COMPUTATIONAL MAPPING OF THE HUMAN-SARS-COV-2 PROTEIN-RNA INTERACTOME

A bioRχiv PREPRINT

Marc Horlacher[1], Svitlana Oleshko[1], Yue Hu[1], Mahsa Ghanbari[2], Ernesto Elorduy Vergara [1], Nikola S. Müller [1], Uwe Ohler [2], Lambert Moyon [1,*], Annalisa Marsico [1,*]

1 Computational Health Center, Helmholtz Center Munich, Germany
2 Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine, Berlin, Germany
* lambert.moyon@helmholtz-muenchen.de and annalisa.marsico@helmholtz-muenchen.de

December 22, 2021

ABSTRACT

It is well known that viruses make extensive use of the host cell's machinery, hijacking it for the purpose of viral replication and interfere with the activity of master regulatory proteins – including RNA binding proteins (RBPs). RBPs recognize and bind RNA molecules to control several steps of cellular RNA metabolism, such as splicing, transcript stability, translation and others, and recognize their targets by means of sequence or structure motifs. Host RBPs are critical factors for viral replication, especially for RNA viruses, and have been shown to influence viral RNA stability, replication and escape of host immune response. While current research efforts have been centered around identifying mechanisms of host cell-entry, the role of host RBPs in the context of SARS-CoV-2 replication remains poorly understood. Few experimental studies have started mapping the SARS-CoV-2 RNA-protein interactome in infected human cells, but they are limited in the resolution and exhaustivity of their output. On the other hand, computational approaches enable screening of large numbers of human RBPs for putative interactions with the viral RNA, and are thus crucial to prioritize candidates for further experimental investigation. Here, we investigate the role of RBPs in the context of SARS-CoV-2 by constructing a first single-nucleotide *in silico* map of human RBP / viral RNA interactions by using deep learning models trained on RNA sequences. Our framework is based on Pysster and DeepRiPe, two deep learning method which use a convolutional neural network to learn sequence-structure preferences of a specific RBP. Models were trained using eCLIP and PAR-CLIP datasets for >150 RBP generated on human cell lines and applied cross-species to predict the propensity of each RBP to bind the SARS-CoV-2 genome. After extensive validation of predicted binding sites, we generate RBP binding profiles across different SARS-CoV-2 variants and 6 other betacoronaviruses. We address the questions of (1) conservation of binding between pathogenic betacoronaviruses, (2) differential binding across viral strains and (3) gain and loss of binding events in novel mutants which can be linked to disease severity and spread in the population. In addition, we explore the specific pathways hijacked by the virus, by integrating host factors linked to these virus-binding RBPs through protein-protein interaction networks or genome wide CRISPR screening. We believe that identifying viral RBP binding sites will give valuable insights into the mechanisms of host-virus interaction, thus giving us a deeper understanding of the life cycle of SARS-CoV-2 but also opening new avenues for the development of new therapeutics.

***K*eywords** SARS-CoV-2 · RBP binding · deep learning

1 Introduction

SARS-CoV-2, causative agent of the recent COVID-19 pandemic, has and still is affecting the lives of billions of people world wide and has been a major threat to public health. Despite the large-scale vaccination effort of the last year, the number of infections remains high, as well as the number of deaths, mainly among the non-vaccinated and most vulnerable individuals. Our diffiulty to control SARS-CoV-2 infections is partly due to the rise of new more infectious viral variants, against which the full efficacy of current vaccinaes is still debated, but also to the lack of effective medications against the virus, such as host antivirals. This highlights the important of gaining a better understanding of the biology of SARS-CoV-2 and design therapeutic strategies.

SARS-CoV-2 belongs to the family of Coronavirineae and, together with SARS-CoV and MERS, it is part of the genus of betacoronaviruses, which exclusively infect mammalian species [V'kovski et al. (2021)]. SARS-CoV-2 is an enveloped positive-sense single-stranded RNA carrying a genome of 30kb, one of the largest among RNA viruses. Upon infection, it deploys an RNA molecule which uses the protein synthesis machinery of the host to express a set of viral proteins crucial for replication [Sola et al. (2015)]. The genomic RNA is translated to produce non structural proteins (nsps) from two open reading frames (ORFs), ORF1a and ORF1b, and it also contains untranslated regions (UTRs) at the 5' and 3' ends of the genomic RNA [V'kovski et al. (2021)]. A recent study revealed the complexity of the SARS-CoV-2 transcriptome, due to numerous discontinuous transcription events [Kim et al. (2020)]. In addition, negative sense RNA intermediates are generated to serve as the template for the synthesis of positive-sense genomic RNA (gRNA) and subgenomic RNAs (sgRNAs) which encode conserved structural proteins (spike protein [S], envelop protein [E], membrane protien [M] and nucleocapsid protein [N]), and several accessory proteins (3a, 6, 7a, 7b, 8 and 10) [Kim et al. (2020)]. During its life cycle, SARS-CoV-2 extensively interacts with host factors in order to facilitate cell entry, transcription of viral RNA and translation of subgenomic mRNAs, virion maturation and evasion of the host's immune response [V'kovski et al. (2021), Davey et al. (2011), Fung and Liu (2019)]. Mechanisms of virus-host interaction are multifaceted and include protein-protein interactions (PPIs), RNA-RNA interaction and binding of host proteins to viral RNAs, which can be both pro-and anti-viral. Studies on SARS-CoV-2 infected cells to date have been focused on the 'entry point' of the virus human epithelial cells, involving the interaction of the viral spike protein S with the human ACE2 receptor and subsequent fusion with the cellular membrane [Kim et al. (2020)], on characterizing the expression or modification changes in the host cell transcriptome and proteome [Stukalov et al. (2021), Gordon et al. (2020b)] and on performing genome wide CRISPR screens to identify host factors essential for the infection process [Wei et al. (2021), Wei et al. (2021)]. Mapping of protein-protein interactions (PPIs) between viral and host proteins has revealed cellular pathways important for SARS-CoV-2 infection. For instance, a recent study identified close to 300 host-virus interactions in the context of SARS-CoV-2 [Gordon et al. (2020b)]. However these studies cannot reveal how viral RNA is regulated during infection.

RNA viruses have the ability to hijack key cellular host pathways by interfering with the activity of master regulatory proteins such as RNA binding proteins (RBPs) [Hentze et al. (2018)]. These are a class of proteins able to bind RNA molecules and control several steps of cellular RNA metabolism, such as splicing, stability, export and translation initiation. RBPs are known to recognize RNA molecules by their nucleotide sequence, as well as their secondary structure, and in most cases RNA targets of an RBP share at least one common local sequence or structure preference – a so-called motif - which facilitates the recognition of the RNA by the protein. RNA viruses, specifically, contain RNA sequences and structures that play critical roles in the virus life cycle and host cell RBPs which interact with these RNA elements influence several steps of this cycle. These include: selective recognition of viral template RNA; recruitment of viral RNA to membrane compartments; synthesis of negative, positive, and subgenomic viral RNAs [Garcia-Moreno et al. (2018)] and viral RNA stability [Li and Nagy (2011b)] and decay, where they behave as essential actors in the cell's antiviral defense [Molleston and Cherry (2017)]. Examples include the Dengue virus [Manokaran et al. (2015)], the Murine Norovirus (MNV) [Vashist et al. (2012)] and others [Li and Nagy (2011a)]. A recent study shows that over 200 human RBPs display differential interaction with RNA upon Sindbis virus (SINV) infection of a human cell line. Here, RBPs stimulated by the infection seem to redistribute to viral replication factories and regulate the capacity of the virus to infect [Garcia-Moreno et al. (2018)]. In a recent proteome-wide study, 342 RBPs were identified to be annotated with gene ontology (GO) terms related to viruses, infection or immunity with a further 130 RBPs being linked to viruses in literature [Garcia-Moreno et al. (2018)]. The ability of viral RNAs to recruit essential host RBPs could explain permissiveness of certain cell types as well as its range of hosts [Li and Nagy (2011b)], which is especially relevant for zoonotic viruses such as SARS-CoV-2.

