Discovering host protein interactions specific for SARS-CoV-2 RNA genome

Roberto Giambruno[1,2,#,$], Elsa Zacco[3,#], Camilla Ugolini[1,#], Andrea Vandelli[3,4,5], Logan Mulroney[1,6], Manfredi D'Onghia[1], Bianca Giuliani[1], Elena Criscuolo[7], Matteo Castelli[7], Nicola Clementi[7,8], Massimo Clementi[7,8], Nicasio Mancini[7,8], Tiziana Bonaldi[9,10], Stefano Gustincich[3], Tommaso Leonardi[1], Gian Gaetano Tartaglia[3,11,$] and Francesco Nicassio[1,$]

1. Center for Genomic Science of IIT@SEMM, Fondazione Istituto Italiano di Tecnologia, 20139 Milano, Italy.
2. Institute of Biomedical Technologies, National Research Council, 20090 Segrate, Italy.
3. Center for Human Technologies, Istituto Italiano di Tecnologia, Genoa, Italy
4. Department of Biochemistry and Molecular Biology, Universitat Autònoma de Barcelona, Bellaterra, 08193 Barcelona, Spain
5. Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain
6. European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK
7. Laboratory of Microbiology and Virology, Vita-Salute San Raffaele University; via Olgettina 58, 20132 Milan, Italy.
8. Laboratory of Medical Microbiology and Virology, IRCCS San Raffaele Scientific Institute; via Olgettina 60, 20132 Milan, Italy
9. Department of Experimental Oncology, European Institute of Oncology IRCCS, 20139 Milano, Italy.
10. Department of Oncology and Hematology-Oncology, University of Milan, Via Festa del Perdono 7, 20122 Milano, Italy
11. Catalan Institution for Research and Advanced Studies, ICREA, Passeig Lluís Companys 23, 08010, Barcelona, Spain

[#]These authors equally contributed

[$]Correspondence to: roberto.giambruno@cnr.it; gian.tartaglia@iit.it; francesco.nicassio@iit.it;

Abstract

SARS-CoV-2, a positive single-stranded RNA virus, interacts with host cell proteins throughout its life cycle. These interactions are necessary for the host to recognize and hinder the replication of SARS-CoV-2. For the virus, to translate, transcribe and replicate its genetic material. However, many details of these interactions are still missing. We focused on the proteins binding to the highly structured 5' and 3' end regions of SARS-CoV-2 RNA that were predicted by the *cat*RAPID algorithm to attract numerous proteins, exploiting RNA-Protein Interaction Detection coupled with Mass Spectrometry (RaPID-MS) technology. The validated interactors, which agreed with our predictions, include pseudouridine synthase PUS7 that binds to both ends of the viral RNA. Nanopore direct-RNA sequencing confirmed that the RNA virus is heavily modified, and PUS7 consensus regions were found in both SARS-CoV-2 RNA end regions. Notably, a modified site was detected in the viral Transcription Regulatory Sequence - Leader (TRS-L) and can influence the viral RNA structure and interaction propensity. Overall, our data map host protein interactions within SARS-CoV-2 UTR regions, pinpointing to a potential role of pseudouridine synthases and post-transcriptional modifications in the viral life cycle. These findings contribute to understanding virus-host dynamics and may guide the development of targeted therapies.

Introduction

The outbreak of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in the human population has dramatically affected life expectancy worldwide. SARS-CoV-2 infection causes the coronavirus disease 2019 (COVID-19), a potentially fatal condition especially for the elderly population and for individuals with underlying health issues. To date, vaccination is the most effective strategy to significantly reduce the incidence of SARS-CoV-2 infection and the onset of severe COVID-19. However, new and fast-spreading SARS-CoV-2 variants might be able to escape the antibody response of vaccinated people (1).

SARS-CoV-2 is a positive-sense single-stranded RNA virus of the Coronoviridae family (2, 3). Its genome is composed of ~30.000 nucleotides and contains 14 open reading frames (ORFs) encoding 16 non-structural proteins (nsp 1 to 16), 12 accessory proteins (ORF3a, ORF3b, ORF3c, ORF3d, ORF6, ORF7a, ORF7b, ORF8, ORF9b, ORF9c, ORF9d and ORF10) and 4 structural proteins: spike (S), envelope (E), membrane (M) and nucleocapsid (N) (4–6). Structural and accessory proteins are encoded in subgenomic RNAs (sgRNAs) that bear a common 5′ leader sequence fused to different body sequences upstream to the encoded ORF (7). During its life cycle, SARS-CoV-2 RNAs interact with several host proteins specifically required for the translation of viral proteins, the transcription of the subgenomic RNAs, the replication of the genome and the generation of new viral particles (8). At the same time, host proteins recognize the presence of the SARS-CoV-2 RNA in the cytoplasm, activating the innate immune response through the interferon signaling pathways (9–11). Although the interaction between viral RNAs and host proteins has been recognized as a key molecular process for virulence, a complete and clear understanding of these interactions and their underlying mechanisms has not yet been achieved. Hence, our understanding on how viral RNA exploits cellular mechanisms for its own advantage is still limited.

Several laboratories have independently mapped the SARS-CoV-2 RNA–host protein interactome in mammalian cell lines that are permissive to SARS-CoV-2 infection (12–15). In these studies, the interactome was analyzed using an RNA-centric approach to specifically purify the viral RNA at 8h-24h post-infection, a time frame in which the virus is actively replicating and produces viral proteins that hijack the host innate immune response (16, 17). Many host RNA-binding proteins (RBPs) have been reported to interact with the SARS-CoV-2 RNA. These RBPs are involved in multiple biological processes such as RNA splicing, RNA metabolism, nonsense-mediated decay, translation and viral processing. However, the

interactions between host proteins and viral RNAs at earlier stages of the viral life cycle and the RNA regions specifically bound by the host proteins remain largely undetected. With this work, we aim to fill these gaps by expressing defined SARS-CoV-2 RNA fragments in human living cells and analyzing those interactions with the host proteins that might occur in the absence of viral proteins. In particular, we focused on the highly structured regions containing the 5′ and the 3′ ends of SARS-CoV-2, which we previously predicted to be interaction hot spots for the host proteins, as observed also for other coronaviruses (18).

To thoroughly investigate the interactome of these key SARS-CoV-2 RNA regions, we used an approach based on the combination of the *in house* algorithm *cat*RAPID and a well-established proximity-ligation approach (19), RNA–Protein Interaction Detection-Mass Spectrometry (RaPID-MS) (20). In addition to the host proteins already reported, we unveiled other RBPs which may have key functions in the early phases of viral infection by interacting with SARS-CoV-2 genomic RNA. Interestingly, we found also the enzyme pseudouridine synthase 7 (PUS7), an RNA modifier which catalyzes the pseudouridylation of RNA transcripts, highlighting a cellular mechanism in charge of introducing post-transcriptional modification on viral RNA. Nanopore direct RNA sequencing of SARS-CoV-2 RNA from infected mammalian cells confirmed the presence of many RNA modifications, including pseudouridylation sites on the viral RNAs. In particular, we found PUS7 consensus sequences heavily modified in the SARS-CoV-2 subgenomic RNAs, with one present in the stem loop 2 of the SARS-CoV-2 5'UTR, within the Transcription Regulatory Sequence - Leader (TRS-L). Biochemical and biophysical assays support a role for this unique modification by influencing both RNA structure and interaction propensity. In conclusion, we propose that the interaction and the activity of cellular pseudouridine synthases may represent a new mechanism able to influence SARS-CoV-2 life cycle.

Materials and methods

Preliminary predictions of protein-RNA interactions

The c*at*RAPID algorithm (21, 22) was used to identify the binding propensity of the 10 SARS-CoV-2 500 bp fragments against a library of 2064 human RBPs. For each fragment, only interactions with associated propensity score > 85th percentile of the fragment score distribution were retained and subsequently, the average score was calculated.

Reagents and plasmids

The following plasmids were purchased from Addgene and used for the RaPID-MS assay: BoxB-plasmid (Addgene #107253); BoxB-EDEN15 plasmid (#107252); BASU RaPID plasmid (#107250).

The following plasmids were a gift from Paul Khavari (Addgene plasmid #107253; #107252; #107250. http://n2t.net/addgene: 50917; 107252; 107250. RRID : Addgene_50917; Addgene 107252; Addgene 107250) (Ramanathan et al., 2018). SARS-CoV-2 RNA fragments, derived from the SARS-CoV-2 RNA sequence MN908947.3, were synthesized by GeneArt (ThermoFisher Scientific) according to the sequence reported in **Supplementary Table S1**, flanked by Esp3I restriction sites. The same strategy was adopted for the Scramble control sequence: 3′-CGTCTCCGCTTTCGACGACAATTTATAAAGACAGCGGTCGAGGGAAGATTTACGAGTTGAATCGAGATGCGCTGATTCGACGCAGTGTCGCGTTGTGGTGAGGTAAATTGATAGGTGTATTTTGCGAGATACAGTGATGAACACTTCATTAACAACATGATTTATACGACGATTACTAGAATTATGAAAAATGAGTCATCTACAAGCGCGTTTTTACATTGCCGTGGTTAATCGTAAGGATAGCACAGTTAACAGCGGACCCCGGCGGACTCGGCCCTATCTGAACGAATTGAGCTCCGTTCGAAATATCTAGTGAATGACCCTCCCCACGTGCCTTGATAAGCCGTGGTATTTCGTATCATACAAGTTCCAGAAGGATGGTTCAACATAGTAGGGTACCGACTGGATAGAACAAACTACTCATGTTTTCGCCGGGGGACGAACGGTAAGCTCCGCTGGGTTGACTTCTTGACCAAAGTATTTGGGTATCCAAACAGTGCCGTTAACAGCCAAGCTAGAGACG-5′.

The RNA fragments were then cloned into the pLEX BoxB-plasmid using the Esp3I sites, as described in literature (20).

pDONR207 SARS-CoV-2 NSP1 was a gift from Fritz Roth (Addgene plasmid # 141255; http://n2t.net/addgene:141255; RRID: Addgene_141255). NSP1 coding sequence was cloned into the mammalian expression pCDNA5-FRT/TO-2xSTREP-3xHA vector, gently provided by G.Superti Furga, through the Gateway cloning system (Invitrogen).
The human coding sequence of PUS7 (variant 1, C-ter tag) cloned into the pDONOR221 vector was purchased from DNASU (HsCD00867933). The sequence was mutagenized to obtain the N-ter tag canonical variant 2 through the QuikChange II XL Site-Directed Mutagenesis Kit (Agilent, 200521) using the following DNA primers: 5'-GGGAAAAAGGCTTTGGCAAATCCAAGAAAACATTCTTGGCC- 3' and 5'-GGCCAAGAATGTTTTCTTGGATTTGCCAAAGCCTTTTTCCC- 3'. The same strategy was used to generate the catalytic inactive mutant PUS7 D294A, using the following primers: 5'-ATTCTCCTACATGGGAACCAAAGCTAAAAGGGCTATAACAGTTC -3' and 5' -GAACTGTTATAGCCCTTTTAGCTTTGGTTCCCATGTAGGAGAAT -3'.
The resultant cDNA was sequence verified and through the Gateway cloning system cloned into the pDEST-cDNA5-FRT/TO-3*Flag-APEX2 N-term, which was a gift from Benjamin Blencowe (Addgene plasmid # 182925; http://n2t.net/addgene :182925; RRID: Addgene_182925).

The following antibodies were used in this study**:** Streptavidin-HRP (Cat. #3999) from Cell Signaling Technology. Anti-Vinculin (Clone H Vin 1 0,2 Ml; Cat. #V9131) from Merck. Anti-HA-11 epitope tag, clone 16B12 from Biolegend (Cat #901501). Anti-Flag-M2 (Merck, F3165).

Cell lines

HEK293T cells were grown in DMEM with Glucose and L-Glutamine (Lonza, BE-12-604Q) supplemented with 10% fetal bovine serum (FBS) Tetracycline free (Euroclone, ECS01822) and 100 U/ml Penicillin and Streptomycin. CaCo-2 cell lines were cultivated in Minimum Essential Medium Eagle (MEM) (Merck, M4655) complemented with 20% FBS Tetracycline free, 2 mM L-Glutamine (Lonza, BE17605E), 1 mM Sodium Pyruvate (Lonza, BE13115E), 0.1 mM not essential amino acids (Lonza, BE13114E) and 100 U/ml Penicillin and Streptomycin (Euroclone, ECB3001D).

Virus isolation and cell infection

SARS-CoV-2 virus was isolated from a mildly symptomatic SARS-CoV-2 infected patient, as described in Ugolini et al. (6). CaCo-2 cells were infected at 80% confluency into a 25 cm^2 tissue culture flask with SARS-CoV-2 at 0.1 Multiplicity of infection (MOI). After 1 h adsorption at 37 °C, cells were washed with PBS, and further cultured at 37 °C for 48 h with 4% FBS. After a PBS wash, enzymatic dissociation was performed for 4–6 min at 37 °C in 1 ml TrypLE (Invitrogen), then cell pellets were washed with ice-cold PBS and lysed with 1 ml of TRIzol (Invitrogen). The samples were stored at -80 °C for subsequent RNA extraction.

