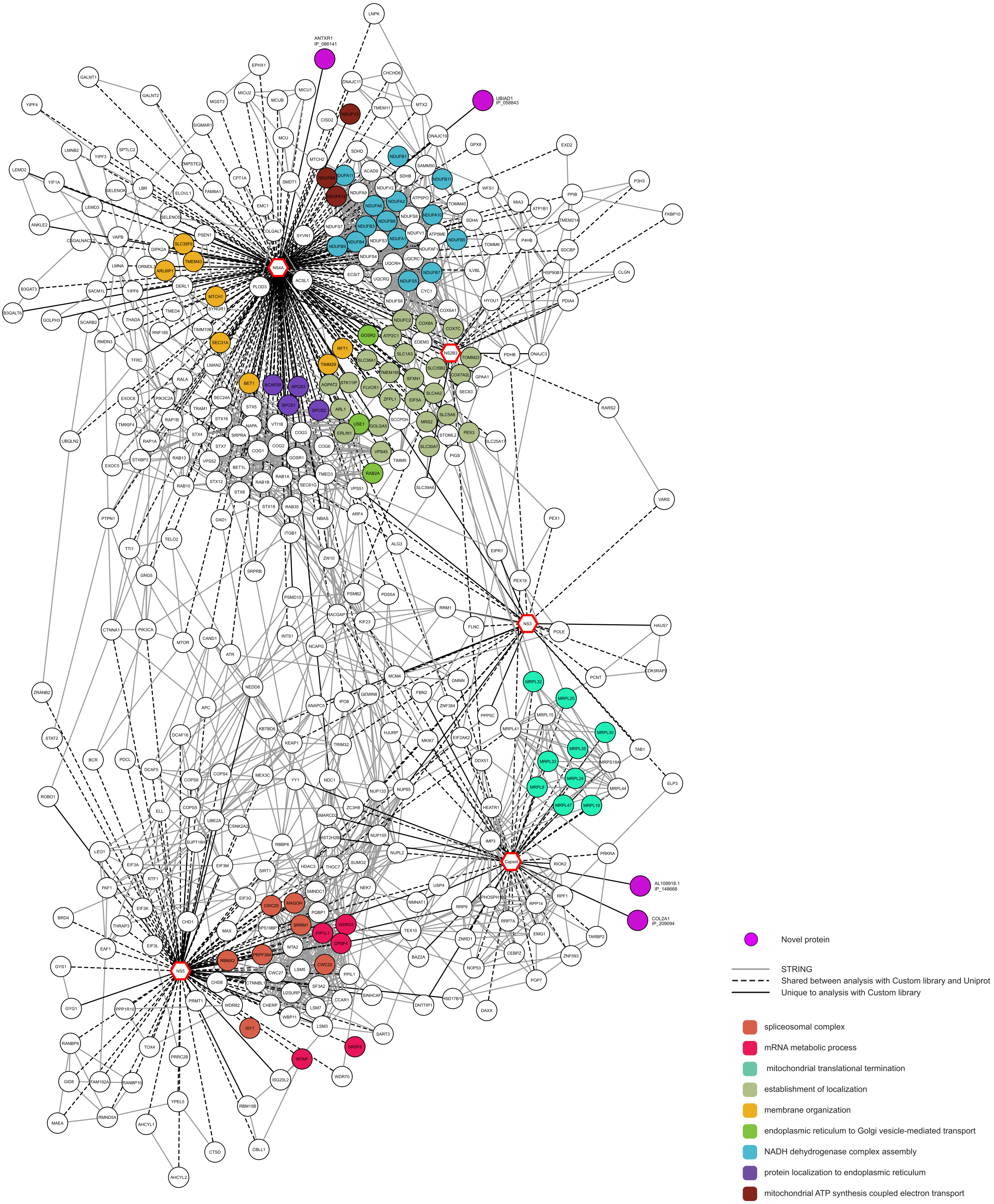
NS4A
NS2B3
NS3
Capsid
NS5
ANTXR1
IP_086141
UBIAD1
IP_058843
AL109918.1
IP_148668
COL2A1
IP_209094
Novel protein
STRING
Shared between analysis with Custom library and Uniprot
Unique to analysis with Custom library
spliceosomal complex
mRNA metabolic process
mitochondrial translational termination
establishment of localization
membrane organization
endoplasmic reticulum to Golgi vesicle-mediated transport
NADH dehydrogenase complex assembly
protein localization to endoplasmic reticulum
mitochondrial ATP synthesis coupled electron transport