Several studies have demonstrated direct associations between human RBPs and betacoronavirus infections. For instance, targeted recruiting of RBPs has been shown in context of SARS-CoV infection, including DEAD-box helicase 1 (DDX1), which facilitates template read-through and thus replication of genomic viral RNA and heterogeneous nuclear ribonucleoprotein A1 (hnRNPA1), which might regulate viral RNA synthesis [Fung and Liu (2019), Luo et al. (2005), Wu et al. (2014)]. Further, 5'-UTR binding of zinc finger CCHC-type andRNA-binding motif 1 (ZCRB1) [Tan

et al. (2012)] and poly(A) tail binding of poly(A)-binding protein (PABP) [Spagnolo and Hogue (2000)] has been shown in infectious bronchitis virus and bovine coronavirus, respectively. As all human coronaviruses harbor 3'-and 5'-UTR regions and poly(A) tails [Liu et al. (2021a)], binding of those RBPs could also facilitate host-virus interaction with SARS-CoV-1 and SARS-CoV-2 (Fung et al. [Fung and Liu (2019)]). It is thus crucial to evaluate the potential of these RNA genomes to hijack host RBPs.

Multiple recent studies show that SARS-CoV-2 RNAs extensively interact with both pro-and anti-viral host RBPs during its life cycle [Flynn et al. (2021), Schmidt et al. (2021), Lee et al. (2021), Labeau et al. (2021)]. Using comprehensive identification of RNA-binding proteins by mass spectrometry (ChIRP-MS), Flynn et al. [Flynn et al. (2021)] identified a total of 229 vRNA-bound host factors in human Huh7.5 cells with prominent roles in protecting the host from virus-induced cell death. Schmidt et al. [Schmidt et al. (2021)] identified 104 vRNA-bound human proteins in the same cell line via RNA antisense purification and quantitative mass spectrometry (RAP-MS), with GO-terms strongly enriched in translation initiation, nonsense-mediated decay and viral transcription. The authors further confirmed the specific location of vRNA binding sites for cellular nucleic acid-binding protein (CNBP) and La-related protein 1 (LARP1) via enhanced cross-linking immunoprecipitation followed by sequencing (eCLIP-seq), which were both associated to restriction of SARS-CoV-2 replication [Schmidt et al. (2021)]. Lee at al. [Lee et al. (2021)] identified 109 vRNA-bound proteins via a modified version of the RAP-MS protocol and linked those RBPs to RNA stability control, mRNA function, and viral process. Further, the authors show 107 of those host factors are found to interact with vRNA of the seasonal betacoronavirus HCoV-OC43, suggesting that the vRNA interactome is highly conserved. Finally, Labeau et al. [Labeau et al. (2021)] used ChIRP-MS to identify 142 host proteins that bind to the SARS-CoV-2 RNA and showed, in contrast to Flynn et al. [Flynn et al. (2021)], that siRNA knockdown of most RBPs cellular expression leads to a significant reduction in viral particles, suggesting that the majority of RBPs represent pro-viral factors. Taken together, there is strong evidence that SARS-CoV-2, like other RNA viruses, relies on the presence of a large number of essential RNA-binding host factors. However, the list of SARS-CoV-2 relevant RBPs have limited overlap and the outcome really depends on the specific cell line utilized in the experiment. A large scale analysis of the propensities of different host RBPs to bind the SARS-CoV-2 genome

Over the last two decades, many experimental methods have been developed to profile RBP-RNA interactions. In vivo binding assays such as crosslinking and immunoprecipitation (IP) followed by sequencing (CLIP-seq) (Hafner et al. [Hafner et al. (2021)]), including PAR-CLIP and eCLIP protocols are arguably the most widely used as they are able to provide sets of functional elements that are directly bound by an RBP of interest [Van Nostrand et al. (2020a)]. The ENCODE Consortium currently maintains the largest collection of eCLIP experiments, spanning 223 experiments across 150 RBPs and two cell lines as of August 2021 [Van Nostrand et al. (2020b)], while PAR-CLIP datasets are collected from [Mukherjee et al. (2019)]. While CLIP-Seq experiments would also allow for precise identification of host factor interaction with viral RNAs, the high cost of profiling interactions across a large number of RBPs quicklckly become prohibitive, as dedicated pull-down and sequencing has to be performed for each RBP individually, therefore such dataset. Further, in order to keep up with the continuous emergence of novel SARS-CoV-2 variants, CLIP-Seq experiments would need to be repeated for the genome of each viral strain in order to account for (or to identify) gain-or loss-of-binding variants. A cheaper and more flexible alternative is to computationally model the binding preference of RBPs using existing CLIP-Seq datasets. To date, an abundance of computational methods to predict binding sites have been developed, many of which make use of recent advances in machine-and deep-learning [Pan et al. (2019)]. Two of those methods, Pysster [Budach and Marsico (2018)] and DeepRIPE [Ghanbari and Ohler (2020)], use convolutional neural network (CNN) which is trained on binding sites derived from CLIP-Seq experiments and learns the sequence preferences of RNA-RBP interactions.

In this study we set up to provide, to our knowledge, the first comprehensive *in-silico* analysis of SARS-CoV-2 RNA - host RBP interactions training the machine learning model pysster on about 150 RBP eCLIP in-vivo dataset, completed by deepRIPE predictions on 59 additional RBPs from PAR-CLIP experiments. We provide a high-quality RBP binding map for SARS-CoV-2, 6 other betacoronaviruses and SARS-CoV-2 variants, representing the propensities of a large number of human RBP to bind the viral RNA. We provide a selected set of human RBPs which are confidently predicted to bind SARS-CoV-2 and with putative functional roles in the infection process as reported from previous studies, as well as novel candidates with no RBP binding experimental evidence so far for SARS-CoV-2. Finally, we identify mutations and rearrangements across 10 different SARS-CoV-2 variants-of-concern, which significantly alter RBP-RNA interactions, as well as perform extensive *in-silico* single-nucleotide perturbation across the SARS-CoV-2 genome to predict SNPs that would lead to gain and/or disruption of RBP binding sites.

2 Results

We obtained binding site information from publicly available eCLIP experiments of 150 RBPs from the ENCODE [Van Nostrand et al. (2020b)] database and pre-processed them to obtain a set of high-quality protein-RNA binding sites for each RBP (Methods), which were subsequently used to generate training samples of bound and unbound sequence on human RNA transcripts (Figure 1A). For each RBP, we trained a CNN classifier to predict protein-RNA interaction on arbitrary input RNA sequences using pysster Budach and Marsico (2018), resulting in 150 pysster models (Figure 1B). Additionally, we used DeepRiPe (Ghanbari and Ohler (2020)) to train 59 models on PAR-CLIP data which were used to validate and complement pysster model predictions. Next, we perform extensive model performance evaluation and remove poorly performing models from downstream analysis (Figure 1C). For SARS-CoV-2 binding predictions, we introduce a scanning approach (Methods) to enable binding prediction at single-nucleotide resolution on RNA sequences of arbitrary length (Figure 1D). Binding predictions are additionally annotated with empirical p-values (Methods). Using single-nucleotide prediction scores and empirical p-values, we construct an in-silico binding map of human RBPs on the SARS-CoV-2 genome (Figure 1E) and score the impact of genomic variants-of-concern identified in 10 SARS-CoV-2 viral variants (Figure 1F). Additionally, we perform an *in-silico* mutagenesis of the SARS-CoV-2 genome to identify hypothetical SNPs that lead to gain or disruption of human RBP binding sites, and compare those with the characterized variants-of-concern. Further, we perform binding prediction of 6 other human coronaviruses and identify conserved binding sites as well as binding sites that are unique to SARS-CoV-2.