RNA extraction and nanopore direct RNA sequencing

RNA was extracted from CaCo-2 cells infected with SARS-CoV-2 using the RNeasy Mini kit (Qiagen) including an in-column DnaseI treatment step, following the manufacturer protocol. The isolated RNA was then processed by direct-RNA sequencing protocol as described in Ugolini et al. (6).

RaPID assay

The RaPID protocol was performed as reported (20) and slightly adapted as described below. HEK293T cells were transfected with plasmid vectors expressing λN-HA-BASU and one of the BoxB-RNA fragments using Lipofectamine™ 3000 transfection reagent (L3000001; ThermoFisher Scientific), according to vendor's instructions. After 48h from transfection, medium was changed and replaced with standard growth medium complemented with 200 µM biotin (Merck, B4639-1G) for 1 hour. Cells were harvested, washed once with PBS 1x and lysed with RIPA buffer (10mM Tris-HCl pH 8.0, 1%Triton-X100, 150mM NaCl, 0.1% SDS, 0.1% NaDeoxycholate, 1mM EDTA) supplemented with 1mM 1,4-Dithiothreitol (DTT), cOmplete™ Protease Inhibitor Cocktail (11697498001; Merck) and PhosSTOP™ (4906845001; Merck). Cell lysates were incubated for 15 minutes on ice and then centrifuged at 15000 g for 15 minutes. The supernatants containing the protein extracts were transferred into fresh 1.5 ml tubes and protein concentration was measured by Bio-Rad Protein Assay Kit using BSA as protein standard (5000002; Bio-Rad). From each sample, 3 mg of protein extract was taken and brought to the same volume (600uL) with RIPA buffer. 5% of input material was taken for further analysis and 150 µL of pre-washed Streptavidin Mag Sepharose® (GE28-9857-99; Merck) were added to the remaining material. Then, samples were rocked over night at 4 °C. The following day, beads were separated from the unbound fractions and 5% of each fraction was collected in fresh tubes. Beads containing the biotinylated proteins were washed 3 times

with 1 mL of Wash Buffer 1 (1% SDS supplemented with 1mM DTT, protease and phosphatase cocktail inhibitors); 1 time with Wash Buffer 2 (0.1% Na-DOC, 1% Triton X-100, 0.5M NaCl, 50mM HEPES pH 7.5, 1µM EDTA supplemented with 1mM DTT, protease and phosphatase cocktail inhibitors) and 1 time with Wash Buffer 3 (0.5% Na-DOC, 150mM NaCl, 0.5% NP-40, 10mM Tris-HCl, 1µM EDTA supplemented with 1mM DTT, protease and phosphatase cocktail inhibitors). All the washes were performed by rocking samples for 5 minutes at 4 °C. Finally, proteins were eluted with Laemmli buffer containing 100 mM DTT, boiled for 5 minutes at 95 °C and processed by *In-gel* digestion. The resultant peptide mixtures were analyzed by Nano-LC-MS/MS analysis (see **Supplementary methods**).

Data analysis of MS data

Proteins were identified and quantified using MaxQuant software v.1.6.0.16. using the Andromeda search engine (23, 24). In MaxQuant, the false discovery rate (FDR) of all peptide identifications was set to a maximum of 1%. The main search was performed with a mass tolerance of 6 ppm. Enzyme specificity was set to Trypsin/P. A maximum of 3 missed cleavages was permitted and the minimum peptide length was fixed at 7 amino acids. Carbamidomethylation of cysteines was set as a fixed modification. The 2021_01 version of the human UniProt reference proteome (UP000005640) was used for peptide identification. Proteins were profiled by quantitative label-free analysis, activating the label-free software MaxLFQ (23) and analyzed using Perseus software (25), plotting the LFQ values in a volcano plot graph, where the proteins enriched with each SARS-CoV-2 RNA fragments were compared to the Scramble RNA control. The p-Value was calculated by Perseus using a two-tailed *t*-Test. Missing values were replaced by the minimum detection value of the matrix. Proteins found as significantly enriched also with the EDEN15 RNA were removed from the final network. The interaction network generated with the statistically significant proteins was visualized using Cytoscape 3.8.1 (26).

Computational evaluation of experimental interactions

The *cat*RAPID algorithm was used to assess the predictive accuracy on the experimental RAPID dataset. Here, for each SARS-CoV-2 fragment-human protein pair, a fragmentation procedure was performed to identify the binding regions. The propensity score is calculated taking into consideration the maximum and the minimum score of all the sub-fragments computed by the algorithm (21) and computing their difference (27). The median LFQ value

for each protein-fragment pair was then normalized taking into account protein abundance in HEK293 cell lines from PAXdb (28) and protein length from UniProt database (29), using the following equation:

Normalized LFQ value = LFQ – $0.6*\log_2(\text{abundance})-0.0001*\log_2(\text{length})$.

This approach was introduced to reduce the experimental bias in LFQ scores with high protein abundance and length.

RNA folding analysis

Predictions of RNA folding were performed using the RNAfold algorithm hosted in ViennaRNA Web Services (30) using the default parameters and displaying the centroid secondary structures.

Gene Ontology (GO) Analysis

The GO analysis was performed on the 73 proteins specifically associated with the 10 SARS-CoV-2 RNA fragments and identified by RaPID-MS using the GOTERM biological process (BP) and molecular function (MF) present in the DAVID 6.8 Bioinformatics Resources (https://david-d.ncifcrf.gov/home.jsp) (31).

RNA secondary structure investigation

Circular dichroism analyses were performed on the RNA oligonucleotide "SL2" of sequence 5' -AACCAACUUUCGAUCUCUUGUAGAUCUGUUCU -3' synthesized by Eurofins, and on the pseudouridylated SL2 sequence 5' -AACCAACUUUCGAUCUCUUG(psi)UAGAUCUGUUCU -3', synthesized by Tebu-bio. The two versions of SL2 were prepared in a water stock at 100 µM and diluted to a concentration of 20 µM in a buffer containing 20 mM potassium phosphate at pH 7.2 and 150 mM KF. The RNA was then incubated for 1 hour at 30 °C to allow correct folding. Data acquisition of near-UV CD spectra was performed in one-millimeter path-length quartz cuvettes on a JASCO-1100 spectropolarimeter supplied with a constant N_2 flush at 3.0 L/min. The experiment was performed in triplicates.

Nanopore DRS and Nanocompore analysis

The nine DRS datasets were compared to the IVT dataset, using the NRCeq assembly and the methods in (6). The dataset of Vero E6 cells, Calu-3 and CaCo-2 sample 1 and 2 are derived from (6) (ENA ID: PRJEB48830); CaCo-2 sample 3 and 4 are deposited with this manuscript (ENA ID: PRJEB53497). Reads were grouped by cell type and their fastq files concatenated. Reads were resquiggled using F5C (v0.6) (32). IVT reads were separately mapped to each reference sgRNA using minimap2 (v2.17-r974-dirty) (33) with the following parameters: -ax map-ont -p 0 -N 10.

Nanocompore (v1.0.4) (34) was used to detect RNA modifications by comparing the IVT reads to a set of reads from each cell line for every reference sgRNA (**Suppl. Table 5, 6, 7**) with the subsequent parameters:

--fasta <NRCeq_assembly_fasta_file> --overwrite --downsample_high_coverage 5000 --allow_warnings --pvalue_thr 0.01 --min_coverage 30 --logit --nthreads 3 --bed <NRCeq_assembly_bed_file> .

Merging analysis

Individual significant k-mers from the different cell lines may be offset by one or two positions, therefore the significant k-mers were transformed into larger nucleotide ranges of nine nucleotides, that include the five nucleotides of the significant k-mer, plus the two neighboring nucleotides present at both its extremities. Any of these nine nucleotides may be a post-transcriptional modified nucleotide detected by Nanocompore. In addition, we minimized ambiguous mapping by counting significant k-mers found on any isoform of VME1 and NCAP as a single isoform according to their genomic coordinates.

Sites were compared to the ”high-confidence sites” defined by Fleming et al. (35) using the single-nucleotide genomic position.

Sites were manually inspected by plotting the distribution of signal intensity and dwell time for each position using functions implemented in Nanocompore (34).

Comparison with RaPID fragments

Both significant k-mers and sites were compared with the RaPID fragments. Genomic coordinates were used for all comparisons with RaPID fragments (**Suppl. Table 1**). In the case of k-mers, we defined a RapID fragment match when the first nucleotide of the k-mer

overlapped with any RaPID fragment, while for sites, we searched for every nucleotide of the site that could overlap any fragment. Genomic regions overlapping two fragments were identified with both fragment names. K-mers were assigned to each fragment based on the genomic position of their first nucleotide.

Genomic RNAs analysis

All infected datasets were mapped to the viral genome reference with minimap2 with the subsequent parameters:

-k 8 -w 1 -ax splice -g 30000 -G 30000 -A1 -B2 -O2,24 -E1,0 -C0 -z 400,200 --no-end-flt -F 40000 -N 32 --splice-flank=no --max-chain-skip=40 -un -p 0.7.

All the positive strand primary alignments, with no deletions, ranging from at least 45 and 29850 in genomic coordinates were extracted from all the datasets. These reads were compared to the IVT reads, pre-aligned to the viral reference genome, through Nanocompore with the subsequent parameters:

--downsample_high_coverage 2100 --allow_warnings --min_coverage 10 --logit.

The downsampling parameter was set to 2100 as it was obtained from the multiplication of the number of gRNA reads (140) for the number of IVT fragments (15), as downsampling is performed over all the dataset.

K-mers were evaluated nucleotide per nucleotide, to assess their overlap with IVT junctions, RaPID fragments, Fleming et al. sites, sgRNAs sites from the previous analysis and Single Nucleotide Polymorphisms (SNPs). The list of SNPs collects data from the literature (see **Supplementary Table 5**). Only k-mers not overlapping any SNP or IVT junction were considered. K-mers were considered significant if having an absolute LOR >=0.5 and a GMM p-value <=0.01.

We then investigated if, despite the low coverage noticed for gRNAs, a comparison between the sgRNAs-specific sites and the gRNAs-specific ones could be performed. Therefore, to make the comparison between the gRNAs and the sgRNAs datasets fair, we pulled together all the data from each cell line analyzed and selected sgRNAs uridine-containing significant k-mers present in at least one canonical transcript model (see **Supplementary Table 10**). To determine if the common sites were randomly selected or if such a number was to be statistically expected, we performed a hypergeometric test on common sites located after the TRS-B spike junction sequence, both for gRNAs and sgRNAs. The significance of the test (p-value = 1.44e-15, see github directory) suggests that, if we increased the coverage of the

gRNAs dataset, we might retrieve the same sites found for the sgRNAs in the region starting from the TRS-B spike junction sequence to the end of the viral genome. The test could not be performed on the region starting at the genomic zero coordinate to the TRS-B spike junction sequence, as, of course, sgRNAs do not cover this portion of the genome.

Results

Prediction of SARS-CoV-2 regions with the highest propensity to interact with host proteins

We previously reported that RNA regions with high structural content tend to tightly interact with a large number of proteins (36, 37), which suggests that structured regions of SARS-CoV-2 RNA genome might play a crucial role in its regulation (18). In line with this finding, we and others have observed that the first and last 1.5 kb of the viral genome recruit specific factors required for viral translation and replication, and that they act as targets for the intracellular host defense response (38, 39). Despite previous attempts, a comprehensive and reproducible map of RNA:host protein interactions is missing (40). We decided to accurately map the regions with the highest propensity to interact with host proteins and, accordingly, divided the initial as well as terminal 1.5 kb regions of the viral genome into 500 nucleotide-long RNA fragments. Each fragment overlaps the following one by 250 nucleotides. A total of five fragments for the 5′ region (numbered 1 to 5) and five for the 3′ region (numbered 6 to 10) was identified (**Supplementary Table 1** and **Fig. 1A**). Through the *cat*RAPID algorithm (21), which estimates the binding affinity of protein-RNA pairs (41, 42), we predicted the interactions of each fragment with human RBPs (over 2000 entries; **Supplementary Table 2**). In this analysis we did not consider the cell-specificity pattern of expression of the RBPs, as the calculations serve to provide an estimate of the protein binding ability of RNA fragments. We identified fragments 1 and 2 and fragments 9 and 10 as those harboring the highest scores (**Fig. 1B**). Considering that the 5′ and 3′ UTRs (fragments 1 and 10) have the strongest scores, we reasoned that these regions might also be the most functionally relevant (43). Altogether, these results provide important indications on the location of structural hot spots potentially relevant for the functional interactions between the virus and the host.