Fig. 5. Network of viral-host protein–protein interactions built from the custom database Network of protein interactions for proteins of the Zika virus (indicated as octogonal nodes with thick red borders) with proteins with a degree higher than 1. The nodes correspond to canonical human proteins (circle with black borders) or alternative proteins (purple circle) identified by AP-MS data analysis. Nodes of canonical proteins are coloured based on their associated gene ontology as indicated on the bottom right corner. Edges represent confident interactions between two proteins as follows: solid black lines are interactions only identified in the current analysis; dashed black lines are interactions shared with the original study (Shah, *et al.*); and full grey lines are host–host protein interactions retrieved from the STRING database. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

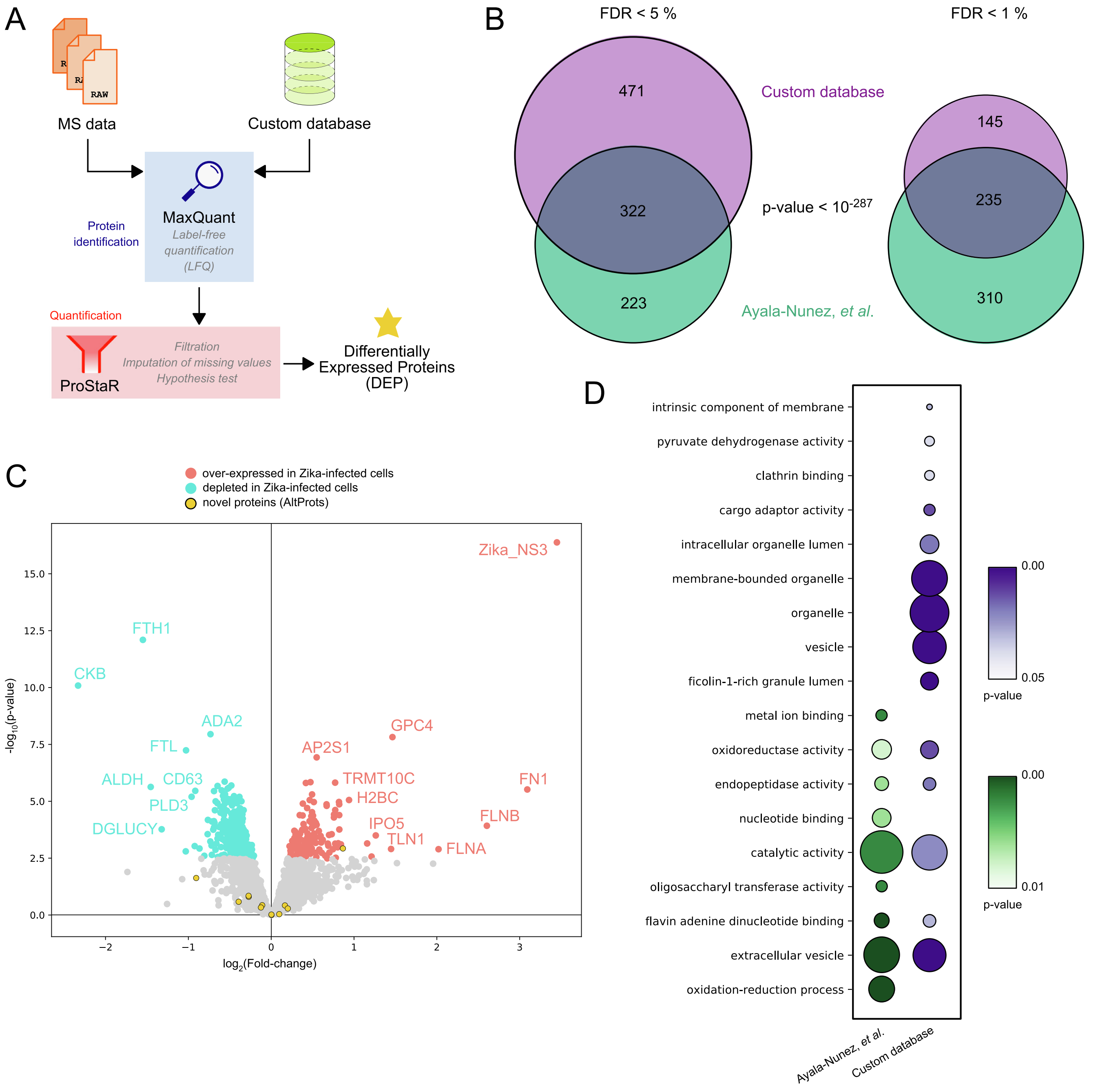

Fig. 6. Proteome profiling with the custom database identifies novel proteins A. Graphical representation of the pipeline used for quantitative proteome profiling analysis. Mass spectrometry (MS) data is interrogated using the custom database defined in Fig. 3. The pipeline contains a step of protein identification (blue) using MaxQuant, followed by protein quantification (red) using ProStaR, to produce a list of differentially expressed proteins (DEPs). B. Overlap of proteins identified in this analysis (custom database) with that of the original study from Ayala-Nunez and colleagues [9]. Significance of the overlap was evaluated using a Fisher's exact test. C. Volcano plot of differentially expressed proteins during Zika virus infection. Proteins found significantly upregulated in the Zika-infected conditions are indicated in red, while those found significantly down-regulated are indicated in blue. Novel proteins are highlighted in yellow with a black edge. D. Gene ontology (GO) enrichment of significantly differentially expressed proteins. All significant enrichments (FDR < 1%, Benjamini-Hochberg correction) are listed across both analyses. The dot size relates to the number of proteins supporting this GO term enrichment, while the color relates to the p-value. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Furthermore, our analysis with a custom database identified a majority of GO terms enriched in the original analysis (5 out of 9, Fig. 6D). Although, we did not identify GO terms relating to oxidation–reduction or nucleotide binding enriched, we did identify terms relating to vesicle formation and granule lumen that were not reported in the original study (Suppl. Table 7). This suggests that the use of a custom database might identify enrichment of more specific terms.

4. Discussion

In this study, we explored publicly available datasets of pathogen-host interactions to further analytical methods aimed at the production of more powerful models of these biological systems. We demonstrated the importance of protein groups handling when comparing MS datasets and protein interaction networks. Furthermore, using a custom database we were able to retrieve biologically relevant protein complexes, and discover alternative proteins as novel interactors of Zika viral proteins. Although more bench work is needed to functionally characterize the identified alternative proteins and validate the interactions with viral proteins, this study lends support to the use of transcriptomic-informed databases in proteomics. The computational framework presented here highlights the biological insights gained from using deeper ORF annotations [17,19–21,65,66]. The application of this framework on published proteomics dataset might shed light on cellular processes previously not considered [26,30,32,67]. Such proteogenomics-like endeavour is facilitated by existing repositories of experimentally supported yet non-annotated ORFs, such as the OpenProt resource or the sORFs repository [28,30].