2.1 Robust Model Predictions in Human and Viral Sequences

Pysster models were extensively evaluated for selecting a subset of high-confidence predictors out of the 150 models. Performance evaluation (Methods) showed a robust precision and recall (PR), with a median area under the PR curve (auPR) of 0.6 (Figure 2A). As models are used for downstream scanning of the viral genome, rather than classification of standalone examples, we further evaluate the performance of the individual models by computing the correlation of the predicted positive-class probabilities with the ground-truth of ENCODE peaks on a held-out set of human transcripts (Methods). Nearly all models showed a significant positive correlation, with a mean median Spearman correlation coefficient (SCC) across transcripts of 0.149 and a maximum median SCC of 0.38 (Figure 2B). This indicates that the trained models are well-suited for the scanning task across the viral genome, as further supported by qualitative analysis on selected hold-out human transcripts (Figure 2C). In general, we observe that models which perform well with respect to the auPR tend to perform well in the context of RNA sequence scanning (Figure 2D). To ensure that downstream analyses are based on a high-quality set of binding site predictions, models with a median SCC of less than 0.1 or a auPR of less than 0.6 were discarded. A total of 63 models were thus kept for predicting on the SARS-CoV-2 genome. For DeepRIPE we relied on the results from [Ghanbari and Ohler (2020)] and retained only those models where informative sequence motifs could be learned during training, leaving a total of 33 RBP models for predicting on SARS-CoV-2 genome.

Further, to confirm that pysster models trained on CLIP-seq data from human cell lines are suitable for cross-species binding-site inference in SARS-CoV-2, we validated our approach for selected RBPs. To this end, we obtained eCLIP datasets for CNBP and LARP1 on both human and SARS-CoV-2 transcripts from Schmidt et al. [Schmidt et al. (2021)]. We preprocessed called binding sites on human transcripts in a similar fashion to the ENCODE binding sites as described in (Methods) and subsequently trained models on generated training data. We then performed prediction along the SARS-CoV-2 RNA sequence and compared the resulting prediction scores with true binding sites as well as the raw eCLIP signal (Figure 2E, 2F). Figure 2E shows a strong correlation between model predictions and eCLIP signal for the CNBP model (SCC=0.332, p-value=0 ; for LARP1: SCC=0.133, p-value=7.96e-12), and shows accumulation of high-scoring positions at location of called binding sites. Further, we observed significantly higher prediction scores for in-binding-site nucleotides versus outside-binding-site nucleotides for both RBPs (Figure 2F; t-test, p-value=0 for CNBP; p-value=2.44e-6 for LARP1). Taken together, these results strongly support the validity of our approach for cross-species in-silico prediction of RBP binding sites.

2.2 A comprehensive *In-Silico* Binding-Map of Human RBPs on SARS-CoV-2

We performed *in-silico* binding site calling by identifying consecutive significant and high-scoring positions within the SARS-CoV-2 genome with both DeepRIPE and pysster high-confidence RBP models (Methods). Based on binding site predictions, we first demonstrate that our model predictions correspond to *bona fide* RBP binding sites on the on the SARS-CoV-2 genome by performing motif analysis, second we build a first computational map of SARS-CoV-2-human RBP interactions and evaluate the enrichment of different RBPs for different viral genomic regions. Finally, we narrow down the number of human RBPs predicted to bind the SARS-CoV-2 genome to a short list of potential functional candidates.

To illustrate how the binding site prediction procedure is implemented, single-nucleotide resolution prediction scores from the RBFOX2 pysster model and the MNBL1 DeepRIPe model are shown as example across the entire SARS-CoV-2 genome (Figure 3A). High-scoring significant positions on the binding profiles, which were found within 10 nt from each other (see Methods) were further merged into larger binding sites. We used integrated gradients to interpret the predictions (see Methods), in order to assess what viral sequence features drive the binding site predictions, and whether they correspond to the binding site preferences of those proteins known from the literature. In detail, for both RBFOX2 and MBNL1 we computed attribution maps pinpointing the individual nucleotides that were most important for the classification of each extended binding site sequence for those RBPs (Figure 3A). For both RBFOX2 and MBNL1 the motifs extracted from the attribution maps in correspondence of the predicted binding sites matched the known motifs for these proteins. For example, we observe the well-known consensus sequence (U)GCAUG recognized by the splicing factor RBFOX2 in the corresponding attribution maps (Figure 3A). Similarly, the attribution maps computed at each binding site of the splicing factor MBNL1 all harbour occurrences of the YGCY motif.

We could train predictive models with both pysster and DeepRIPE for 20 proteins where both eCLIP and PAR-CLIP datasets were available. We filtered 10 out of these 20 RBPs, whose models did not fulfill the quality criteria implemented above for either one or the other method. In order to evaluate the robustness of our viral binding site predictions across different datasets and across different machine learning model architectures, we compared pysster and DeepRIPE predictions on the SARS-CoV-2 genome for the remaining 10 proteins, namely TARDBP, CSTF2, IGF2BP1, PUM2, CSTF2T, QKI, IGF2BP2, IGF2BP3, CPSF6 and FXR2. We observed significant signal correlations at the level of single nucleotide prediction score for 8 out of the 10 RBPs, ranging from a maximum of 0.64, for the TARDBP protein (Figure 3B) to a minimum of 0.15 for CPSF6. In addition, 9 out of 10 of the RBPs also exhibited a significant correlation at the level of single-nucleotide binding site prediction. We noticed in general a higher consensus of binding site predictions for those with harbouring a well-defined RNA sequence motif, such as QKI, TARDBP, PUM2, CSTF2T and CSTF2. As proof of concept, we show that attribution maps computed in correspondence of consensus binding sites between pysster and DeepRIPE recapitulate well the known motifs of QKI and TARDBP RBPs (Figure 3c), highlighting the robustness of model predictions for RBP models of good quality. In addition, robust RBP binding site predictions where the two methods agree are valuable candidates for future experimental validation.

Confident from our attribution map analysis that our predictions might reflect true propensities of selected viral regions to be bound by human RBPs, we set out to build the full *in-silico* map of human RBPs, for which a high confidence prediction model could be trained, on SARS-CoV-2. We clustered pysster and DeepRIPE predicted binding sites separately (Figure 3B, Figure S2) on the basis of the dominant SARS-CoV-2 bound genomic region (Figure 3C) (see Methods). Both the deepRIPE and pysster-based binding heatmaps show clusters of proteins exhibiting extensive binding across the whole SARS-CoV-2 genome, independently of the genomic element, while other RBPs localize or are enriched in specific genomic elements (Figure 3C, lower panel). For example, RBPs from cluster 1 from the pysster heatmap do not show a strong preferential localization pattern and include factors involved in diverse processes of the RNA metabolism, including RNA splicing, stability, localization, export and others. Cluster 2 contains RBPs with preferential binding for the viral genome 5'UTR, mainly splicing regulators such as RBFOX2, PRPF8 and NONO, as well as the DDX3X helicase, which was recently identified as host target against SARS-CoV-2 infection [Ciccosanti et al. (2021)], and TARDBP, a protein that localizes to P-bodies and stress granules, whose binding to SARS-CoV-2 5' UTR was also identified in a recent study [Mukherjee and Goswami (2020)]. This protein could be implicated in facilitating translation of viral proteins. Notably, the cluster 7 contains RBPs displaying binding preference to viral 5' UTR, Orf7b and Orf8. Some of these RBPs are part of the splicesome complex, including GPKOW, SF3A3, SF3B4, AQR and A2AF2. RBPs which preferentially bind the viral 3' UTR include the FTO protein, an RNA demethylase factor, known to regulate viral infection and HIV-1 protein expression [Tirumuru et al. (2016)], as well as viral infection regulators SND1 and SRSF7, known to be involved in mRNA 3' end processing, and RBPs FUBP3 and SAFB2. Finally, cluster 6 to 10 from the pysster heatmap mainly include proteins with extensive binding across the whole SARS-CoV-2 genome including the genomic elements encoding for subgenomic RNAs for ptoein E (cluster 6), spike protein S (cluster 8) and protein N (cluster 9).