Identification of the human interactome for the 5′ and 3′ ends of SARS-CoV-2 genome

To analyze the RNA interaction with the host proteins experimentally, we exploited the RaPID-MS approach (20) using the 10 above-described fragments generated by *cat*RAPID (**Fig. 1A and Supplementary Fig. 1**). The fragments were cloned within BoxB sequences (**Materials and Methods**), used as tags for the RaPID assay, and were expressed into HEK293T cells as host cell line (**Fig. 1C**). As a negative control, we generated a scrambled sequence of 500 nt of similar GC content compared to the other fragments (named 'Scramble' henceforth). The

BoxB-tagged EDEN15 RNA was selected as positive control (20). We verified that: i) each plasmid expressing an RNA fragment was co-expressed with a plasmid coding for the biotin protein ligase λN-HA-BASU by FACS analysis (**Supplementary Fig. 2)**; ii) λN-HA-BASU was expressed correctly and able to biotinylate protein substrates in the presence of exogenous biotin by Western Blot (WB) analysis (**Supplementary Fig. 3A**). The resulting biotinylated proteins were purified under denaturing conditions using streptavidin beads, as confirmed by WB analysis (**Supplementary Fig. 3B**) and analyzed by liquid chromatography tandem mass spectrometry (LC-MS/MS). In total, 3 independent biological replicates of RaPID-MS were performed using the 10 SARS-CoV-2 fragments, as well as the Scramble and the EDEN15 control RNAs. We identified a total of 1296 proteins interacting with our selected fragments, a resource reported in **Supplementary Table 3**. A comparison between our dataset and the ones generated by previous studies indicates an overlap of about 40% (**Supplementary Table 3)**, demonstrating good agreement with literature (44, 45).

As we are interested in finding host proteins that specifically bind to SARS-CoV-2 regions rather than other RNAs, we compared the list of interactors obtained with each individual fragment with those obtained with our negative control Scramble (**Supplementary Fig. 4A and 4B, and Supplementary Table 3**). We removed RBPs that were found as interacting to the same extent of Scramble (**Supplementary Table 3)**, including PTBP1, G3BP1 and SYNCRIP, which were previously reported to specifically interact with the SARS-CoV-2 RNA (12–15). We were able to verify the specificity of our approach, as we identified CELF1 as the interactor of our positive control EDEN15 (**Supplementary Fig. 4B**), as expected (46). In total, we isolated 73 proteins significantly and specifically interacting with SARS-CoV-2 RNA fragments, hereafter called 'RaPID-MS dataset' (**Fig. 2**). In the RaPID-MS dataset, we have 10 proteins, previously reported as interactors of SARS-CoV-2 RNA, namely FAM120A, HAT1, LSG1, SHMT1, SYNE2 and the ribosomal proteins RPL14, RPL18A, RPL24, RPL35, RPS6. The remaining proteins were identified for the first time (**Supplementary Fig. 5A and 5B)**. The highest number of interacting proteins were in fragments 1, 4, 7 and 10. Conversely, fragment 2 displayed only two specific interactions (**Fig. 2A**).

To balance the possible bias that could be introduced by the RaPID technique, we considered protein length and abundance to compute the median enrichment value of each significant interactor (**Fig. 2B**). As for other studies of SARS-CoV-2 RNA interactome (40), we retrieved ribosomal proteins (N=7) among the most highly enriched interactors (**Fig. 2B**), explained by the fact that some of the fragments contain open reading frames (ORFs). We expected the

RaPID-MS dataset to be enriched for direct RNA-protein interactions. Indeed, 73% (53/73) interactors are annotated as RBPs according to the RBPome database (https://rbpbase.shiny.embl.de/) (green circles in **Fig. 2C**), while the remaining are members of the protein-complexes biotinylated in proximity of the target. We noted that while some proteins (N=26) were shared among 2 or more fragments, most interactions (N=47) are specific for just one RNA fragment (**Fig. 2C**). In some cases, the proteins interact with the overlapping regions included within the fragments, reinforcing our conclusions. Examples are WNK3 and AP4M1, which were found in association with the neighboring fragments 4 and 5 and fragments 7 and 8, respectively. Interestingly, three proteins (CEP350, RGPD1 and GPKOW) were exclusively interacting with the two terminal fragments (1 and 10), suggesting they might bind 5′ and 3′ ends either independently or by participating in viral RNA circularization, as recently described (47). Gene set enrichment analysis (GSEA) for gene ontology (GO) processes highlighted that the retrieved interactors belonged to cellular pathways involved in viral transcription, nonsense mediated mRNA decay (NMD), rRNA processing, translational initiation, mRNA splicing and mRNA export from the nucleus (**Supplementary Fig. 5C**).

Correlation between computationally predicted and experimentally validated RNA-host protein interactors

To properly compare predicted and experimental interactions, the median enrichment value of each experimentally identified interactor was normalized by protein length and abundance (**Supplementary Table 4**). We restricted the computational prediction analysis to human RBPs reported in the *cat*RAPID library (21), here named RaPID-RBP (**Supplementary Table 4**), and assessed to what extent the predicted binding propensities are in agreement with the experimental results obtained in the RAPID-MS dataset. In particular, we verified whether the strongest positive (i.e. interacting) and strongest negative (i.e. non-interacting) predicted protein-RNA pairs could be identified in the list of experimentally validated interactions. We evaluated the performance of our prediction using the Area Under the ROC Curve (AUC). Overall, *cat*RAPID reached an AUC of over 0.80. The AUC increased from 0.58 to 0.83 when applying the predictions to the experimental scores from the top 20% (i.e. the 20% strongest positives vs the 20% strongest negatives) to the top 2.5% (i.e. the 2.5% strongest positives vs the 2.5% strongest negatives) (**Fig. 3A** and **Supplementary Table 4**). When considering each fragment separately, the prediction performances increased even further for the terminal regions of the viral genome (fragments 1 and 10) and for some internal overlapping fragments

(fragments 4, 5, 6), reaching an AUC of 0.95 for the top 1% ranked experimental cases (**Fig. 3B and Supplementary Table 4**). Overall, this analysis suggests that predicted and the experimental data are well correlated.

Prompted by this observation, we exploited the *cat*RAPID score to produce a rank of the RBP interactome of SARS-CoV-2 RNA identified by RAPID-MS. The $\log_2$ *cat*RAPID score for all the possible viral RNA-host protein interactions was between 3.22 and 6.70, with a median value of 4.62 (**Fig. 3C and Supplementary Table 4**). The *cat*RAPID predicted RNA-protein interactions of each viral RNA fragment were differently distributed within these intervals (**Supplementary Fig. 6A-B**). In particular, proteins above a $\log_2$ *cat*RAPID score of 5.22, corresponding to the 85th percentile and the graphical inflection point of the distribution, are significantly enriched for interactions with fragment 10, 4 and 5 (**Fig. 3C and Supplementary Table 4**). These three fragments together displayed 373 predicted interactions that were not found with the RNA control Scramble, including 13 interactions involving 9 RBPs significantly associated with the SARS-CoV-2 RNA fragments by RaPID-MS (**Fig. 3D**). Among them, i) FAM120A, LSG1 and RPS6 that were already reported as host proteins reproducibly interacting with SARS-CoV-2 RNA (40); ii) the splicing factors GPKOW, SF3A1, NOP58 and the NDM factor SMG7 that were computationally predicted to specifically associate to the genomic viral RNA (48, 49); iii) the RNA modifier PUS7, that catalyzes the isomerization of uridine into pseudouridine in cellular tRNAs and mRNAs, reported here for the first time. In the RaPID-MS dataset, PUS7 specifically associates with fragments 1, 4, 7 and 10, all harboring a PUS7 consensus sequence. Importantly, pseudouridine has been found as the most abundant modification in SARS-CoV-2 RNA (10) and another member of the pseudouridine synthase family, PUS1, has been reported to interact with SARS-CoV-2 RNA (14). These observations strongly suggest that, in human cells, the formation of pseudouridine residues on SARS-CoV-2 genome could be catalyzed by the activity of cellular RNA-independent pseudouridine synthases, such as PUS1 and PUS7.

SARS-CoV-2 subgenomic RNAs bear multiple putative pseudouridylated sites

To directly investigate occurrence of RNA modifications in SARS-CoV-2, including pseudouridine sites, we employed Nanopore direct RNA sequencing (DRS) on cell lines infected with SARS-CoV-2 and analyzed the data with the Nanocompore software package (**Materials and Methods**). Nanocompore searches for RNA post-transcriptional modifications by comparing the ionic current features generated by Nanopore DRS from two different

experimental conditions: a test sample and a reference devoid of the modification of interest (or with a reduced number of them) (34). Due to the physics of nanopore sequencing, Nanocompore compares the ionic current features of five nucleotides (named as k-mer) at a time. To identify putative pseudouridine sites, we filtered the results by selecting k-mers that contained at least one uridine and identified as significant by Nanocompore (p-value ≤ 0.01, absolute value of the log odds ratio (LOR) ≥ 0.5). Initially, we used this technology to search for pseudouridine sites in the full-length genomic RNA (gRNA). We compared gRNA reads from CaCo2, Calu-3 and Vero E6 infected with SARS-CoV-2 with a baseline reference, which consists of unmodified SARS-CoV-2 RNA obtained by in vitro transcription (IVT) (50). Nanocompore identified 63 significant k-mers (p-value ≤ 0.01, abs (LOR) ≥ 0.5), of which 58 having a uridine at the center of the identified sequence or in the first or second neighbor nucleotide (**Fig. 4, Supplementary Fig. 7A-B and Supplementary Table 5**). Among them, 6 k-mers were located within RaPID fragments 3-4, 5, 7-8, 8-9 and 9-10 (**Supplementary Table 5** and **Supplementary Fig. 7B**). However, in the analyzed samples we noticed a coverage of only 140 reads spanning the whole 30 kb-long genome. Therefore, we extended our analysis to the subgenomic RNAs (sgRNAs) that were more abundant, as previously quantified (6). We analyzed all the 14 canonical sgRNAs of SARS-CoV-2 (**Fig. 4A** and **Supplementary Table 6-8**) used by the virus to translate the structural and accessory proteins required to produce new virions. In the attempt of characterizing sgRNAs, we have previously employed Nanopore ReCappable Sequencing (NRCeq), a new technique that can identify capped full-length RNAs (6). Thus, we compared viral sgRNA from infected Vero E6, CaCo-2 and Calu-3 cells to the unmodified SARS-CoV-2 IVT RNA, used as reference. Sequenced reads were aligned to the most up-to-date viral reference transcriptome (6) and the ionic current features from the RNA reads were realigned to each transcriptomic position of the reference using Nanopolish (51). Then, we used Nanocompore to identify RNA post-transcriptional modifications as marked by differences in the electrical signal between the viral sgRNA reads and the IVT RNA reads (p-value ≤ 0.01, abs (LOR) ≥ 0.5). In total, we identified 1164 (CaCo-2), 430 (Calu-3) and 627 (Vero E6) significant uridine-containing k-mers, distributed among the 14 canonical SARS-CoV-2 reference sgRNAs (**Fig. 4A, Supplementary Fig. 8** and **Tables 6-8**).

To focus on sites consistently modified across samples, we considered only significant modified uridines that were identified in at least two out of three SARS-CoV-2 infected cell lines (see **Materials and Methods**). Overall, we obtained 457 candidate modified regions across the 10 different SARS-CoV-2 sgRNAs encoding for canonical ORFs (**Fig. 4B)**. Two of

these regions were recently reported as a pseudouridylated sites in SARS-CoV-2 RNA (35) (**Supplementary Table 9**). In addition, from a supplementary analysis comparing the significant uridine-modified k-mers found in the SARS-CoV-2 sgRNAs and gRNAs, we retrieved an overlap of 17 sites (**Supplementary Fig. 7C** and **Supplementary Table 5**).

To shortlist the sites that have a higher likelihood of being pseudouridines (ψ) modified by PUS7, we selected those harboring the RNA consensus sequence of PUS7. We found that 53 sites had the more generic PUS7 consensus sequence UNUAR (red lollipops of **Fig. 4B)**, while eight had the more restrictive UGUAR motif (red lollipops with displayed sequence of **Fig. 4B)**. Six sites are located within RaPID fragments 1, 9 and 10, of which 1 and 10 were bound by PUS7 in our interactomic data (**Supplementary Table 9** and **Fig. 2C**). The modified sites containing the UGUAR consensus were manually inspected to confirm that the distributions of ionic current intensities and dwell times were different between the IVT and SARS-CoV-2 reads in a window of nine k-mers centered on the central U of the UGUAR motif (**Supplementary Fig. 9)**. Interestingly, we observed that three of these sites were present within the Transcription Regulatory Sequence - Leader (TRS-L) at the 5′ UTR of the sgRNAs encoding for ORF10, NS6, and VME1 (**Fig. 4B** and **Supplementary Table 9**).

PUS7-dependent pseudouridylation of TRS-L could modulate the binding of NSP1 to the viral 5'UTR

In total, Nanocompore analysis of the sgRNAs identified multiple putative post-transcriptionally modified sites in the TRS-L of SARS-CoV-2 (**Fig. 5A**). Conversely, the 5'UTR of the gRNA resulted devoid of any modification (**Supplementary Table 5**). One of the modified sites is within the m6A consensus motif DRACH, which has recently been reported to be methylated by METTL3 and to regulate the translation rate of SARS-CoV-2 proteins (52).