The discovery of unannotated ORFs as differentially expressed proteins (DEPs) or highly confident interacting proteins (HCIPs) upon viral infection has previously been shown with the human cytomegalovirus, retroviruses and flu lysates [17,19,20]. Some of the identified alternative proteins originated from pseudogenes, which should be re-classified as paralogs. Various pseudogenes were previously shown to produce functional proteins, and many have been involved in human diseases [68–71]. Moreover, the evolutionary arms race in which virus and host are involved imposes pressure for constant evolution of immune evasion or interaction strategies. Thus, the evolution strategies of viral-host protein interactions involve gene duplication (paralogy) [72–74]. Although mostly studied in the viral genome, this paralogy strategy is not exclusive to the virus [74]. This suggests that many alternative proteins from pseudogenes might be important in viral-host protein interaction networks.

Furthermore, using publicly available RNA-seq datasets, a custom protein database can be created which enriches biologically relevant identifications. Both the sORFs repository and the OpenProt resource allow custom downloads of novel ORFs [30,75]. The use of customized databases optimized for the size of the search space and the biological relevance of included proteins have been shown to outperform standard database strategies [76,77]. The results presented in this study show that biologically relevant information can be gained from the use of a custom protein database. Although the most confident protein identifications were reported by analyses with both the SwissProt database and the custom database, the latter provided additional identifications leading to an enrichment in GO terms relevant for viral processes and supported by previous experimental evidence [14,78,79].

Notably, we retrieved an enrichment in proteins known to associate with the nuclear chromatin in the capsid interactors (Fig. 4). These are proteins involved in the neuronal development [80–84]. Interestingly, the capsid is the only Zika viral protein reported to associate with proteins involved in the neuronal development [3]. These proteins represent important interactors to better understand ZIKV-mediated microcephaly. Interestingly, in the original study and other datasets, nucleolin (NCL) was identified as strong capsid interactor [3,7,8]. However, in our analysis, NCL was indeed identified but did not pass filters due to a lack of specificity and low abundance (Suppl. Table 5).

Our strategy identified biologically relevant canonical proteins and novel proteins. However, the RNomics and proteomics datasets in this study were not paired and displayed differences in cellular background and infection stage. This might prevent the identification of other biologically relevant proteins; on the other hand, it also indicates that the identified interactors show a high degree of conservation across cellular backgrounds. Using paired data might lead to other protein identifications, although maybe more specific to a cellular background, experimental design or infection stage [3]. Furthermore, the AP-MS dataset used here over-expressed viral proteins individually which may prevent the detection of biologically relevant protein that necessitate the combined effect of all viral proteins.