2.3 A high-confidence catalog of predicted human RBPs - SARS-CoV-2 RNA interactions and their functional implications

A total of 92 RBPs (59 from pysster models, 33 from DeepRipe model) were found to present at least one predicted binding site in the SARS-CoV-2 genome We compiled a list of candidate RBPs with putative important functional roles in SARS-CoV-2 infection (Table 1). In detail, we inspected the scientific literature for either evidence of involvement of those host RBPs in SARS-CoV-2 host-pathogen PPI networks which regulate host immune response [Gordon et al. (2020b), Stukalov et al. (2021)] or evidence of being identified as essential for SARS-CoV-2 infection from genome-wide CRISPR screens [Kim et al. (2020), Wei et al. (2021)]. To strengthen the evidence about putative candidate RBPs, we also included the results from two studies showing experimentally direct evidence of some host

Table 1: Overlaps between Pysster models, DeepRIPE models, and proteins from external sources (established proteins)

	Pysster			DeepRIPE		Overlaps with proteins from external sources										Pathway
RBP	**auROC**	**auPR**	**medPIP**	**au-ROC**	**AP**	**Kim et al. (2021)**	**Gordon et al. (2020a)**	**Wei et al. (2021)**	**Schneider et al. (2021)**	**Schmidt et al. (2021)**	**Vandelli et al. (2020)**	**Bojkova et al. (2020)**	**Lee et al. (2021)**	**Stukalov et al. (2021)**	**Sun et al. (2021)**	**SIGNOR**
AATF	0.92	0.66	0.1				X									16-17
BCLAF1	0.92	0.6	0.36						X[4]				X[15]			1, 6-7, 9-10, 14, 16-17, 20, 22-24, 31, 33-34, 37
CAPRIN1				0.76	0.22				X[234]				X[1416]		X[18]	1, 8, 11, 17
CPSF6	0.89	0.61	0.18	0.79	0.26											
CPSF7				0.79	0.54						X[6]		X[1516]		X[18]	
CSTF2	0.93	0.81	0.14	0.82	0.3										X[1718]	
DDX3X	0.96	0.78	0.32						X[23]	X		X[10]	X[15]			
ELAVL1				0.9	0.73			X			X[67]	X[8]	X[1516]			
EWSR1	0.93	0.62	0.22	0.85	0.2			X[1]				X[9]	X[1216]		X[18]	
EXOSC5	0.9	0.64	0.09				X						X[14]			11
FAM120A	0.92	0.62	0.24									X[8]	X[1516]		X[1718]	
FIP1L1				0.8	0.3				X[4]				X[1516]			
FMR1	0.94	0.67	0.18						X[4]	X		X[10]				
G3BP1	0.93	0.64	0.31		X	X			X[234]			X[8]	X[111516]			1, 3, 5, 9-11, 13-21, 23-24, 26-27, 29-32, 34-35, 38
GRSF1	0.93	0.71	0.18					X	X[4]							
HNRNPA1	0.94	0.74	0.11						X[23]				X[15]		X[1718]	
HNRNPU	0.91	0.61	0.1				X		X[4]				X[12141516]		X[18]	1, 7
IGF2BP1	0.91	0.66	0.12	0.83	0.19				X[23]	X	X[6]		X[1416]		X[17]	
IGF2BP2	0.91	0.65	0.13	0.84	0.29				X[2]	X		X[8]			X[17]	
IGF2BP3	0.88	0.56	0.08	0.84	0.42					X		X[8]	X[14]		X[1718]	
ILF3	0.93	0.74	0.15						X[4]		X[7]	X[9]	X[1516]		X[17]	
KHDRBS1	0.97	0.86	0.18									X[9]	X[1516]		X[1718]	
KHSRP	0.9	0.65	0.15							X			X[1516]		X[1718]	1-19, 21-38
MATR3	0.94	0.7	0.23					X	X[4]			X[8]	X[1516]			
MOV10				0.83	0.41		X		X[23]			X[8]	X[141516]		X[1718]	1-2, 7-8, 17, 20-21, 24, 27, 29, 37
NIP7	0.92	0.69	0.15								X[7]					
NOLC1	0.93	0.76	0.1						X[4]	X	X[7]		X[1516]			
NONO	0.92	0.6	0.15	0.93	0.38		X				X[7]	X[89]	X[1516]		X[17]	
NOP56				0.92	0.69								X[1516]		X[1718]	
PTBP1	0.94	0.8	0.37								X[67]	X[8]	X[1516]		X[1718]	
QKI	0.97	0.87	0.31	0.97	0.64										X[1718]	
RBM22	0.91	0.72	0.2												X[1718]	
RPS3	0.94	0.63	0.28						X[234]			X[10]	X[12141516]			
SAFB2	0.93	0.7	0.11									X[9]	X[15]		X[18]	
SERBP1	0.94	0.74	0.07						X[234]		X[7]	X[8]	X[1516]			
SND1	0.95	0.75	0.21						X[234]		X[5]	X[8]	X[15]		X[1718]	
SRSF1	0.94	0.7	0.28								X[67]		X[1516]			
SRSF7	0.92	0.69	0.17				X		X[4]				X[16]			
SUGP2	0.9	0.63	0.15						X[4]							
TARDBP	0.98	0.92	0.28	0.95	0.73			X							X[18]	1-12, 14-22, 24-25, 27-29, 31-38
TIAL1	0.95	0.8	0.13						X[4]			X[10]			X[1718]	
U2AF2	0.95	0.77	0.15								X[67]				X[1718]	
XPO5	0.91	0.65	0.1	0.84	0.29		X	X[1]	X[4]				X[11]			
YBX3	0.92	0.71	0.1						X[2]			X[8]				
ZNF800	0.93	0.62	0.25						X[4]	X			X[15]			
ZRANB2	0.9	0.64	0.1						X[4]							

RBPs [Schmidt et al. (2021)] of being part of the SARS-CoV-2 bound interactome [Schmidt et al. (2021)] or predicted to bind SARS-CoV-2 based on the PRISMNet deep learning model, which integrate RNA structural features from experimental icSHAPE data and in vivo RBP-RNA bindings to predict high-confidence binding sites [Sun et al. (2021)]]. In addition, we integrated evidence of direct involvement of those RBPs in SARS-CoV-2 infection, as reported in the SIGNOR database, a highly curated resource of pathways and manually curated genes involved in SARS-CoV-2 [Licata et al. (2020)]. A total of 44 out of the 92 RBPs had evidence of experimentally determined or predicting viral RNA binding or had been reported by at least two other studies listed in Table 1, including the SIGNOR database. On the other hand, 25 out of the 92 RBPs have not been reported to bind SARS-CoV-2, neither to be associated to SARS-CoV-2 pathways in SIGNOR, and can be considered novel promising candidates for future investigation (Table 2).

Table 2: Overlaps between Pysster models, DeepRIPE models, and proteins from external sources (novel proteins)

	Pysster			DeepRIPE		Overlaps with proteins from external sources										Pathway
RBP	**auROC**	**auPR**	**medPIP**	**au-ROC**	**AP**	**Kim et al. (2021)**	**Gordon et al. (2020a)**	**Wei et al. (2021)**	**Schneider et al. (2021)**	**Schmidt et al. (2021)**	**Vandelli et al. (2020)**	**Bojkova et al. (2020)**	**Lee et al. (2021)**	**Stukalov et al. (2021)**	**Sun et al. (2021)**	**SIGNOR**
AGGF1	0.91	0.71	0.17							X			X[15]			
CPSF1				0.77	0.23			X					X[1516]			
CPSF3				0.8	0.12		X	X		X						
DKC1	0.96	0.89	0.21							X			X[1516]			
DND1				0.82	0.46											
ELAVL2				0.93	0.61											
ELAVL3				0.94	0.72											
ELAVL4				0.93	0.58											
FTO	0.92	0.63	0.27													
FXR1	0.92	0.6	0.26	0.86	0.26	X						X[8]				
FXR2	0.94	0.67	0.23	0.8	0.18	X						X[10]				
GPKOW	0.92	0.71	0.16					X								
LARP7	0.87	0.67	0.07				X			X			X[1516]			
MBNL1				0.98	0.94								X[16]			
NOP58				0.93	0.68				X[4]							
PABPN1	0.89	0.6	0.17													
PRPF8	0.95	0.74	0.33										X[14]			
PUM2	0.95	0.8	0.16	0.95	0.72							X[10]	X[1516]			
RBFOX2	0.96	0.8	0.24										X[1516]			
SF3A3	0.96	0.84	0.23													
SF3B4	0.98	0.88	0.26			X		X								
SRSF9	0.92	0.65	0.18										X[141516]			
TRA2A	0.96	0.8	0.29										X[141516]			