Another modified site is, instead, present in the PUS7 consensus sequence UGUAR, thus pointing towards a putative functional role of PUS7-dependent pseudouridylation of the SARS-CoV-2 5'UTR. We sought to understand the impact that the presence of a pseudouridine on the TRS-L might have. The modified site is present at the highly accessible uridine 54 (U54) within the stem loop 2 (SL2) (**Fig. 5A**), the most conserved region of the 5'UTR of Coronaviruses (53). We obtained the chemically synthesized SL2 RNA sequence either unmodified or carrying a pseudouridylated U54. Structural investigation performed via circular dichroism

revealed that the presence of ψ slightly increases the propensity of SL2 to form more stable secondary structures (**Supplementary Fig. 10A**), likely affecting the conformation of the SL2 pentaloop sequence (54). Since RNA secondary structure is one of the drivers of the interactions with proteins, we reasoned that the presence of a pseudouridylated SL2 might favor its interaction with its main binder, the viral protein NSP1. NSP1 binds to the SARS-CoV-2 leader sequence to protect and promote the translation of viral RNA transcripts (8, 55, 56). By means of biolayer interferometry, we estimated the impact of this pseudouridylation on the dissociation constant (K_d) with a recombinant NSP1. We observed that NSP1 is able to bind the *in vitro*-synthesized SL2 RNA sequence, with or without pseudouridylation, with a K_d in the high nanomolar range. However, the presence of ψ increased 2-fold the binding affinity, improving it from 300 nM to 170 nM (**Supplementary Fig. 10B**). The same trend of enhanced binding was assessed in cells through UV-RIP analysis performed with STREP-HA-tagged NSP1. We found that NSP1 is able to bind to the BoxB-RNA fragment 1 carrying the SARS-CoV-2 5'UTR (**Fig. 5B**). This RNA-protein interaction is enhanced by the ectopic expression of PUS7, while no differences were observed by the expression of a catalytic defective version of PUS7 (D294A) (57) (**Fig. 5B**). These results hint at a possible role of PUS7-mediated pseudouridylation of SL2, a region included in the 5' UTR that is typically bound by NSP1 to favor the translation of viral sgRNAs

Discussion

The identification of host proteins interacting with SARS-CoV-2 RNA is of fundamental importance for the development of molecules that block virus replication in human cells. Host proteins transiently interact both with the SARS-CoV-2 genomic and subgenomic RNAs to contribute to the different steps of the viral life cycle including translation, transcription and replication. Four independent studies have characterized the host and viral protein interactome of SARS-CoV-2 RNA genome in human cells at different time points, providing a global picture of the interactions occurring in infected cells. However, there is a relatively poor overlap (Jaccard index 0.051) between the interactors retrieved in these studies (44, 45), most likely attributable to the different permissive host cells used, the time of infection of the assay and the strength of interactions, and the type of crosslinking agents adopted to enrich for RNA-protein interactions (40). Moreover, in infected cells, host proteins can concurrently interact with the viral genomic RNA, subgenomic viral RNAs or both.

To unveil the SARS-CoV-2 RNA-host protein interactions occurring at the initial stages of viral infection in human cells, we decided to exploit the proximity ligation technology RaPID-MS (20), specifically focusing on the most predicted structured regions of SARS-CoV-2 RNAs. Even though our analysis was performed in non-infected cells, we identified the proteins FAM120A, HAT1, LSG1, RPL14, RPL18A, RPL24, RPL35, RPS6 SHMT1 and SYNE2 that were reported by the other studies that worked with SARS-CoV-2 infected cells, confirming the reliability of our approach. Compared to other approaches, our strategy has two main advantages: i) it does not involve any crosslinking step, minimizing the number of proteins indirectly associated with SARS-CoV-2 RNA; ii) it detects viral RNA-host protein interactions that occur during the early phase of infection, when viral proteins are poorly expressed and the virus is strictly dependent on the host proteins. For instance, among the retrieved SARS-CoV-2 RNA fragment-host protein interactions, we found three proteins, CEP350, GPKOW and RGPD1, that are exclusively interacting with the fragments containing the 5′ and 3′ UTR regions. These proteins may be required for the genome cyclization process that possibly regulates SARS-CoV-2 RNA discontinuous transcription, as recently reported (5, 47).

The fact that numerous RBPs were identified as non-specific RNA binders suggests that several interactions participate in host mechanisms triggered by the accumulation of viral RNAs in infected cells. The accumulation of viral RNA-host RBP interactions is likely caused by the formation of stress granules (SGs) that are required for the sequestration of viral RNA by the host to suppress viral replication (56). Indeed, several of the RBPs defined by us as promiscuous binders are involved in SGs formation and composition, such as the SG nucleating factors G3BP1 and G3BP2 and the proteins CAPRIN1, PUM1 and PUM2, respectively (58).

SARS-CoV-2 RNA is heavily post-transcriptionally modified (59), although the role of every RNA modification has not been fully described. In particular, m6A is known to be deposited by METTL3 on SARS-CoV-2 RNA and impair the binding of the viral RNA sensor RIG-I, thus bypassing the innate cellular immunity response (10, 60). Similarly, pseudouridine, which is documented to be the most abundant modification identified in SARS-CoV-2 RNA, has been hypothesized to aid the virus to hijack the host immune sensors and evade the innate immune response (10). The writer enzyme responsible for this modification is still unknown, although PUS1, a member of the pseudouridine synthase family, was found to be associated with SARS-CoV-2 RNA (14). In our RaPID-MS analysis we found that another member of the pseudouridine synthase family, PUS7, binds to the RNA fragments present at both the 5′ and

3′ end regions of the viral genome. The interacting RNA fragments carry the PUS7 consensus sequence UGUAR, corroborating the evidence that PUS7 might be one of the enzymes responsible for the pseudouridylation of SARS-CoV-2 RNA. In agreement, our DRS analysis on mammalian cells infected with SARS-CoV-2 revealed the presence of 53 sites containing PUS7 consensus sequences that display a significant difference in electrical signal distribution compared to the unmodified *in vitro* transcribed RNA. These data suggest that PUS7 is specifically recruited to its UGUAR consensus sequences in the 5' and 3' ends of the viral RNA, where it can catalyze the formation of pseudouridine, with potential regulatory effects on the RNA secondary structure or its interaction with host proteins.

Interestingly, three of the UGUAR sites identified by Nanocompore overlapped with a region of the TRS-L within stem loops 2 of the SARS-CoV-2 5′ UTR. We believe that this is likely a general event occurring for all the SARS-CoV-2 sgRNAs. The fact that we only detected the pseudouridine sites in three of the ten canonical SARS-CoV-2 ORF-grouped sgRNAs is probably due to the 3′ end coverage bias of DRS reads (61) and the minimum k-mer coverage requirements of Nanocompore. It is possible that this modified site would have been identified in all the sgRNAs if we had improved the 5′ end coverage. The pseudouridylation of the TRS-L may promote specific conformational changes that regulate the secondary structure as well as its interaction propensity towards host and viral proteins. We reported that PUS7-dependent pseudouridylation of SARS-CoV-2 SL2 favors the binding of NSP1 to the viral 5′ UTR. It is plausible to speculate that other proteins might change their binding affinity towards a post-transcriptionally modified 5′ UTR. However, the biological impact that this modification plays on SARS-CoV-2 RNA biology remains elusive.

The combination of the interactomic analyses and RNA modification profiling of SARS-CoV-2 RNA proposes a putative role for pseudouridine in SARS-CoV-2 biology that will require further experimental validation. Additional investigation is needed to determine whether the inhibition of PUS7 activity can be exploited to impair SARS-CoV-2 life cycle at any stage or whether other pseudouridine synthases can modify SARS-CoV-2 RNA, even within the PUS7 consensus sequence. In particular, PUS1 has been recently described to be able to modify UGUAG-containing RNA sequences (62). Considering the high viral copy numbers present in cells where SARS-CoV-2 is actively replicating, it is reasonable to imagine that several pseudouridine synthases might act on the viral RNAs at the same time. Thus, a comprehensive understanding of the molecular mechanism regulating the pseudouridylation of SARS-CoV-2

RNA could be crucial for the development of innovative strategies to impair viral replication in infected human cells.

Code Availability

All the analyses are available at the Github directory: https://github.com/camillaugolini-iit/SARS-pseudoU.git.

Data Availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (63) partner repository with the dataset identifier PXD034941". Reviewer account details: **Username:** reviewer_pxd034941@ebi.ac.uk; **Password:** 1Me1DL7H.
The RNA sequencing datasets generated in this study have been deposited at the European Nucleotide Archive (ENA). Data is available under the ID: PRJEB53497. Reviewers can temporarily visualize the UCSC Genome Browser tracks at the following webpage: https://genome.ucsc.edu/s/camilla.ugolini/SARS%2DCoV%2D2%20putative%20pseudo.

Acknowledgment

We thank Lorenzo di Tucci e Marco Rabozzi for developing an accelerated version of Nanopolish for preliminary analysis. The research leading to these results has been supported by the European Research Council (RIBOMYLOME_309545 and ASTRA_855923), the H2020 projects (IASIS_727658 and INFORE_825080). E.Z. received funding from MINDED fellowship of the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 754490.

Figure legends:

Figure 1. Experimental approach to investigate SARS-CoV-2 interactome.

A) Schematic representation of the SARS-CoV-2 RNA fragments selected to be studied with the RaPID-MS strategy. The scheme shows the fragments names, positions within the SARS-CoV-2 genome and degree of overlap between them. **B)** *cat*RAPID predictions of the 10 selected RNA fragments with the catalog of human RNA-binding proteins. Fragments belonging to the first and last 1.5 Kbp of SARS-CoV-2 genome are colored in different shades of pink and green, respectively. Error bars for each fragment correspond to the average value ± the standard error (SE). The gray dashed line indicates the trend of the *cat*RAPID score in the chart. **C)** Description of the technique RaPID-MS, in which the RNA fragment of interest is expressed in cells flanked by BoxB stem loops. BoxB is specifically recognized by the co-transfected λN peptide fused to the biotin ligase BASU. Upon biotin addition to the growth medium, BASU biotinylates the host protein interactors attracted by the RNA of interest.

Figure 2. Defining SARS-CoV-2 interaction network with the human proteome.

A) Number of specific interactors identified for each SARS-CoV-2 RNA fragment (labeled with the letter "F" follow by a number). **B)** Enrichment distribution of the specific interactors for all considered fragments compared to the control "Scramble". Data are normalized for protein length and abundance, as specified in the Materials and Methods section. **C)** The network of proteins identified by RaPID-MS. Only proteins significantly enriched over the control "Scramble" are displayed. Data are derived from the analysis of 3 independent biological replicates. In pink are displayed the RNA fragments, while in green the retrieved interactors. Proteins circled in green are RNA-binding proteins, according to the RBPome database (https://rbpbase.shiny.embl.de/).

Figure 3. *cat*RAPID performances on the RaPID-RBP datasets

A) *cat*RAPID performances on the RaPID-RBP dataset. The predictive power of the method is calculated for different percentages of the dataset. *cat*RAPID performance is calculated on LFQ experimental value, normalized by taking into account the abundance of the proteins in PAXdb and protein length in Uniprot. **B)** *cat*RAPID performance on the RaPID-RBP dataset, focusing on the single RNA fragments. For each fragment, the AUC at the different percentage of the dataset is shown. The performances are evaluated as in **A** and the fragments are ordered according to the respective genomic position. **C) Left.** Scatter chart of the analyzed RBPs

ranked according to their relative log_2 *cat*RAPID score. The dashed lines indicate the 85th percentile. **Right.** Boxplot representation of the whole distribution divided in the indicated percentiles. **D)** Scatter chart of the analyzed RBPs ranked according to their relative log2 *cat*RAPID score, displaying only proteins present in the 85th percentile of the distribution.

Figure 4. **Nanocompore analysis identified significant k-mers in the analyzed cell lines.** **A)** UCSC Genome Browser (*http://genome.ucsc.edu*) annotation of the modified k-mers containing a uridine found in SARS-CoV-2 infected CaCo-2, Calu-3 and Vero E6 cells and distributed over 14 different reference sgRNAs. From the top to the bottom, IVT and RaPID fragments, SARS-CoV-2 NRCeq assembly and RefSeq SARS-CoV-2 ORFs are present as a reference. **B)** Graphical representation of nucleotide ranges shared between at least two cell lines (red circles). Here, sites have been grouped per ORFs encoded by each reference sgRNA of the assembly and the sequence of those carrying an UGUAR motif is displayed.