CRediT authorship contribution statement

Sebastien Leblanc: Software, Formal analysis, Visualization. **Marie A. Brunet:** Conceptualization, Methodology, Software, Validation, Formal analysis, Visualization, Resources, Writing - original draft.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

S.L. is supported by a CIHR scholarship for graduate studies. We thank X. Roucou and the members of his lab for thoughtful discussions and guidance. We also thank R. Gaudin, NJ. Krogan, G. Ming and all members of their team, who shared their data to public repositories and made this study possible. We thank the team at Calcul Québec and Compute Canada for their support with the use of the supercomputer mp2 from Université de Sherbrooke. Operation of the mp2 supercomputer is funded by the Canada Foundation of Innovation (CFI), le ministère de l'Économie, de la science et de l'innovation du Québec (MESI) and les Fonds de Recherche du Québec.

Author contributions

S.L. conducted the network analyses and generated the corresponding figures under supervision of MA.B. MA.B. designed the study, conducted the transcriptomic and proteome profiling analyses and generated the corresponding figures. MA.B. wrote the paper; all authors approved the final manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.10.010.

References

[1] Guarner J, Hale GL. Four human diseases with significant public health impact caused by mosquito-borne flaviviruses: West Nile, Zika, dengue and yellow fever. Semin Diagn Pathol 2019;36(3):170–6. https://doi.org/10.1053/j.semdp.2019.04.009.

[2] Lessler J, Chaisson LH, Kucirka LM, Bi Q, Grantz K, Salje H, et al. Assessing the Global Threat from Zika Virus. Science 2016;353:aaf8160. https://doi.org/10.1126/science.aaf8160.

[3] Scaturro P, Kastner AL, Pichlmair A. Chasing Intracellular Zika Virus Using Proteomics. Viruses 2019;11. https://doi.org/10.3390/v11090878.

[4] Duffy MR, Chen T-H, Hancock WT, Powers AM, Kool JL, Lanciotti RS, Pretrick M, Marfel M, Holzbauer S, Dubray C, Guillaumot L, Griggs A, Bel M, Lambert AJ, Laven J, Kosoy O, Panella A, Biggerstaff BJ, Fischer M, Hayes EB. Zika virus outbreak on Yap Island, Federated States of Micronesia. N Engl J Med 2009;360(24):2536–43.

[5] Cao-Lormeau V-M, Blake A, Mons S, Lastère S, Roche C, Vanhomwegen J, Dub T, Baudouin L, Teissier A, Larre P, Vial A-L, Decam C, Choumet V, Halstead SK, Willison HJ, Musset L, Manuguerra J-C, Despres P, Fournier E, Mallet H-P, Musso D, Fontanet A, Neil J, Ghawché F. Guillain-Barré Syndrome outbreak associated with Zika virus infection in French Polynesia: a case-control study. The Lancet 2016;387(10027):1531–9. https://doi.org/10.1016/S0140-6736(16)00562-6.

[6] Schuler-Faccini L, Ribeiro EM, Feitosa IML, Horovitz DDG, Cavalcanti DP, Pessoa A, et al. Possible Association Between Zika Virus Infection and Microcephaly – Brazil, 2015. MMWR Morb Mortal Wkly Rep 2016;65:59–62. https://doi.org/10.15585/mmwr.mm6503e2.

[7] Scaturro P, Stukalov A, Haas DA, Cortese M, Draganova K, Płaszczyca A, et al. An orthogonal proteomic survey uncovers novel Zika virus host factors. Nature 2018;561:253–7. https://doi.org/10.1038/s41586-018-0484-5.

[8] Shah PS, Link N, Jang GM, Sharp PP, Zhu T, Swaney DL, Johnson JR, Von Dollen J, Ramage HR, Satkamp L, Newton B, Hüttenhain R, Petit MJ, Baum T, Everitt A, Laufman O, Tassetto M, Shales M, Stevenson E, Iglesias GN, Shokat L, Tripathi S, Balasubramaniam V, Webb LG, Aguirre S, Willsey AJ, Garcia-Sastre A, Pollard KS, Cherry S, Gamarnik AV, Marazzi I, Taunton J, Fernandez-Sesma A, Bellen HJ, Andino R, Krogan NJ. Comparative flavivirus-host protein interaction mapping reveals mechanisms of dengue and zika virus pathogenesis. Cell 2018;175 (7):1931–1945.e18. https://doi.org/10.1016/j.cell.2018.11.028.