[1] also included in the PPI network

[2] SARS-CoV-2 RNA interacting proteins

[3] proteins included in the PPI network (network based on STRING v.11 interactions between human proteins in the expanded SARS-CoV-2 RNA interactome)

[4] differentially expressed proteins (SARS-CoV-2 infected and uninfected Huh7 cells)

[5] proteins that were reduced during SARS-CoV-2 infection

[6] proteins that increased during SARS-CoV-2 infection

[7] additional potential inhibitors of SARS-CoV-2 replication

[8] statistically significant interactors enriched in both probe I and probe II experiment

[9] statistically significant interactors enriched in only probe I experiment

[10] statistically significant interactors enriched in only probe II experiment

[11] proteins included in virus-host PPI network of SARS-CoV-2 in A549 cells

[12] transcripts that significantly change upon SARS-CoV-2 infection

[13] proteins that significantly change upon SARS-CoV-2 infection

[14] ubiquitination site significantly changes upon SARS-CoV-2 infection

[15] phosphorylation site significantly changes upon SARS-CoV-2 infection

[16] included in results of enrichment analysis

[17] RBPs predicted to bind the 5'UTR of SARS-CoV-2

[18] RBPs predicted to bind the 3'UTR of SARS-CoV-2

Table 3: Pathways from SIGNOR COVID-19 Browser

Id	Pathway	Id	Pathway	Id	Pathway
1	Innate response to dsRNA	14	MAPK activation	27	NF-KB Canonical
2	Acute Myeloid Leukemia	15	Phagosome (KEGG)	28	Non-small-cell lung cancer
3	Fibrosis	16	Protein processing in ER (KEGG)	29	Pancreatic ductal adenocarcinoma
4	RNA degradation (KEGG)	17	Autophagy	30	PI3K/AKT
5	AMPK Signaling	18	Colorectal Carcinoma	31	TNF alpha
6	Glioblastoma Multiforme	19	Death Receptor Signaling	32	T cell activation
7	IL1 Receptor	20	EGFR	33	TGFbeta Signaling
8	Stress granules	21	FSGS	34	Toll like receptors
9	Virus entry	22	Sonic Hedgehog	35	WNT Signaling and Myogenesis
10	Apoptosis	23	Insulin Receptor	36	WNT:WNT Signaling
11	ER stress	24	Inhibition of Apoptosis	37	TCA cycle (KEGG)
12	Glycolysis-Gluconeogenesis (KEGG)	25	Malignant Melanoma	38	VEGF
13	Lysosome (KEGG)	26	MTOR Signaling		

2.4 Variants of Concern Show Gain-and Loss-of-Binding Events

Multiple waves of SARS-CoV-2 infections have spread across the globe, some of which linked to the apparition of viral variants as compared to the genome of the reference SARS-CoV-2 virus. These variants have been experimentally characterized as more efficient than other lineages, explaining in part how they became the most successful at spreading locally or globally. These viral variants can be categorized from their genomic sequence, notably by the identification of specific genetic variants, systematically found across thousands of sequenced samples from patients infected by the viral variant in a specific location and time period. These sequence mutations have been investigated for their potential to impact the sequence at the protein level for genes of particular importance in the entry or the replication cycle of the virus, in particular for those located within the gene of the spike protein. This lead to the definition of sets of variants of concerns, mutations of the sequence defined by the World Health Organization as genetic variants with "evidence for increased transmissibility, virulence, and/or decreased diagnostic, therapeutic, or vaccine efficacy.", each set associated to a specific viral variant. Although such mutation may be selected from their protein coding impact conferring a favorable change for the virus, it is also possible that these mutations affect the binding of host RBPs, thus modulating their hijacking of these host factors to their benefit.

To investigate this, we compiled a total of 229 variants-of-concern from 10 viral variants, such as alpha and delta variants, initially gathered from the UCSC genome browser for SARS-CoV-2 [Fernandes et al. (2020)]. Of these, 31 mutations were shared by more than one viral variant, with the mutations A23403G, C3037T, C14408T consistently found across all viral variant. While accounting for shared mutations between viral variant results in a set of 150 unique mutations, we investigated each of the 229 mutation in context of the mutation markup of the respective viral variant, as duplicate mutations can interact with other (unique) mutations within the viral variant, thus affecting RBP binding in different ways. To do so, we spiked each set of variants of concern in the sequence of the reference genome of SARS-CoV-2, and calculated delta scores for each mutation, which were further used to identify events of gain or loss of binding (Methods). We identified 141 events which resulted in gain-or loss-of-binding (GOB or LOB) across 82 unique mutations and 45 RBPs, as illustrated in Figure 4A. We observe notably a higher number of GOB (82) vs. LOB (59) events. Among those, the highest number of events was observed for RBPs YBX3 (9 events from 9 distinct mutations, of which 5 are loss of binding) and IGF2BP2 (9 events from 9 distinct mutations, of which 3 are loss of binding). Interestingly, we observed that some groups of RBPs exhibit similar mutation-impact profiles, such as YBX3, RPS3, SND1, and SRSF7 (belonging to the previously identified cluster 10, Figure 3C). This highlights that RBPs with similar sequence preference profiles may be sensitive to the same set of mutations. Finally, we note that some mutations are not only shared between some viral variant, but also co-occur with other variants-of-concern in their proximity. Such groups of mutations appear to affect in similar fashion the binding sites of RBPs. For instance, we see that the G3BP1 RNA binding protein presents a potential gain of binding event in the 5'UTR of the viral variants delta, epsilon, eta, gamma, iota, kappa, lambda and mu (Figure 4A). This gain of binding is detected from multiple mutations located nearby (G210T; C241T; T733C), and two of these are shared across viral variants (G210T: delta, kappa; C241T: delta, epsilon, gamma, iota, lambda, mu).

To further filter the binding sites impact events, we applied a threshold on the absolute delta score distribution, so as to keep at most 20% of the total of mutations - RBP pairs. Doing so reduced the ensemble from 141 to 28 selected events (19.8%), representing binding sites affected by variants of concerns that change the prediction score value of more than 0.28 (absolute value). These events concern 15 RBPs and 22 unique mutations. Figure 4B shows that these events are concentrated within the gene of the spike protein, as well as the gene of the N protein, while Figure 4C highlights which RBPs are affected by those mutations. This accumulation of events per viral component does not

reflect the relative proportion of these viral components over the full size of the genome. Among RBPs with noticeable numbers of binding impact, we find ZRANB2 for the eta viral variant, as well as SRSF7 in the delta and kappa viral variants.