Figure 5. PUS7 modifies the SL2 of SARS-CoV-2 TRS-L favoring the binding of NSP1. **A)** The secondary structure of SARS-CoV-2 TRS-L with the Nanocompore p-value (NS6 sgRNA vs IVT, GMM-logit test) overlaid as a color scale. The color represents the lowest p-value among those of the 5 overlapping k-mers. Only k-mers with a p-value ≤ 0.01 are colored according to the relative scale. **B) Left.** UV-RIP of HA-NSP1 co-expressed in HEK293 cells together with BoxB-Frag.1 expressing vector and either APEX2-FLAG, APEX2-FLAG-PUS7 or APEX2-FLAG-PUS7 D294A. The RNA bound to HA-NSP1 was eluted, retrotranscribed into cDNA and analyzed by RT-qPCR analysis using specific primers for RPLP0 and BoxB-Frag.1. The data shown are the average of 4 biological replicates (n =4) and are expressed as relative enrichment over each respective input. Statistical analysis was performed using non-parametric, 2-tailed *t*-Test. *= Values with a P value < 0.05. **Right.** Representative image of the four WB analysis conducted on an aliquot of the UV-RIP input material (10%), showing the expression levels of the exogenous protein transiently co-expressed in HEK293 cells.

References

1. Mistry,P., Barmania,F., Mellet,J., Peta,K., Strydom,A., Viljoen,I.M., James,W., Gordon,S. and Pepper,M.S. (2022) SARS-CoV-2 Variants, Vaccines, and Host Immunity. *Front. Immunol.*, **12**, 809244.

https://doi.org/10.3389/fimmu.2021.809244

2. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses (2020) The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol*, **5**, 536–544.

https://doi.org/10.1038/s41564-020-0695-z

3. Harrison,A.G., Lin,T. and Wang,P. (2020) Mechanisms of SARS-CoV-2 Transmission and Pathogenesis. *Trends in Immunology*, **41**, 1100–1115.

https://doi.org/10.1016/j.it.2020.10.004

4. Zhang,Y.-Z. and Holmes,E.C. (2020) A Genomic Perspective on the Origin and Emergence of SARS-CoV-2. *Cell*, **181**, 223–227.

https://doi.org/10.1016/j.cell.2020.03.035

5. Malone,B., Urakova,N., Snijder,E.J. and Campbell,E.A. (2022) Structures and functions of coronavirus replication–transcription complexes and their relevance for SARS-CoV-2 drug design. *Nat Rev Mol Cell Biol*, **23**, 21–39.

https://doi.org/10.1038/s41580-021-00432-z

6. Ugolini,C., Mulroney,L., Leger,A., Castelli,M., Criscuolo,E., Williamson,M.K., Davidson,A.D., Almuqrin,A., Giambruno,R., Jain,M., *et al.* (2022) Nanopore ReCappable sequencing maps SARS-CoV-2 5′ capping sites and provides new insights into the structure of sgRNAs. *Nucleic Acids Research*, **50**, 3475–3489.

https://doi.org/10.1093/nar/gkac144

7. Sola,I., Almazán,F., Zúñiga,S. and Enjuanes,L. (2015) Continuous and Discontinuous RNA Synthesis in Coronaviruses. *Annu. Rev. Virol.*, **2**, 265–288.

https://doi.org/10.1146/annurev-virology-100114-055218

8. Banerjee,A.K., Blanco,M.R., Bruce,E.A., Honson,D.D., Chen,L.M., Chow,A., Bhat,P., Ollikainen,N., Quinodoz,S.A., Loney,C., *et al.* (2020) SARS-CoV-2 Disrupts Splicing, Translation, and Protein Trafficking to Suppress Host Defenses. *Cell*, **183**, 1325-1339.e21.

https://doi.org/10.1016/j.cell.2020.10.004

9. Yin,X., Riva,L., Pu,Y., Martin-Sancho,L., Kanamune,J., Yamamoto,Y., Sakai,K., Gotoh,S., Miorin,L., De Jesus,P.D., *et al.* (2021) MDA5 Governs the Innate Immune Response to SARS-CoV-2 in Lung Epithelial Cells. *Cell Reports*, **34**, 108628.

https://doi.org/10.1016/j.celrep.2020.108628

10. Li,N., Hui,H., Bray,B., Gonzalez,G.M., Zeller,M., Anderson,K.G., Knight,R., Smith,D., Wang,Y., Carlin,A.F., *et al.* (2021) METTL3 regulates viral m6A RNA modification and host cell innate immune responses during SARS-CoV-2 infection. *Cell Reports*, **35**, 109091.

https://doi.org/10.1016/j.celrep.2021.109091

11. Kim,Y.-M. and Shin,E.-C. (2021) Type I and III interferon responses in SARS-CoV-2 infection. *Exp Mol Med*, **53**, 750–760.

https://doi.org/10.1038/s12276-021-00592-0

12. Kamel,W., Noerenberg,M., Cerikan,B., Chen,H., Järvelin,A.I., Kammoun,M., Lee,J.Y., Shuai,N., Garcia-Moreno,M., Andrejeva,A., *et al.* (2021) Global analysis of protein-RNA interactions in SARS-CoV-2-infected cells reveals key regulators of infection. *Molecular Cell*, **81**, 2851-2867.e7.

https://doi.org/10.1016/j.molcel.2021.05.023

13. Lee,S., Lee,Y., Choi,Y., Son,A., Park,Y., Lee,K.-M., Kim,J., Kim,J.-S. and Kim,V.N. (2021) The SARS-CoV-2 RNA interactome. *Molecular Cell*, **81**, 2838-2850.e6.

https://doi.org/10.1016/j.molcel.2021.04.022

14. Flynn,R.A., Belk,J.A., Qi,Y., Yasumoto,Y., Wei,J., Alfajaro,M.M., Shi,Q., Mumbach,M.R., Limaye,A., DeWeirdt,P.C., *et al.* (2021) Discovery and functional interrogation of SARS-CoV-2 RNA-host protein interactions. *Cell*, **184**, 2394-2411.e16.

https://doi.org/10.1016/j.cell.2021.03.012

15. Schmidt,N., Lareau,C.A., Keshishian,H., Ganskih,S., Schneider,C., Hennig,T., Melanson,R., Werner,S., Wei,Y., Zimmer,M., *et al.* (2021) The SARS-CoV-2 RNA–protein interactome in infected human cells. *Nat Microbiol*, **6**, 339–353.

https://doi.org/10.1038/s41564-020-00846-z

16. Lei,X., Dong,X., Ma,R., Wang,W., Xiao,X., Tian,Z., Wang,C., Wang,Y., Li,L., Ren,L., *et al.* (2020) Activation and evasion of type I interferon responses by SARS-CoV-2. *Nat Commun*, **11**, 3810.

https://doi.org/10.1038/s41467-020-17665-9

17. Xia,H., Cao,Z., Xie,X., Zhang,X., Chen,J.Y.-C., Wang,H., Menachery,V.D., Rajsbaum,R. and Shi,P.-Y. (2020) Evasion of Type I Interferon by SARS-CoV-2. *Cell Reports*, **33**, 108234.

https://doi.org/10.1016/j.celrep.2020.108234

18. Vandelli,A., Monti,M., Milanetti,E., Armaos,A., Rupert,J., Zacco,E., Bechara,E., Delli Ponti,R. and Tartaglia,G.G. (2020) Structural analysis of SARS-CoV-2 genome and predictions of the human interactome. *Nucleic Acids Research*, **48**, 11270–11283.

https://doi.org/10.1093/nar/gkaa864

19. Giambruno,R. and Nicassio,F. (2022) Proximity-dependent biotinylation technologies for mapping RNA-protein interactions in live cells. *Front. Mol. Biosci.*, **9**, 1062448.

https://doi.org/10.3389/fmolb.2022.1062448

20. Ramanathan,M., Majzoub,K., Rao,D.S., Neela,P.H., Zarnegar,B.J., Mondal,S., Roth,J.G., Gai,H., Kovalski,J.R., Siprashvili,Z., *et al.* (2018) RNA–protein interaction detection in living cells. *Nat Methods*, **15**, 207–212.

https://doi.org/10.1038/nmeth.4601

21. Armaos,A., Colantoni,A., Proietti,G., Rupert,J. and Tartaglia,G.G. (2021) *cat* RAPID *omics v2.0* : going deeper and wider in the prediction of protein–RNA interactions. *Nucleic Acids Research*, **49**, W72–W79.

https://doi.org/10.1093/nar/gkab393

22. Bellucci,M., Agostini,F., Masin,M. and Tartaglia,G.G. (2011) Predicting protein associations with long noncoding RNAs. *Nat Methods*, **8**, 444–445.

https://doi.org/10.1038/nmeth.1611

23. Cox,J., Hein,M.Y., Luber,C.A., Paron,I., Nagaraj,N. and Mann,M. (2014) Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Molecular & Cellular Proteomics*, **13**, 2513–2526.

https://doi.org/10.1074/mcp.M113.031591

24. Cox,J. and Mann,M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*, **26**, 1367–1372.

https://doi.org/10.1038/nbt.1511

25. Tyanova,S., Temu,T., Sinitcyn,P., Carlson,A., Hein,M.Y., Geiger,T., Mann,M. and Cox,J. (2016) The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods*, **13**, 731–740.

https://doi.org/10.1038/nmeth.3901

26. Su,G., Morris,J.H., Demchak,B. and Bader,G.D. (2014) Biological Network Exploration with Cytoscape 3. *Current Protocols in Bioinformatics*, **47**.

https://doi.org/10.1002/0471250953.bi0813s47

27. Lang,B., Armaos,A. and Tartaglia,G.G. (2019) RNAct: Protein–RNA interaction predictions for model organisms with supporting experimental data. *Nucleic Acids Research*, **47**, D601–D606.

https://doi.org/10.1093/nar/gky967

28. Wang,M., Herrmann,C.J., Simonovic,M., Szklarczyk,D. and Mering,C. (2015) Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics*, **15**, 3163–3168.

https://doi.org/10.1002/pmic.201400441

29. The UniProt Consortium, Bateman,A., Martin,M.-J., Orchard,S., Magrane,M., Agivetova,R., Ahmad,S., Alpi,E., Bowler-Barnett,E.H., Britto,R., *et al.* (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, **49**, D480–D489.

https://doi.org/10.1093/nar/gkaa1100

30. Kerpedjiev,P., Hammer,S. and Hofacker,I.L. (2015) Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics*, **31**, 3377–3379.

https://doi.org/10.1093/bioinformatics/btv372

31. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, **4**, 44–57.

https://doi.org/10.1038/nprot.2008.211

32. Gamaarachchi,H., Lam,C.W., Jayatilaka,G., Samarakoon,H., Simpson,J.T., Smith,M.A. and Parameswaran,S. (2020) GPU accelerated adaptive banded event alignment for rapid comparative nanopore signal analysis. *BMC Bioinformatics*, **21**, 343.

https://doi.org/10.1186/s12859-020-03697-x

33. Li,H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.

https://doi.org/10.1093/bioinformatics/bty191

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC6137996

34. Leger,A., Amaral,P.P., Pandolfini,L., Capitanchik,C., Capraro,F., Miano,V., Migliori,V., Toolan-Kerr,P., Sideri,T., Enright,A.J., *et al.* (2021) RNA modifications detection by comparative Nanopore direct RNA sequencing. *Nat Commun*, **12**, 7198.

https://doi.org/10.1038/s41467-021-27393-3

35. Fleming,A.M., Mathewson,N.J., Howpay Manage,S.A. and Burrows,C.J. (2021) Nanopore Dwell Time Analysis Permits Sequencing and Conformational Assignment of Pseudouridine in SARS-CoV-2. *ACS Cent. Sci.*, **7**, 1707–1717.

https://doi.org/10.1021/acscentsci.1c00788

36. Sanchez de Groot,N., Armaos,A., Graña-Montes,R., Alriquet,M., Calloni,G., Vabulas,R.M. and Tartaglia,G.G. (2019) RNA structure drives interaction with proteins. *Nat Commun*, **10**, 3246.

https://doi.org/10.1038/s41467-019-10923-5

37. Cerase,A., Calabrese,J.M. and Tartaglia,G.G. (2022) Phase separation drives X-chromosome inactivation. *Nat Struct Mol Biol*, **29**, 183–185.

https://doi.org/10.1038/s41594-021-00697-0

38. Schubert,K., Karousis,E.D., Jomaa,A., Scaiola,A., Echeverria,B., Gurzeler,L.-A., Leibundgut,M., Thiel,V., Mühlemann,O. and Ban,N. (2020) SARS-CoV-2 Nsp1 binds the ribosomal mRNA channel to inhibit translation. *Nat Struct Mol Biol*, **27**, 959–966.

https://doi.org/10.1038/s41594-020-0511-8

39. Brant,A.C., Tian,W., Majerciak,V., Yang,W. and Zheng,Z.-M. (2021) SARS-CoV-2: from its discovery to genome structure, transcription, and replication. *Cell Biosci*, **11**, 136.

https://doi.org/10.1186/s13578-021-00643-z

40. Baggen,J., Vanstreels,E., Jansen,S. and Daelemans,D. (2021) Cellular host factors for SARS-CoV-2 infection. *Nat Microbiol*, **6**, 1219–1232.

https://doi.org/10.1038/s41564-021-00958-0

41. Cid-Samper,F., Gelabert-Baldrich,M., Lang,B., Lorenzo-Gotor,N., Ponti,R.D., Severijnen,L.-A.W.F.M., Bolognesi,B., Gelpi,E., Hukema,R.K., Botta-Orfila,T., *et al.* (2018) An Integrative Study of Protein-RNA Condensates Identifies Scaffolding RNAs and Reveals Players in Fragile X-Associated Tremor/Ataxia Syndrome. *Cell Reports*, **25**, 3422-3434.e7.