[9] Ayala-Nunez NV, Follain G, Delalande F, Hirschler A, Partiot E, Hale GL, Bollweg BC, Roels J, Chazal M, Bakoa F, Carocci M, Bourdoulous S, Faklaris O, Zaki SR, Eckly A, Uring-Lambert B, Doussau F, Cianferani S, Carapito C, Jacobs FMJ, Jouvenet N, Goetz JG, Gaudin R. Zika virus enhances monocyte adhesion and transmigration favoring viral dissemination to neural cells. Nat Commun 2019;10(1). https://doi.org/10.1038/s41467-019-12408-x.

[10] Tang H, Hammack C, Ogden SC, Wen Z, Qian X, Li Y, et al. Zika virus infects human cortical neural precursors and attenuates their growth. Cell Stem Cell 2016;18:587–90. https://doi.org/10.1016/j.stem.2016.02.016.

[11] Coyaud E, Ranadheera C, Cheng D, Gonçalves J, Dyakov BJA, Laurent EMN, St-Germain J, Pelletier L, Gingras A-C, Brumell JH, Kim PK, Safronetz D, Raught B. Global interactomics uncovers extensive organellar targeting by zika virus. Mol Cell Proteomics 2018;17(11):2242–55.

[12] Rolfe AJ, Bosco DB, Wang J, Nowakowski RS, Fan J, Ren Yi. Bioinformatic analysis reveals the expression of unique transcriptomic signatures in Zika virus infected human neural stem cells. Cell Biosci 2016;6(1). https://doi.org/10.1186/s13578-016-0110-x.

[13] Schwanhäusser B, Busse D, Li Na, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. Global quantification of mammalian gene expression control. Nature 2011;473(7347):337–42. https://doi.org/10.1038/nature10098.

[14] Mazeaud C, Freppel W, Chatel-Chaix L. The Multiples Fates of the Flavivirus RNA Genome During Pathogenesis. Front Genet 2018;9. https://doi.org/10.3389/fgene.2018.00595.

[15] Roth H, Magg V, Uch F, Mutz P, Klein P, Haneke K, et al. Flavivirus Infection Uncouples Translation Suppression from Cellular Stress Responses. MBio 2017;8. https://doi.org/10.1128/mBio.02150-16.

[16] Reid DW, Campos RK, Child JR, Zheng T, Chan KWK, Bradrick SS, et al. Dengue virus selectively annexes endoplasmic reticulum-associated translation machinery as a strategy for co-opting host cell protein synthesis. J Virol 2018;92. https://doi.org/10.1128/JVI.01766-17.

[17] Stern-Ginossar N, Weisburd B, Michalski A, Le VTK, Hein MY, Huang S-X, Ma M, Shen B, Qian S-B, Hengel H, Mann M, Ingolia NT, Weissman JS. Decoding human cytomegalovirus. Science 2012;338(6110):1088–93.

[18] Ingolia N, Brar G, Stern-Ginossar N, Harris M, Talhouarne GS, Jackson S, Wills M, Weissman J. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. Cell Reports 2014;8(5):1365–79.

[19] Agarwal S, Harada J, Schreifels J, Lech P, Nikolai B, Yamaguchi T, Chanda SK, Somia NV. Isolation, characterization, and genetic complementation of a cellular mutant resistant to retroviral infection. Proc Natl Acad Sci 2006;103 (43):15933–8.

[20] Razooky BS, Obermayer B, O'May JB, Tarakhovsky A. Viral Infection Identifies Micropeptides Differentially Regulated in smORF-Containing lncRNAs. Genes 2017;8:206. https://doi.org/10.3390/genes8080206.

[21] Brunet MA, Levesque SA, Hunting DJ, Cohen AA, Roucou X. Recognition of the polycistronic nature of human genes is critical to understanding the genotype-phenotype relationship. Genome Res 2018. https://doi.org/10.1101/gr.230938.117.

[22] Andrews SJ, Rothnagel JA. Emerging evidence for functional peptides encoded by short open reading frames. Nat Rev Genet 2014;15(3):193–204. https://doi.org/10.1038/nrg3520.

[23] Pueyo JI, Magny EG, Couso JP. New peptides under the s(ORF)ace of the genome. Trends Biochem Sci 2016;41(8):665–78. https://doi.org/10.1016/j.tibs.2016.05.003.

[24] Ingolia NT. Ribosome Footprint Profiling of Translation throughout the Genome. Cell 2016;165(1):22–33. https://doi.org/10.1016/j.cell.2016.02.066.

[25] Chen J, Brunner A-D, Cogan JZ, Nuñez JK, Fields AP, Adamson B, Itzhak DN, Li JY, Mann M, Leonetti MD, Weissman JS. Pervasive functional translation of noncanonical human open reading frames. Science 2020;367(6482):1140–6. https://doi.org/10.1126/science.aay0262.