To display the full range of delta scores, figure 4D and 4E present the mutational profile of two lineages with the highest-spreading efficiency: the alpha and delta viral variants. For each viral variant, the top 10 mutational events are highlighted, according to their absolute delta score values. In both viral variants, half of these top events correspond to actual gain or loss of binding (black rectangles) while the other half do not change the threshold-passing status (score or p-value) from REF to ALT (purple rectangles). This second set represents nonetheless mutations with very strong impacts on the potential binding of the considered RBPs. We further illustrate some of these binding impact events through the computation of attribution maps, as presented in the Figure 3. We compare the logo plots between reference sequence and mutated sequence to both confirm the detection of a known sequence motif (as reported in the mCrossBase database [**?**]), and evaluate whether these can be used to characterize a mutation impact for gain and loss of binding. For instance in the alpha viral variant, a C>T mutation at position 27,972 leads to the predicted gain of a PUM2 binding site. Its detection by the model translates into a detected logo TGTACTTA that is similar to a known binding motif for this RBP (TGTATATA), where the two unfavourable nucleotides "CT" compensated by the 6 other nucleotides, including the T introduced by the mutation. The same observation can be made for the loss of binding of QKI from the A>T mutation at position 23,063, where the attribution map confirms the loss of a known motif. We also highlight a set of mutations occurring jointly from position 28,271 to position 28,282, predicted to affect the binding of FTO. Although only one of them is identified in the selected top 10, we see they all have the same direction in impact (loss of binding). The attribution map shows a greater amount of negatively scored nucleotides in the interval for the mutated sequence, as compared to the reference sequence, although we could not recover here a known motif for FTO. We note here that attribution maps may not always enable the detection of a small, clear motif, as the models consider a large window, and may be able to detect higher order sequence preferences.

Considering the delta variant, a loss of binding for the RBP SRSF7 in the 5' UTR can be explored through the attribution map, where two variants of concern (G210T, C241T) are predicted to be loss-of-binding within a 31 nucleotides interval. This same RBP presents two gain of binding events in the gene of the S protein (C>G at position 23,604), and in the ORF7a (T>C 27,638). A very strong loss-of-binding is predicted for PUM2, within the gene of the protein S (T>G at position 22,917). Again we can recover the motif at this position, which is lost within the mutated sequence. We also investigate a loss of binding for the RBP HNRNPC, occurring at position 25,469 from a C to T transversion in the ORF3a. The mutation introduces a new T nucleotide which contribute to a T-rich site favorable for HNRNPC binding.

2.5 Probing of SARS-CoV-2 via In Silico Mutations

With the ever quicker emergence of new SARS-CoV-2 variants, a fast assessment of their virological impact is of key importance. Here, we perform in-silico perturbation of the SARS-CoV-2 genome followed by subsequent variant RBP-binding impact scoring to generate a set of hypothetical mutations which might interfere with human RBP binding. To this end, we systematically scan the SARS-CoV-2 genome by generating the set of all possible point mutations and select the mutation with the highest impact at each genomic position (3.12). Analogously to the strain mutation analysis, we then select a subset of mutation that result in gain-or loss-of-binding. These mutations are depicted in Figure 5a. While in general binding of RBPs is highly volatile upon introduction of point-mutation, with the most impacted RBP (AQR) showing a total of 1628 change-of-binding events, some RBPs (e.g. FTO and XRN2) show a surprising robustness against genomic perturbation.

We investigated whether observed point-mutation of the 10 variants-of-concern analyzed in section 2.4 exhibit an increased or decreased effect on RBP binding compared to the set of all possible point-mutations on SARS-CoV-2. We find that both GOB and LOB events in observed point-mutations have a significantly lower absolute delta score ($p < 2.2e^{-16}$ and $p < 2.918e^{-07}$ respectively; Wilcoxon test). We hypothesize that since the majority of RBP-binding impacting variants are likely to be deleterious for the virus, the set of observed variants is enriched in variants that don't, or only marginally, affect RBP binding.

2.6 RBP-Binding across Human coronaviruses

We investigated to which extent binding sites of human RBPs are conserved across related human coronaviruses. For this purpose, we obtained genomes and genomic annotations of 6 SARS-CoV-2-related human coronaviruses, namely SARS-CoV-1, MERS, HCoV-OC43, HCoV-NL63, HCoV-HKU1, HCoV-229E (Methods). Binding sites were identified across each viral genome using all 63 high-confidence Pysster models, as performed for SARS-CoV-2 (Methods). Figure 6A shows the general binding propensity of RBPs across viral genomes of the 7 coronaviruses. In general,

RBP binding is conserved across coronaviruses, with pathogenic viruses (SARS-CoV-1, SARS-CoV-2 and MERS) showing a highly similar binding pattern. Further, a total of 62/63 RBPs (except FKBP4) were predicted to harbor a binding site in at least one virus, while the total number of binding RBPs vary slightly across different viruses. We observe a greater variability of RBP binding with shared genomic regions across coronaviruses, for instance in the 5' and 3' UTRs. Viral UTRs are known to play an important role in both pro-and anti-viral responses and recent evidence suggests that evolution of the 3' UTR is contributing to increased viral diversity [Farkas et al. (2021)]. Indeed, the 3' UTR of SARS-CoV-2 shows a severe truncation when compared to SARS-CoV-1 and MERS. Given that viral UTRs are not under selecting pressure with respect to a translated protein, they might be more prone to acquire mutations that modulate regulation through host RBPs. Figure 6B and 6C show RBP binding respectively to the 3' and 5' UTRs of all selected coronaviruses. While the severe pathogenic viruses show conserved binding on the 5' UTR and cluster strictly, a depletion of RBP binding sites is observed in the 3' UTR of SARS-CoV-2 when compared to SARS-CoV-1 and MERS.

To investigate gain-and loss-of-binding in viral UTRs across severe pathogenic human coronaviruses, we performed multiple sequence alignments of viral 3' and 5' UTRs and compared the predicted Pysster score across the three viruses (Methods).

2.6.1 Loss of FXR2-Binding in SARS-CoV-2 3' UTR

Figure 6E shows 3' UTR binding of FXR2, a paralog of FMRP (fragile X mental retardation protein). FMRP was shown to broadly bind along the entirety the 3' UTR of the Zika virus (ZIKV) [Soto-Acosta et al. (2018)]. Interestingly, our model predicted extensive binding of FXR2 along the 3' UTR of SARS-CoV-1 and MERS, while SARS-CoV-2 showed a complete lack of predicted FXR2 binding sites, owing to its significantly shorter 3' UTR. However, While FMRP was suggested to act as a ZIKV restriction factor by blocking viral RNA translation, a significantly reduced ZIKV infection was observed upon knockdown of FXR2 [Soto-Acosta et al. (2018)].

2.6.2 Druggable FTO Binds to the 3' UTR of SARS-CoV-1, MERS and SARS-CoV-2

Recently, the m6A methylation machinery was identified to play a key role in the context of SARS-CoV-2, with studies reporting widespread alterations of the viral and host RNA m6A methylome following infection [Xu et al. (2021)]. Further, altered expression levels of methyltransferase-like 3 (METTL3) and fat mass and obesity-associated protein (FTO) linked to viral replication [Zhang et al. (2020)]. FTO is a demethylase (eraser) enzyme with enriched binding in the 3' UTR of mRNAs in mammals [Meyer et al. (2012)]. FTO has previously been suggested as a potential drug target against COVID-19 [Zannella et al. (2021)], as targeted knockdown has been shown to significantly decrease SARS-CoV-2 infection [Zhang et al. (2020), Zannella et al. (2021), Burgess et al. (2021)]. We therefore investigated predicted binding of FTO to the 3' UTR of SARS-CoV-2 and related viruses. Indeed, we observed that SARS-CoV-1, SARS-CoV-2 and MERS, as well as the less pathogenic viruses HCoV-HKU1 and HCoV-OC43 harbor at least one FTO binding site in their 3' UTR (Figure 6B). Further, Figure 6D shows that while SARS-CoV-1 and MERS harbor multiple shared FTO binding sites along their 5' UTR, SARS-CoV-2 only harbors one FTO binding site at the 3' end of its 5' UTR which is exclusively shared with SARS-CoV-1.

2.6.3 Newly Acquired TARDBP Binding in the SARS-Cov-2 5' UTR

We next focus on TARDBP (also known as TDP-43) (Figure 5F), which was predicted to bind the 5' UTR of a SARS-CoV-2 mutant in a recent study [Mukherjee and Goswami (2020)]. In eukaryotes, TARDBP binds to the 5' and 3' ends of pre-mRNA transcripts and plays an important role in mRNA splicing, translation and transportation [Kuo et al. (2014)]. TARDBP has further been implicated to play a role in viral replication and pathogenesis in the context of coxsackievirus B3 infection. In contrast to the findings of Mukherjee et al. [Mukherjee and Goswami (2020)], our model identified a TARDBP binding site at the genomic range of 89-98 in the wild-type reference of SARS-CoV-2. Interestingly, in addition to observing a lack of predicted binding signal of TARDBP on the 5' UTR of SARS-CoV-1 and MERS (Figure 5F), we found a complete lack of TARDBP binding to any of the 6 other investigated coronaviruses. This suggests that 5' UTR TARDBP binding potential is newly acquired in SARS-CoV-2 and may affect its virulence.