https://doi.org/10.1016/j.celrep.2018.11.076

42. Agostini,F., Cirillo,D., Bolognesi,B. and Tartaglia,G.G. (2013) X-inactivation: quantitative predictions of protein interactions in the Xist network. *Nucleic Acids Research*, **41**, e31–e31.

https://doi.org/10.1093/nar/gks968

43. Armaos,A., Zacco,E., Sanchez de Groot,N. and Tartaglia,G.G. (2021) RNA-protein interactions: Central players in coordination of regulatory networks. *BioEssays*, **43**, 2000118.

https://doi.org/10.1002/bies.202000118

44. Vandelli,A., Vocino,G. and Tartaglia,G.G. (2022) Phase Separation Drives SARS-CoV-2 Replication: A Hypothesis. *Front. Mol. Biosci.*, **9**, 893067.

https://doi.org/10.3389/fmolb.2022.893067

45. Iselin,L., Palmalux,N., Kamel,W., Simmonds,P., Mohammed,S. and Castello,A. (2022) Uncovering viral RNA–host cell interactions on a proteome-wide scale. *Trends in Biochemical Sciences*, **47**, 23–38.

https://doi.org/10.1016/j.tibs.2021.08.002

46. Edwards,J.M., Long,J., de Moor,C.H., Emsley,J. and Searle,M.S. (2013) Structural insights into the targeting of mRNA GU-rich elements by the three RRMs of CELF1. *Nucleic Acids Res*, **41**, 7153–7166.

https://doi.org/10.1093/nar/gkt470

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3737555

47. Ziv,O., Price,J., Shalamova,L., Kamenova,T., Goodfellow,I., Weber,F. and Miska,E.A. (2020) The Short- and Long-Range RNA-RNA Interactome of SARS-CoV-2. *Molecular Cell*, **80**, 1067-1077.e5.

https://doi.org/10.1016/j.molcel.2020.11.004

48. Horlacher,M., Oleshko,S., Hu,Y., Ghanbari,M., Cantini,G., Schinke,P., Vergara,E.E., Bittner,F., Mueller,N.S., Ohler,U., *et al.* (2021) Computational Mapping of the Human-SARS-CoV-2 Protein-RNA Interactome Bioinformatics.

https://doi.org/10.1101/2021.12.22.472458

49. Akbari,B., Ahmadi,E., Zamir,M.R., Shaker,M.S. and Noorbakhsh,F. (2022) SARS-CoV-2 Helicase might interfere with cellular nonsense-mediated RNA decay, insights from a bioinformatics study Bioinformatics.

https://doi.org/10.1101/2022.05.30.494036

50. Kim,D., Lee,J.-Y., Yang,J.-S., Kim,J.W., Kim,V.N. and Chang,H. (2020) The Architecture of SARS-CoV-2 Transcriptome. *Cell*, **181**, 914-921.e10.

https://doi.org/10.1016/j.cell.2020.04.011

51. Simpson,J.T., Workman,R.E., Zuzarte,P.C., David,M., Dursi,L.J. and Timp,W. (2017) Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods*, **14**, 407–410.

https://doi.org/10.1038/nmeth.4184

52. Aly,A., Scott,G., Calderon,M. and Haghighi,A.P. (2022) N6-Adenosine Methylation of SARS-CoV-2 5'-UTR Regulates Translation Molecular Biology.

https://doi.org/10.1101/2022.10.17.512569

53. Miao,Z., Tidu,A., Eriani,G. and Martin,F. (2021) Secondary structure of the SARS-CoV-2 5'-UTR. *RNA Biology*, **18**, 447–456.

https://doi.org/10.1080/15476286.2020.1814556

54. Liu,P., Li,L., Keane,S.C., Yang,D., Leibowitz,J.L. and Giedroc,D.P. (2009) Mouse Hepatitis Virus Stem-Loop 2 Adopts a uYNMG(U)a-Like Tetraloop Structure That Is Highly Functionally Tolerant of Base Substitutions. *J Virol*, **83**, 12084–12093.

https://doi.org/10.1128/JVI.00915-09

55. Vankadari,N., Jeyasankar,N.N. and Lopes,W.J. (2020) Structure of the SARS-CoV-2 Nsp1/5′-Untranslated Region Complex and Implications for Potential Therapeutic Targets, a Vaccine, and Virulence. *J. Phys. Chem. Lett.*, **11**, 9659–9668.

https://doi.org/10.1021/acs.jpclett.0c02818

56. Mendez,A.S., Ly,M., González-Sánchez,A.M., Hartenian,E., Ingolia,N.T., Cate,J.H. and Glaunsinger,B.A. (2021) The N-terminal domain of SARS-CoV-2 nsp1 plays key roles in suppression of cellular gene expression and preservation of viral gene expression. *Cell Rep*, **37**, 109841.

https://doi.org/10.1016/j.celrep.2021.109841

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC8481097

57. Guzzi,N., Cieśla,M., Ngoc,P.C.T., Lang,S., Arora,S., Dimitriou,M., Pimková,K., Sommarin,M.N.E., Munita,R., Lubas,M., *et al.* (2018) Pseudouridylation of tRNA-Derived Fragments Steers Translational Control in Stem Cells. *Cell*, **173**, 1204-1216.e26.

https://doi.org/10.1016/j.cell.2018.03.008

58. Onomoto,K., Yoneyama,M., Fung,G., Kato,H. and Fujita,T. (2014) Antiviral innate immunity and stress granule responses. *Trends in Immunology*, **35**, 420–428.

https://doi.org/10.1016/j.it.2014.07.006

59. Izadpanah,A., Rappaport,J. and Datta,P.K. (2022) Epitranscriptomics of SARS-CoV-2 Infection. *Front. Cell Dev. Biol.*, **10**, 849298.

https://doi.org/10.3389/fcell.2022.849298

60. Burgess,H.M., Depledge,D.P., Thompson,L., Srinivas,K.P., Grande,R.C., Vink,E.I., Abebe,J.S., Blackaby,W.P., Hendrick,A., Albertella,M.R., *et al.* (2021) Targeting the m ^{6} A RNA modification pathway blocks SARS-CoV-2 and HCoV-OC43 replication. *Genes Dev.*, **35**, 1005–1019.

https://doi.org/10.1101/gad.348320.121

61. Workman,R.E., Tang,A.D., Tang,P.S., Jain,M., Tyson,J.R., Razaghi,R., Zuzarte,P.C., Gilpatrick,T., Payne,A., Quick,J., *et al.* (2019) Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat Methods*, **16**, 1297–1305.

https://doi.org/10.1038/s41592-019-0617-2

62. Purchal,M.K., Eyler,D.E., Tardu,M., Franco,M.K., Korn,M.M., Khan,T., McNassor,R., Giles,R., Lev,K., Sharma,H., *et al.* (2022) Pseudouridine synthase 7 is an opportunistic enzyme that binds and modifies substrates with diverse sequences and structures. *Proc. Natl. Acad. Sci. U.S.A.*, **119**, e2109708119.

https://doi.org/10.1073/pnas.2109708119

63. Perez-Riverol,Y., Bai,J., Bandla,C., García-Seisdedos,D., Hewapathirana,S., Kamatchinathan,S., Kundu,D.J., Prakash,A., Frericks-Zipper,A., Eisenacher,M., *et al.* (2022)

The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Research*, **50**, D543–D552.

https://doi.org/10.1093/nar/gkab1038

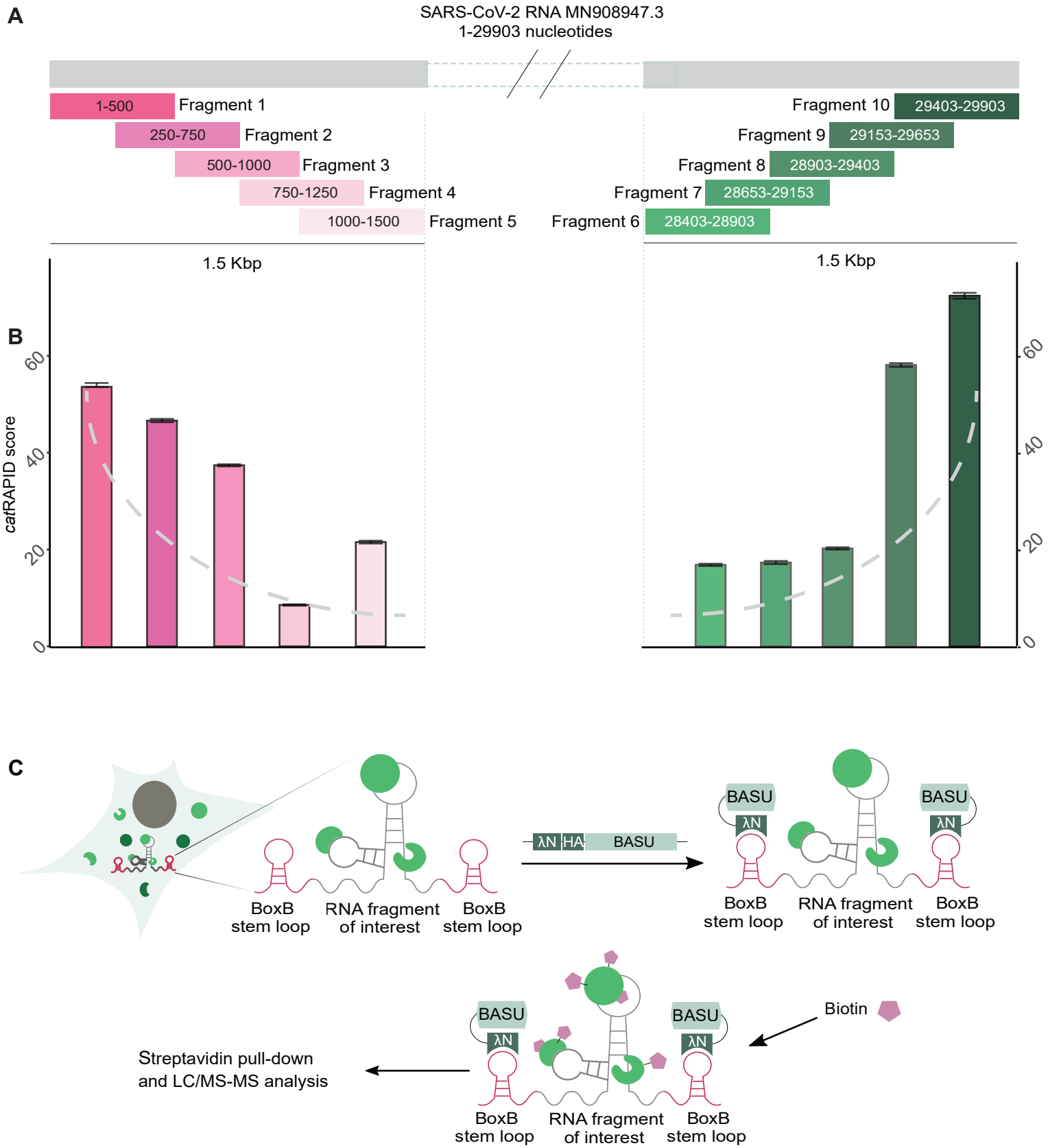

A
SARS-CoV-2 RNA MN908947.3
1-29903 nucleotides
1-500
Fragment 1
250-750
Fragment 2
500-1000
Fragment 3
750-1250
Fragment 4
1000-1500
Fragment 5
Fragment 6
28403-28903
Fragment 7
28653-29153
Fragment 8
28903-29403
Fragment 9
29153-29653
Fragment 10
29403-29903
1.5 Kbp
1.5 Kbp
B
catRAPID score
0
20
40
60
C
BoxB
stem loop
RNA fragment
of interest
BoxB
stem loop
λN
HA
BASU
BASU
λN
BASU
λN
Biotin
Streptavidin pull-down
and LC/MS-MS analysis

Figure 1.