[26] Samandi S, Roy AV, Delcourt V, Lucier J-F, Gagnon J, Beaudoin MC, et al. Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. ELife Sciences 2017;6:e27860. https://doi.org/10.7554/eLife.27860.

[27] Orr MW, Mao Y, Storz G, Qian S-B. Alternative ORFs and small ORFs: shedding light on the dark proteome. Nucleic Acids Res 2019. https://doi.org/10.1093/nar/gkz734.

[28] Olexiouk V, Van Criekinge W, Menschaert G. An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. Nucleic Acids Res 2018;46:D497–502. https://doi.org/10.1093/nar/gkx1130.

[29] Hao Y, Zhang L, Niu Y, Cai T, Luo J, He S, et al. SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. Brief Bioinformatics 2017. https://doi.org/10.1093/bib/bbx005.

[30] Brunet MA, Brunelle M, Lucier J-F, Delcourt V, Levesque M, Grenier F, et al. OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. Nucleic Acids Res 2019;47:D403–10. https://doi.org/10.1093/nar/gky936.

[31] Olexiouk V, Menschaert G. Identification of small novel coding sequences, a proteogenomics endeavor. Adv Exp Med Biol 2016;926:49–64. https://doi.org/10.1007/978-3-319-42316-6_4.

[32] Brunet MA, Roucou X. Mass Spectrometry-Based Proteomics Analyses Using the Openprot Database to Unveil Novel Proteins Translated from Non-Canonical Open Reading Frames. J Vis Exp 2019.

[33] Bateman A, Martin MJ, O'Donovan C, Magrane M, Alpi E, Antunes R, et al. UniProt: the universal protein knowledgebase. Nucleic Acids Res 2017;45: D158–69. https://doi.org/10.1093/nar/gkw1099.

[34] Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al. Ensembl 2020. Nucleic Acids Res 2020;48:D682–8. https://doi.org/10.1093/nar/gkz966.

[35] O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 2016;44(D1): D733–45. https://doi.org/10.1093/nar/gkv1189.

[36] Hulstaert N, Shofstahl J, Sachsenberg T, Walzer M, Barsnes H, Martens L, Perez-Riverol Y. ThermoRawFileParser: modular, scalable, and cross-platform RAW file conversion. J. Proteome Res. 2020;19(1):537–42. https://doi.org/10.1021/acs.jproteome.9b00328.

[37] Vaudel M, Burkhart JM, Zahedi RP, Oveland E, Berven FS, Sickmann A, Martens L, Barsnes H. PeptideShaker enables reanalysis of MS-derived proteomics data sets. Nat Biotechnol 2015;33(1):22–4. https://doi.org/10.1038/nbt.3109.

[38] Vaudel M, Barsnes H, Berven FS, Sickmann A, Martens L. SearchGUI: an open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. Proteomics 2011;11(5):996–9. https://doi.org/10.1002/pmic.201000595.

[39] Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 2008;26(12):1367–72. https://doi.org/10.1038/nbt.1511.

[40] Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. Mol Cell Proteomics 2014;13(9):2513–26.

[41] Verschueren E, Von Dollen J, Cimermancic P, Gulbahce N, Sali A, Krogan NJ. Scoring large-scale affinity purification mass spectrometry datasets with MiST. Curr Protocols Bioinf 2015;49(1). https://doi.org/10.1002/0471250953.bi0819s49.

[42] Sowa ME, Bennett EJ, Gygi SP, Harper JW. Defining the human deubiquitinating enzyme interaction landscape. Cell 2009;138(2):389–403. https://doi.org/10.1016/j.cell.2009.04.042.

[43] Jäger S, Cimermancic P, Gulbahce N, Johnson JR, McGovern KE, et al. Global landscape of HIV–human protein complexes. Nature 2012;481(7381):365–70. https://doi.org/10.1038/nature10719.

[44] Huttlin EL, Ting L, Bruckner RJ, Gebreab F, Gygi MP, Szpyt J, et al. The BioPlex network: a systematic exploration of the human interactome. Cell 2015;162 (2):425–40. https://doi.org/10.1016/j.cell.2015.06.043.

[45] Wieczorek S, Combes F, Lazar C, Giai Gianetto Q, Gatto L, Dorffer A, et al. DAPAR & ProStaR: software to perform statistical analyses in quantitative discovery proteomics. Bioinformatics 2017;33:135–6. https://doi.org/10.1093/bioinformatics/btw580.

[46] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15–21. https://doi.org/10.1093/bioinformatics/bts635.

[47] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 2010;28(5):511–5.

[48] Tanimoto TT. An elementary mathematical theory of classification and prediction. International Business Machines Corporation; 1958.