3 Material and Methods

3.1 Data and Preprocessing

Enhanced CLIP (eCLIP) datasets were obtained from the ENCODE project database, which comprises 223 eCLIP experiments of 150 RBPs across two cell lines, HepG2 and K562. For RBPs with experiments in both cell lines, we selected only data from the HepG2 cell line for downstream analysis, as previous studies demonstrated that those are of higher quality, compared to K562. Thus, a total of 150 eCLIP datasets were processed for downstream model training and evaluation. Peaks of each RBP were taken directly from ENCODE and preprocessing was performed as follows. For each of the two replicates of a given eCLIP experiment, peaks were first intersected with protein-coding transcript locations obtained from the GENCODE database (Release 35) and only those peaks overlapping with a mRNA transcript were retained. Next, the 5' end of each peak was defined as the cross-linked site, as it usually corresponds to the highest accumulation of Reverse Transcription truncation events. A 400bp window was then centered around the cross-linked site for each peak, defining the input window of the downstream model. Input windows of both replicates were intersected reciprocally with a required overlap fraction of 0.75, ensuring that only those peaks which are present in both replicates are considered for the downstream training set construction. The two sets of input windows (corresponding to each replicate) were then merged and at most the top 50,000 peaks with a read-start count FC of 2.0 above the control (SMInput) experiment were selected for each RBP.

PAR-CLIP datasets used for training Deep-Ripe models were collected from different publications and processed with the same pipeline as described in [Mukherjee and Goswami (2020)].

3.2 Training Set Construction

We describe here the training set construction specifically for Pysster models. Please refer to [Ghanbari and Ohler (2020)] for Deep-Ripe models. For each RBP measured with eCLIP, a classification dataset of bound (positive) and unbound (negative) RNA sequences was constructed. Positive samples were directly obtained by taking corresponding reference sequences of the expanded peak regions from the previous step, while two distinct sets of negative samples were generated. First, 400bp long regions which did not overlap with binding sites of the given RBP were sampled from transcripts harboring at least one binding site. The later constraint requires that the transcript is expressed in the experimental cell type and thus ensures that the set of sampled sites does not contain regions that would be RBP-binding under different experimental conditions. The second set of negative samples was generated by sampling binding sites of other RBPs. This ensures that any CLIP-seq biases (such as uradin bias during UV-C cross-linking [38, 39]) are present in both positive and negative samples. Ensuring that CLIP-seq biases are present in both positive and negative samples prevents the model from performing a biases-based sample discrimination during the training. The outlined procedure yields a three-class training set, where class 1 corresponds to positive samples and class 2 and 3 correspond to negative samples. Samples of class 2 and 3 were samples at a 3:1 ratio with respect to class 1 in order to enlarge the training dataset and to prior the models towards unbound regions, which represent the majority among regions in in vivo RNA sequences.

3.3 Pysster Model

The pysster Python library [Budach and Marsico (2018)] was used for implementation of the model which consists of three 1D convolutional layers, each with 150 filters of size 18, followed by a single fully connected layer with 100 units. For feature-vector down-sampling, a maximum pooling layer is added after every convolutional layer. Lastly, a fully connected layer with 3 units, one for each of the three output classes, is added. Dropout [Srivastava et al. (2014)] was applied to each except for the output layer. The model was trained via the Adam optimizer [Kingma and Ba (2014)] with a batch size of 512 and a learning rate of 0.001. For each RBP, we trained for at most 500 epochs and stopped training in case the validation loss did not improve within the last 10 epochs.

3.4 Binary Classification Threshold

As pysster models are trained as a 3-class classification problem with class imbalance, we re-calibrate the model for the binary classification task by introducing a binary decision threshold t_m on the predicted class-1 probability scores. For each model m, t_m is defined as the threshold which maximizes the F1 performance of the model with respect to bound / unbound classification obtained by pooling class 2 and 3 samples into a common 'unbound' class.

3.5 Single-Nucleotide Resolution Prediction

The class-1 prediction score of the pysster models corresponds to the probability that input RNA sequence is bound by the RBP of interest. By design, this score is assigned to the entire 400bp input sequence, although RBP binding sites are much more local, usually spanning only a few nucleotides [Dominguez et al. (2018)]. We thus need a different approach to obtain single-nucleotide binding site probabilities along an RNA sequence. To this end, we employ a one-step sliding-window approach to scan over a given RNA sequence, where the predicted class-1 probability score is assigned to the center nucleotide of the input window. This approach is valid since by training set design, the binding site is expected at the center of the input sequence during training. In order to obtain predictions over the entire RNA sequence, the 5' and 3' ends are 0-padded.

3.6 Performance Evaluation and Model Selection

3.6.1 Precision-Recall and F1 Performance

We used the pysster framework to randomly split the training set generated in 3.2 (for each RBP) into train, validation and test sets at a ratio of 70:15:15 before training, respectively. As the validation loss was monitored for the purpose of early-stopping, the precision-recall (PR) and F1-score performance of the models was evaluated on the test set. Models with an area under the PR curve (auPR) of less than 0.6 were deemed poor quality and thus excluded from the downstream analysis.

3.6.2 Performance in Practice

Training datasets are sampled at a fixed positive-negative ratio which is hardly reflected by the ratio of bound and unbound sites of RNA transcripts found in vivo. For instance, we observe that for some human transcripts and RBPs, binding sites are not observed over several kilobases, while other regions, such as the UTRs, harbor a dense clustering of protein-RNA interactions. Further, by design, positive samples in the training dataset contain the RBP binding-site at the central position of the input window, while in practice a binding site may or may not be contained at any input window position. We therefore require a means to more accurately estimate the performance of the training models in a more practical context, namely, the de-novo prediction of RBP binding-sites along a several kilobase long RNA. We introduce the concept of performance-in-practice (PIP), which measures how well the single-nucleotide prediction scores of the model correlate with binding sites identified by ENCODE. For a given RNA sequence, the PIP of a model is the Spearman correlation coefficient (SCC) between the truncated prediction scores $p_{i,trunc}$ and a binary vector obtained by labeling all positions that fall within ENCODE binding sites with 1 and 0 otherwise. Here, $p_{i,trunc}$ refers to a modified version of the prediction score p_i defined as $p_{i,trunc} = p_i$ if $p_i > t_m$ else 0, where t_m is a threshold obtained for each model as outlined in Section 3.4. For each model, we perform extensive PIP analysis as follows. First, we select the set of transcripts which contain at least one binding site for the given RBP. From this set, we uniformly draw 100 transcripts without replacement as hold-out transcripts. Subsequently, we intersect positive and negative training samples with the hold-out transcripts and discard all samples that overlap with any of the hold-out transcripts before retraining pysster on the remaining training samples. We use the resulting models to predict along the hold-out transcripts as described in section 3.5 and compute the PIP score for each hold-out transcript. Finally, models with a median PIP score of less than 0.1 were excluded from downstream analysis.

3.7 Estimating Significance of Prediction Scores

To directly control the false-positive rate of binding site prediction on the viral sequence, we estimate prediction score significance via an RNA sequence permutation test. In order to obtain a null-distribution of predictions (class-1) scores, we first compute the di-nucleotide frequencies on the viral RNA. Subsequently, we perform frequency-weighted sampling of di-nucleotides to construct a set of N 400b null-distributed input samples and obtain a set of null-distributed prediction scores by predicting on those input samples. Note that a dedicated null-distribution is computed for each model. We chose N = 10,000 as it gives an adequate balance between computational burden and accuracy of the null-distribution estimation. Finally, for any sequence window from the viral genome scanned with a model, we compute a p-value for the prediction score, as the fraction of scores from the null distribution s_{null} for which $s_{i,null} > s_{obs}$, i=1,...,N.