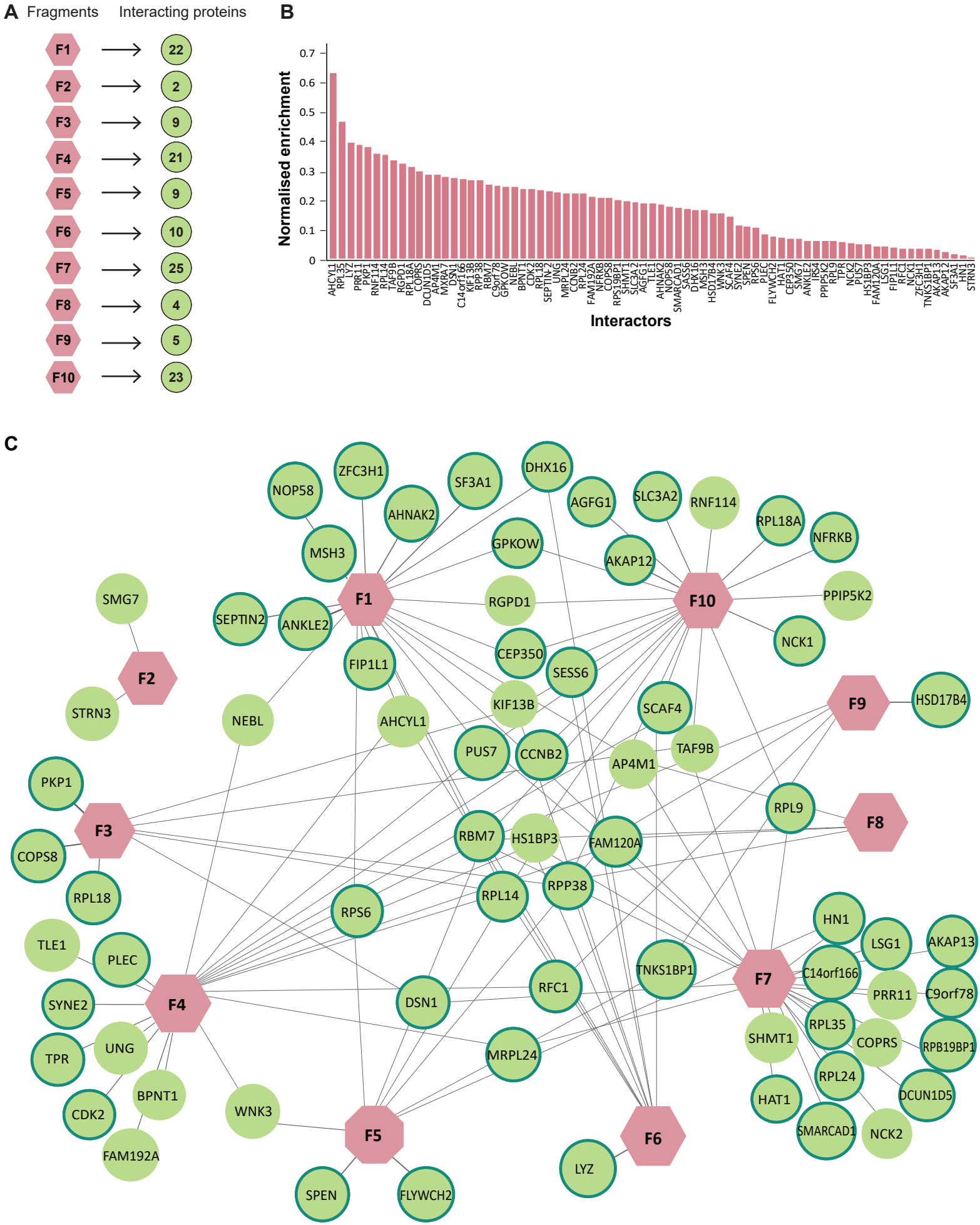

A
Fragments
Interacting proteins
F1 → 22
F2 → 2
F3 → 9
F4 → 21
F5 → 9
F6 → 10
F7 → 25
F8 → 4
F9 → 5
F10 → 23
B
Normalised enrichment
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
AHCYL1 RPL35 LYZ PRR11 PKP1 RNF114 RPL14 TAF9B RGPD1 RPL18A COPRS DCUN1D5 AP4M1 MYRA7 DSN1 C14orf166 KIF13B RPP38 RBM7 C9orf78 GPKOW NEBL BPNT1 CDK2 RPL18 SEPTIN-2 UNG MRPL24 CCNB2 RPL24 FAM192A NFRKB COPS8 RPS19BP1 SHMT1 SLC3A2 AGFG1 TLE1 AHNAK2 NOP58 SMARCAD1 SASS6 DHX16 MSH3 HSD17B4 WNK3 SCAF4 SYNE2 SPEN RPS6 PLEC FLYWCH2 HAT1 CEP350 SMG7 ANKLE2 IRS4 PPIP5K2 RPL9 TPR NCK2 PUS7 HS1BP3 FAM120A LSG1 FIP1L1 RFC1 NCK1 ZFC3H1 TNKS1BP1 AKAP13 AKAP12 SF3A1 HN1 STRN3
Interactors
C
F1
F2
F3
F4
F5
F6
F7
F8
F9
F10
ZFC3H1
NOP58
SF3A1
DHX16
AHNAK2
AGFG1
SLC3A2
RNF114
RPL18A
NFRKB
MSH3
GPKOW
AKAP12
SMG7
SEPTIN2
ANKLE2
RGPD1
PPIP5K2
NCK1
FIP1L1
CEP350
SESS6
STRN3
NEBL
AHCYL1
KIF13B
SCAF4
HSD17B4
PUS7
CCNB2
TAF9B
AP4M1
PKP1
RPL9
COPS8
RBM7
HS1BP3
FAM120A
RPL18
RPP38
RPL14
RPS6
HN1
LSG1
AKAP13
TLE1
PLEC
TNKS1BP1
C14orf166
RFC1
PRR11
C9orf78
SYNE2
DSN1
SHMT1
RPL35
COPRS
RPB19BP1
TPR
UNG
MRPL24
RPL24
DCUN1D5
BPNT1
HAT1
CDK2
WNK3
SMARCAD1
NCK2
FAM192A
LYZ
SPEN
FLYWCH2

Figure 2.

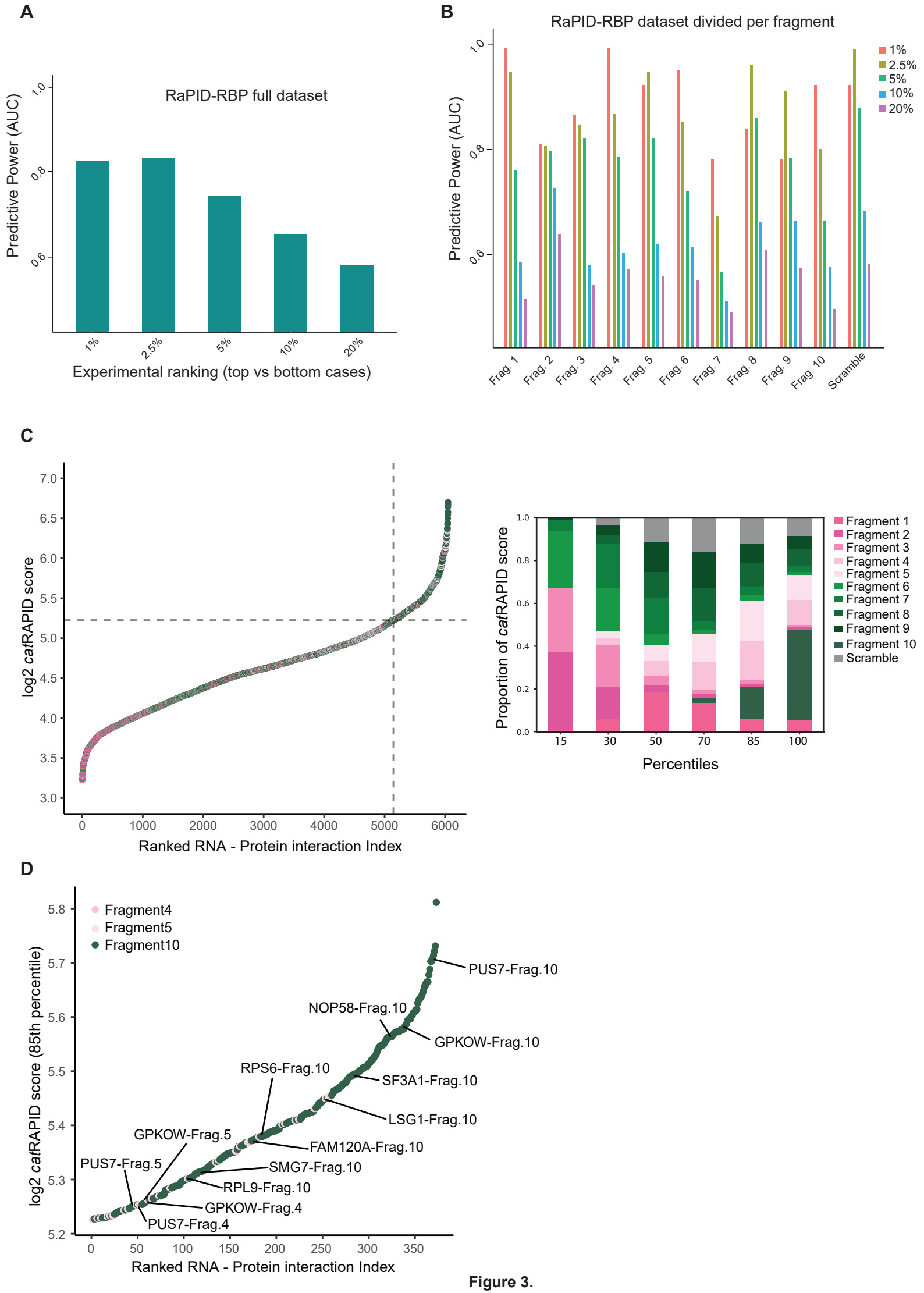
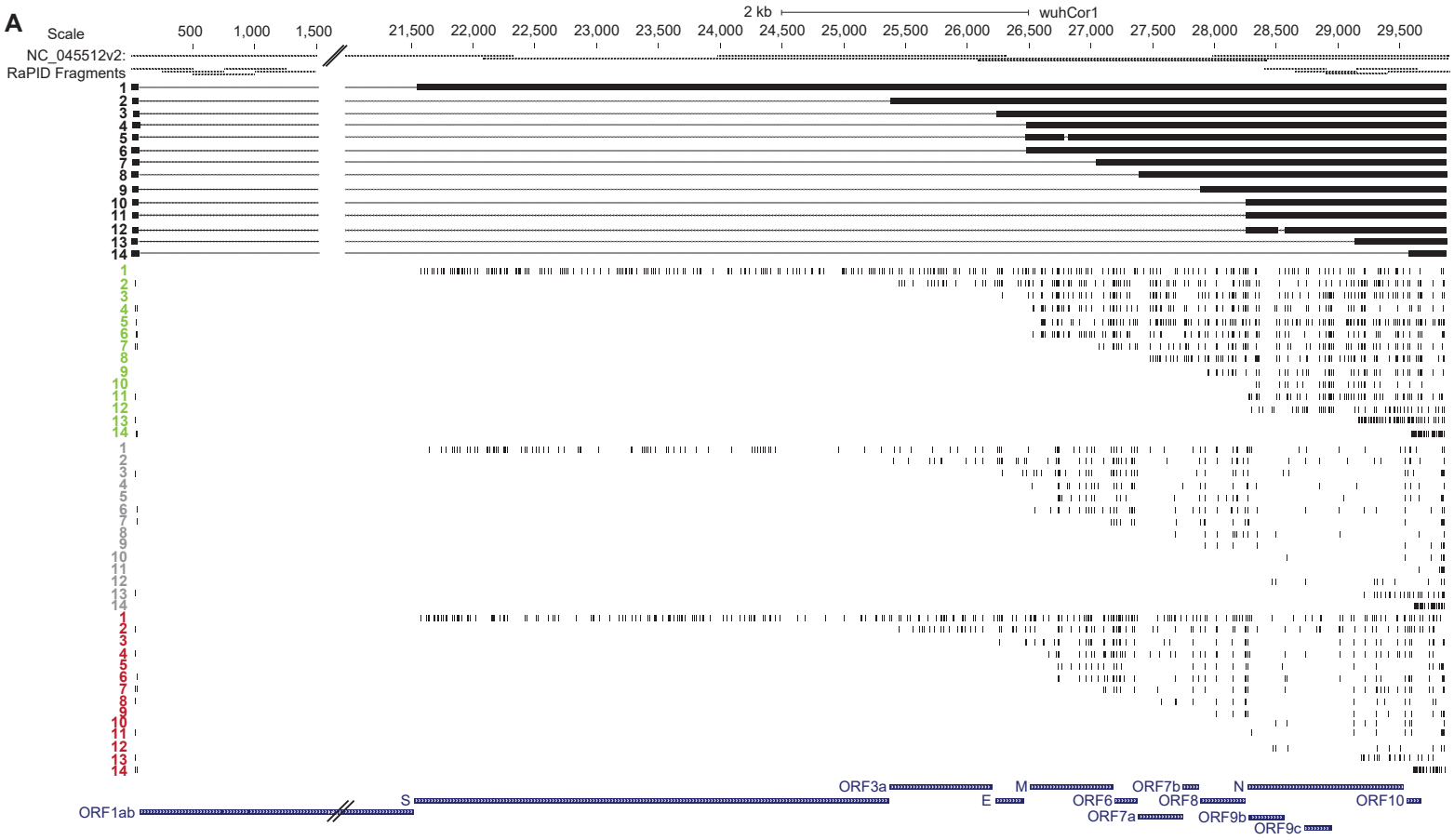

A
RaPID-RBP full dataset
Predictive Power (AUC)
1.0
0.8
0.6
1%
2.5%
5%
10%
20%
Experimental ranking (top vs bottom cases)
B
RaPID-RBP dataset divided per fragment
Predictive Power (AUC)
1.0
0.8
0.6
1%
2.5%
5%
10%
20%
Frag. 1
Frag. 2
Frag. 3
Frag. 4
Frag. 5
Frag. 6
Frag. 7
Frag. 8
Frag. 9
Frag. 10
Scramble
C
log2 catRAPID score
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
0
1000
2000
3000
4000
5000
6000
Ranked RNA - Protein interaction Index
Proportion of catRAPID score
1.0
0.8
0.6
0.4
0.2
0.0
15
30
50
70
85
100
Percentiles
Fragment 1
Fragment 2
Fragment 3
Fragment 4
Fragment 5
Fragment 6
Fragment 7
Fragment 8
Fragment 9
Fragment 10
Scramble
D
Fragment4
Fragment5
Fragment10
log2 catRAPID score (85th percentile)
5.8
5.7
5.6
5.5
5.4
5.3
5.2
PUS7-Frag.10
NOP58-Frag.10
GPKOW-Frag.10
RPS6-Frag.10
SF3A1-Frag.10
LSG1-Frag.10
GPKOW-Frag.5
FAM120A-Frag.10
PUS7-Frag.5
SMG7-Frag.10
RPL9-Frag.10
GPKOW-Frag.4
PUS7-Frag.4
0
50
100
150
200
250
300
350
Ranked RNA - Protein interaction Index

Figure 3.