[49] Klopfenstein DV, Zhang L, Pedersen BS, Ramírez F, Warwick Vesztrocy A, Naldi A, Mungall CJ, Yunes JM, Botvinnik O, Weigel M, Dampier W, Dessimoz C, Flick P, Tang H. GOATOOLS: A Python library for Gene Ontology analyses. Sci Rep 2018;8(1). https://doi.org/10.1038/s41598-018-28948-z.

[50] Rowles DL, Terhune SS, Cristea IM. Discovery of host-viral protein complexes during infection. Methods Mol Biol 2013;1064:43–70. https://doi.org/10.1007/978-1-62703-601-6_4.

[51] Morris JH, Knudsen GM, Verschueren E, Johnson JR, Cimermancic P, Greninger AL, Pico AR. Affinity purification–mass spectrometry and network analysis to understand protein-protein interactions. Nat Protoc 2014;9(11):2539–54. https://doi.org/10.1038/nprot.2014.164.

[52] Shteynberg D, Nesvizhskii AI, Moritz RL, Deutsch EW. Combining results of multiple search engines in proteomics. Mol Cell Proteomics 2013;12 (9):2383–93.

[53] Huttlin EL, Bruckner RJ, Paulo JA, Cannon JR, Ting L, Baltier K, Colby G, Gebreab F, Gygi MP, Parzen H, Szpyt J, Tam S, Zarraga G, Pontano-Vaites L, Swarup S, White AE, Schweppe DK, Rad R, Erickson BK, Obar RA, Guruharsha KG, Li K, Artavanis-Tsakonas S, Gygi SP, Harper JW. Architecture of the human interactome defines protein communities and disease networks. Nature 2017;545(7655):505–9. https://doi.org/10.1038/nature22366.

[54] Huttlin EL, Bruckner RJ, Navarrete-Perea J, Cannon JR, Baltier K, Gebreab F, et al. Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. BioRxiv 2020:2020.01.19.905109. https://doi.org/10.1101/2020.01.19.905109.

[55] Jeong K, Kim S, Bandeira N. False discovery rates in spectral identification. BMC Bioinf 2012;13(Suppl 16):S2. https://doi.org/10.1186/1471-2105-13-S16-S2.

[56] Nesvizhskii AI. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J Proteomics 2010;73(11):2092–123. https://doi.org/10.1016/j.jprot.2010.08.009.

[57] Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. Nat Methods 2014;11(11):1114–25. https://doi.org/10.1038/nmeth.3144.

[58] Hu B, Huo Y, Yang L, Chen G, Luo M, Yang J, Zhou J. ZIKV infection effects changes in gene splicing, isoform composition and lncRNA expression in human neural progenitor cells. Virol J 2017;14(1). https://doi.org/10.1186/s12985-017-0882-6.

[59] Roosendaal J, Westaway EG, Khromykh A, Mackenzie JM. Regulated cleavages at the west nile virus NS4A-2K-NS4B junctions play a major role in rearranging cytoplasmic membranes and golgi trafficking of the NS4A protein. JVI 2006;80(9):4623–32. https://doi.org/10.1128/JVI.80.9.4623-4632.2006.

[60] Hackett BA, Dittmar M, Segrist E, Pittenger N, To J, Griesman T, et al. Sirtuin Inhibitors Are Broadly Antiviral against Arboviruses. MBio 2019;10. https://doi.org/10.1128/mBio.01446-19.

[61] Kulmanov M, Hoehndorf R. DeepGOPlus: improved protein function prediction from sequence. Bioinformatics 2020;36:422–9. https://doi.org/10.1093/bioinformatics/btz595.

[62] Liu L, Zhang Y, Cao W. Highly expressed lncRNA LOC730101 promotes lung cancer cell growth through Wnt canonical pathway. Biochem Biophys Res Commun 2017;493(2):992–7. https://doi.org/10.1016/j.bbrc.2017.09.104.

[63] Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, et al. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res 2020;48:D265–8. https://doi.org/10.1093/nar/gkz991.

[64] Cardon T, Salzet M, Franck J, Fournier I. Nuclei of HeLa cells interactomes unravel a network of ghost proteins involved in proteins translation. Biochim Biophys Acta (BBA) – General Subjects 2019;1863(10):1458–70. https://doi.org/10.1016/j.bbagen.2019.05.009.

[65] Merino-Valverde I, Greco E, Abad M. The microproteome of cancer: from invisibility to relevance. Exp Cell Res 2020;392(1):111997. https://doi.org/10.1016/j.yexcr.2020.111997.

[66] Delcourt V, Staskevicius A, Salzet M, Fournier I, Roucou X. Small proteins encoded by unannotated ORFs are rising stars of the proteome, confirming shortcomings in genome annotations and current vision of an mRNA. Proteomics 2017. https://doi.org/10.1002/pmic.201700058.