A
Scale
2 kb
wuhCor1
500
1,000
1,500
21,500
22,000
22,500
23,000
23,500
24,000
24,500
25,000
25,500
26,000
26,500
27,000
27,500
28,000
28,500
29,000
29,500
NC_045512v2:
RaPID Fragments
ORF1ab
S
ORF3a
E
M
ORF6
ORF7a
ORF7b
ORF8
ORF9b
ORF9c
N
ORF10
CaCo-2 transcripts
Calu-3 transcripts
Vero E6 transcripts
1)3f2f5640-da4a-4237-952f-cd4a41e84234|315
2)e94116d8-72eb-4252-819c-ad06fd9870d4|4145
3)a142ee65-6aca-4eef-b6f0-30e884517173|916
4)fc4d6218-3c79-466d-84e4-705bee2dacef|13813
5)4167c2ea-24da-4951-a5e2-58927d55509a|143
6)ade569a0-3419-462e-b6c5-77d6f75a8ed5|263
7)5820681b-6191-45dd-959c-96a8097aafdd|2255
8)80d2b168-232a-4773-b01e-769f6dc24857|13267
9)bd49f404-9323-4537-9590-eafa3b7fb46f|3539
10)b1b33a2b-f576-4824-892e-def4cff7aa6f|86333
11)61a628be-eac5-4042-9c08-9674a992a139|1817
12)43380068-c713-4bd8-881a-d124d54688fd|227
13)de81ef19-655d-4ced-a9cc-cb8384001058|107
14)efad7b96-ac2e-4ce1-9b83-863ffdb18eac|116

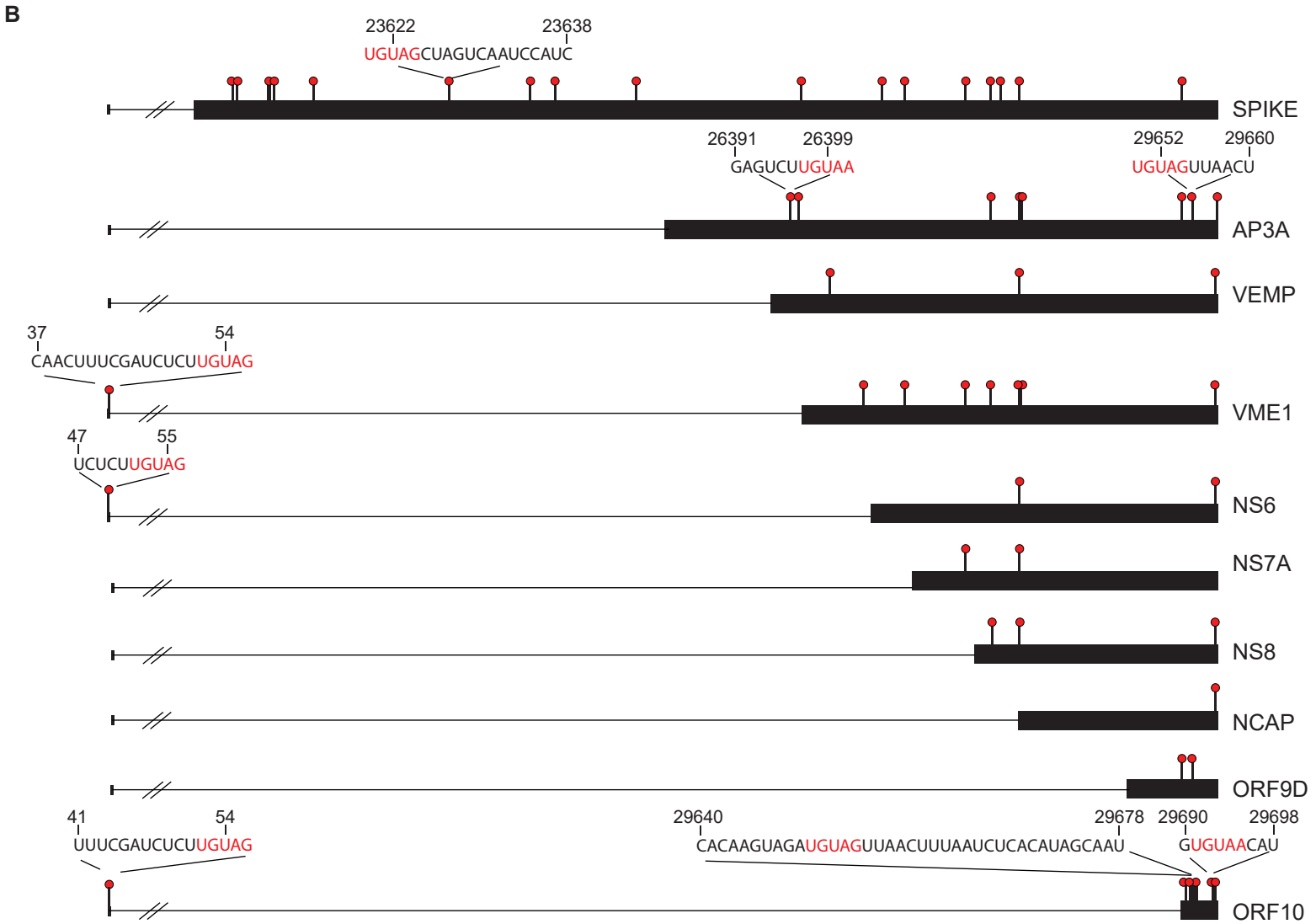
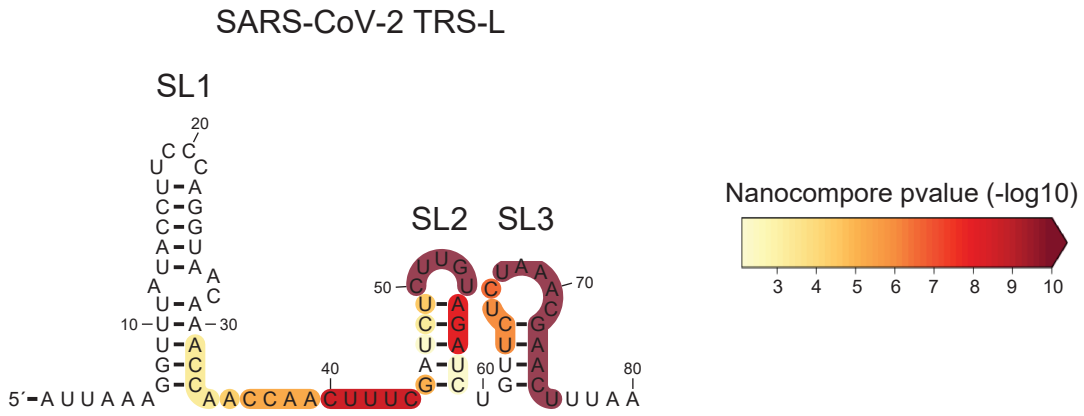

B
23622
23638
UGUAGCUAGUCAAUCCAUC
SPIKE
26391
26399
GAGUCUUGUAA
29652
29660
UGUAGUUAACU
AP3A
VEMP
37
54
CAACUUUCGAUCUCUUGUAG
VME1
47
55
UCUCUUGUAG
NS6
NS7A
NS8
NCAP
ORF9D
41
54
UUUCGAUCUCUUGUAG
29640
CACAAGUAGAUGUAGUUAACUUUAAUCUCACAUAGCAAU
29678
29690
29698
GUGUAACAU
ORF10

Figure 4.

A


SARS-CoV-2 TRS-L
SL1
SL2
SL3
5´-AUUAAA
AACCAACUUUC
Nanocompore pvalue (-log10)
3 4 5 6 7 8 9 10

B

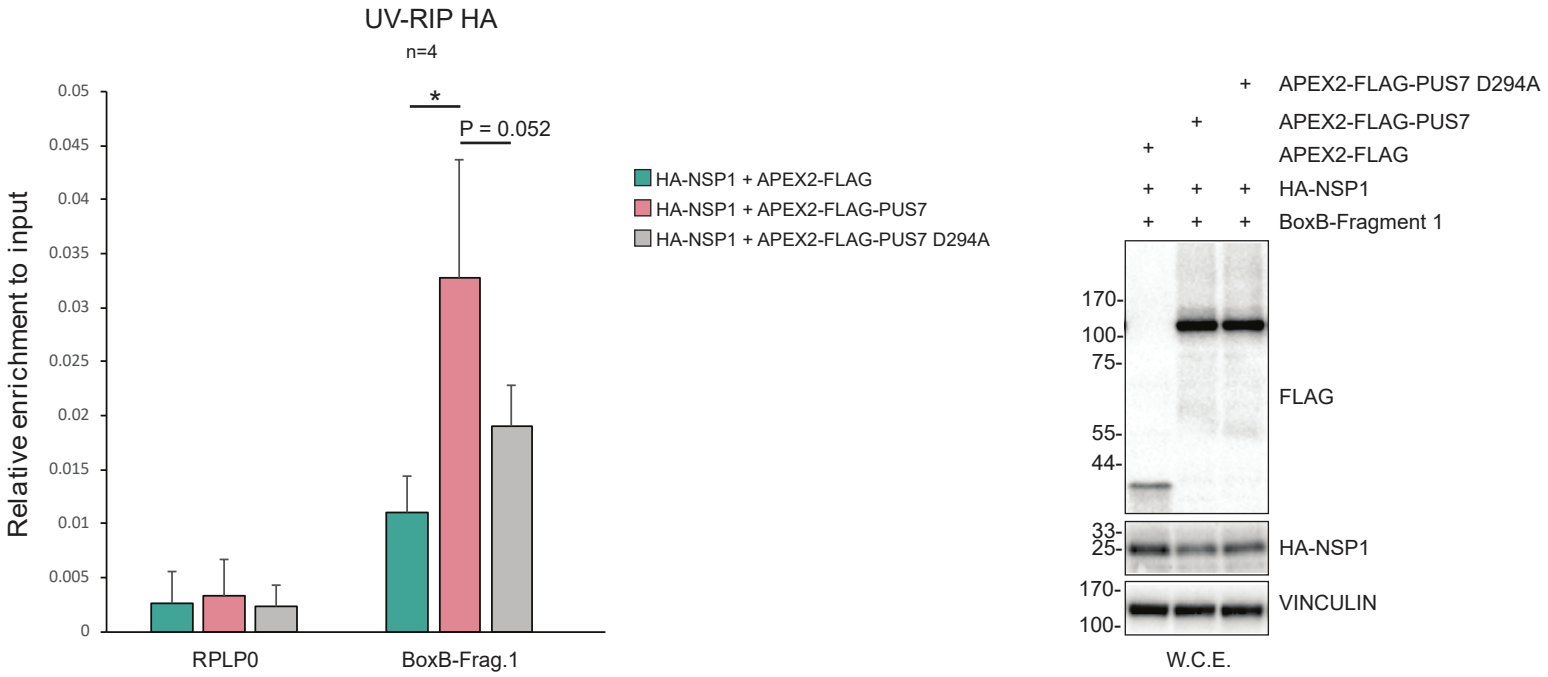

UV-RIP HA
n=4
*
P = 0.052
Relative enrichment to input
0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
RPLP0
BoxB-Frag.1
HA-NSP1 + APEX2-FLAG
HA-NSP1 + APEX2-FLAG-PUS7
HA-NSP1 + APEX2-FLAG-PUS7 D294A
APEX2-FLAG-PUS7 D294A
APEX2-FLAG-PUS7
APEX2-FLAG
HA-NSP1
BoxB-Fragment 1
170-
100-
75-
55-
44-
FLAG
33-
25-
HA-NSP1
170-
100-
VINCULIN
W.C.E.

Figure 5