[67] Deutsch EW, Bandeira N, Sharma V, Perez-Riverol Y, Carver JJ, Kundu DJ, et al. The ProteomeXchange consortium in 2020: enabling 'big data' approaches in proteomics. Nucleic Acids Res 2020;48:D1145–52. https://doi.org/10.1093/nar/gkz984.

[68] Dubois M-L, Meller A, Samandi S, Brunelle M, Frion J, Brunet MA, Toupin A, Beaudoin MC, Jacques J-F, Lévesque D, Scott MS, Lavigne P, Roucou X, Boisvert F-M. UBB pseudogene 4 encodes functional ubiquitin variants. Nat Commun 2020;11(1). https://doi.org/10.1038/s41467-020-15090-6.

[69] Cheetham SW, Faulkner GJ, Dinger ME. Overcoming challenges and dogmas to understand the functions of pseudogenes. Nat Rev Genet 2020;21(3):191–201. https://doi.org/10.1038/s41576-019-0196-1.

[70] Chen Xu, Wan L, Wang W, Xi W-J, Yang A-G, Wang T. Re-recognition of pseudogenes: From molecular to clinical applications. Theranostics 2020;10(4):1479–99. https://doi.org/10.7150/thno.40659.

[71] Singh PP, Affeldt S, Malaguti G, Isambert H. Human Dominant Disease Genes Are Enriched in Paralogs Originating from Whole Genome Duplication. PLoS Comput Biol 2014;10. https://doi.org/10.1371/journal.pcbi.1003754.

[72] Ratmann O, Wiuf C, Pinney JW. From evidence to inference: Probing the evolution of protein interaction networks. HFSP Journal 2009;3:290–306. https://doi.org/10.2976/1.3167215.

[73] Shackelton LA, Holmes EC. The evolution of large DNA viruses: combining genomic information of viruses and their hosts. Trends Microbiol 2004;12(10):458–65. https://doi.org/10.1016/j.tim.2004.08.005.

[74] Brito AF, Pinney JW. Protein–Protein Interactions in Virus–Host Systems. Front Microbiol 2017;8. https://doi.org/10.3389/fmicb.2017.01557.

[75] Olexiouk V, Menschaert G. Using the sORFs.Org Database. Current Protocols in Bioinformatics 2019;65:e68. https://doi.org/10.1002/cpbi.68.

[76] Reisinger F, Martens L. Database on Demand – An online tool for the custom generation of FASTA-formatted sequence databases. PROTEOMICS 2009;9:4421–4. https://doi.org/10.1002/pmic.200900254.

[77] Knudsen GM, Chalkley RJ. The Effect of Using an Inappropriate Protein Database for Proteomic Data Analysis. PLoS One 2011;6. https://doi.org/10.1371/journal.pone.0020873.

[78] Chatel-Chaix L, Cortese M, Romero-Brey I, Bender S, Neufeldt CJ, Fischl W, et al. Dengue Virus Perturbs Mitochondrial Morphodynamics to Dampen Innate Immune Responses. Cell Host & Microbe 2016;20:342–56. https://doi.org/10.1016/j.chom.2016.07.008

[79] Miorin L, Maestre AM, Fernandez-Sesma A, García-Sastre A. Antagonism of type I interferon by flaviviruses. Biochem Biophys Res Commun 2017;492(4):587–96. https://doi.org/10.1016/j.bbrc.2017.05.146.

[80] Landry-Truchon K, Houde N, Boucherat O, Joncas F-H, Dasen JS, Philippidou P, Mansfield JH, Jeannotte L. HOXA5 plays tissue-specific roles in the developing respiratory system. Development 2017;144(19):3547–61.

[81] Jeannotte L, Gotti F, Landry-Truchon K. Hoxa5: A Key Player in Development and Disease. J Dev Biol 2016;4. https://doi.org/10.3390/jdb4020013.

[82] Herskovits AZ, Guarente L. SIRT1 in neurodevelopment and brain senescence. Neuron 2014;81:471–83. https://doi.org/10.1016/j.neuron.2014.01.028.

[83] Fujita Y, Yamashita T. Sirtuins in Neuroendocrine Regulation and Neurological Diseases. Front Neurosci 2018;12. https://doi.org/10.3389/fnins.2018.00778.

[84] Chen T, Zhou L, Yuan Y, Fang Y, Guo Y, Huang H, et al. Characterization of Bbx, a member of a novel subfamily of the HMG-box superfamily together with Cic. Dev Genes Evol 2014;224:261–8. https://doi.org/10.1007/s00427-014-0476-x.