A similar approach was applied for the DeepRipe models, where the null-distributed input samples were generated using second order Markov models.

3.8 Identifying RBP Binding-Sites

We identify RBP binding sites on viral RNA sequence using predicted single-nucleotide binding probabilities (Section 3.5) together with estimated p-values (Section 3.6). For each RBP, we classify each nucleotide in the viral RNA as "bound" if and only if the predicted probability score is equal or greater than the estimated binary threshold t_m (Section 3.4) and the score is found to be significant. Each region with a consecutive stretch of "bound" nucleotides of at least length 2 are then defined as a RBP binding site. Neighboring binding sites that are spaced by less than 10 nucleotides are merged to a single binding site.

3.9 Base-wise Feature Attribution via Integrated Gradients

To gain insight into which RNA sub-sequence are driving factors RBP-binding, we compute sequence importance scores using Integrated Gradients (IGs) [Sundararajan et al. (2017)],[Ghanbari and Ohler (2020)]. Starting from an input baseline, IG performs step-wise linear path interpolation between the baseline and the actual input sequence and computes the gradients of the interpolated input and an output neuron, thus avoiding the gradient saturation problem commonly observed when computing feature importance using gradients x input attribution [Sundararajan et al. (2017)]. Here, we choose 0-vector (all nucleotides set to 0) as the baseline and perform 50 baseline-input interpolation steps. To obtain sequence importance scores for a given binding site, we compute IGs with respect to an input window centered around the binding site.

3.10 DeepRiPe Model

We used a modified version of the DeepRiPe deep neural network [Ghanbari and Ohler (2020)], in which predictions are purely sequence-based. Using a sliding window of size 150 and stride 1, we predicted the binding probability at each nucleotide along the viral sequence in analogy to the procedure described in Section 3.5. Empirical p-values for each position are obtained as outlined in Section 3.7.

3.11 Analyzing Mutations in Variants-of-Concern

Variant information of 10 SARS-CoV-2 viral variants (alpha, beta, delta, epsilon, eta, gamma, iota, kappa, lambda, mu) was obtained from the UCSC genome-browser for the SARS-CoV-2 virus [Fernandes et al. (2020)]. For each strain, a mutated FASTA file was generated using the set of corresponding variants. We then center a window at the reference position of each genomic variant and extract the mutated input sequence for subsequent prediction via each model. We note that in case of genomic variants that are in close proximity with each other, extracted windows might contain multiple mutations. This is crucial, as only the combination of multiple variants might lead to gain-or loss of binding. The resulting prediction score on the alternative allele (ALT) is then compared with the prediction score of the same window on the reference sequence (REF). To quantify the impact of each mutation, we compute a *delta* score between the prediction score of ALT and REF sequence:

$$\Delta_{score} = score_{ALT} - score_{REF}.$$

Mutations with a positive delta score sign represent 'gain-of-binding' (GOB) events, while mutations with negative sign represent 'loss-of-binding' (LOB) events. As we are interested in mutations that impact RBP binding to the virus, we further compile a set of change-of-binding (COB) mutations which are defined as mutations where (in case of LOB) the REF score is passing the binding site score threshold and p-value (Methods; 3.4 and 3.7) while the ALT does not, or vice versa (in case of gain of binding). Here, we use a more permissive significance level of 0.05 as p-value threshold.

3.12 *In Silico* Probing of Gain-and Loss-of-Binding Point Mutations

We perform *in silico* probing of the effects of point-mutations on RBP binding across the SARS-CoV-2 genome as follows. At each viral genome position, the reference base was mutated to each of the three alternative bases. Subsequently, prediction was performed on the input windows derived from each ALT allele using all high-quality models. Finally, in analogue to 3.11, a ALT - REF delta score is computed and a set of COB mutations is compiled. For visualization purpose, we display only the mutation with the highest absolute delta at each position.

3.13 Comparative Analysis of Human coronaviruses

Besides SARS-CoV-2, we obtained reference sequences for 6 other human coronaviruses, including SARS-CoV-1, MERS, HCoV-229E, HCoV-HKU1, HCoV-NL63 and HCoV-OC43 from NCBI [Sayers et al. (2021)]. Using all models that pass the performance thresholds (Section 3.6), we perform single-nucleotide binding prediction along each viral

RNAs (Section 3.5). Next we compute prediction empirical p-values for each viral sequence as outlined in Section 3.7) by generating a dedicated null distribution of scores for each virus and RBP. RBP binding sites across viruses were identified as described in Section 3.8. We compute binding site enrichment for each RBP, virus and a sub-set shared viral genomic locations (ORF1AB, E, N, M, S, 5' UTR, 3' UTR) by intersecting the predicted set of binding sites of each virus with its RefSeq annotations. To compute multiple sequence alignments (MSA) between genomic elements of betacoronaviruses, we use the ClustalO [Sievers and Higgins (2014)] algorithm with default parameters.

4 Discussion

RNA viruses contain RNA sequences and structures that play critical roles in the virus life cycle. Host cell RNA Binding proteins (RBPs) which interact with these RNA elements influence viral RNA stability, replication and escape of host immune response. An exhaustive screening of potential human RNA Binding Proteins 'sequestrated' by the virus to fulfill its purposes or involved in the host immune response is missing. While recent experimental studies based on ChIRP [Flynn et al. (2021)] or antisense RNA purification followed by mass spectrometry [Schmidt et al. (2021)] enable a proteome wide identification of RBPs binding SARS-CoV-2, they do not provide a direct readout or protein-RNA interactions and are limited to the cell lines where the experiment is performed. In addition, once the relevant RBPs have been identified, genomic techniques such as CLIP-seq need to be employed to retrieve the near single-nucleotide resolution interaction profile of each RBP with the viral genome once at a time. This is time consuming and CLIP-seq experiments in infected cells for detecting host-pathogen interaction detection have been limited so far to only two RBPs, CNBP and LARP1 identified in schmidt et al. [Schmidt et al. (2021)] as putative regulators of SARS-CoV-2 infections through viral RNA binding [Schmidt et al. (2021)]. This highlights that computational approaches able to screen large numbers of human proteins for putative interactions with the viral RNA are crucial to prioritize candidates for further experimental investigation. In this study we construct a first human-virus RBP-RNA interaction map for SARS-CoV-2 leveraging *in vivo* high-throughput data available for a large number of human RBPs, and predict the propensity of each RBP to bind the viral RNA with convolutional neural networks. We perform accurate quality control of our models by excluding low performing RBP models or models where no clear motif could be learned, and optimize them to extrapolate predictions from the specific experiment where the models are trained on to another context, namely the viral RNA. We demonstrate for several proteins from that the sequence patterns resulting from the attribution maps were consistent with previously reported motifs, and that the predicted binding sites on SARS-CoV-2 are therefore trustworthy.

Compared to recent computational studies performing motif scanning over the entire SAR-CoV-2 genome [Srivastava et al. (2014), Bartas et al. (2020)] or mapping host RBP-viral RNA interactions based on models trained on computed sequence and structure descriptors of RNA-RBP complexes [Vandelli et al. (2020)], our method has several advantages. First, through the use of convolutional Neural Networks (CNNs), we can learn more complex sequence patterns associated to RBP binding site's , other than linear motifs, and therefore be more powerful in predicting new binding sites, and second, we use *in vivo* CLIP-seq data for training, therefore aiming at improving the prioritization of functional interactions. A recent study used a deep learning model trained on *in vivo* structural data, as well as *in vivo* RBP binding data to predict 42 functionally related host proteins binding to the UTRs of the SARS-CoV-2 genome [Sun et al. (2021)], with a follow-up validation of some of those interactions the author set an important milestone in the field. However, their provided RBP-bound regions on the viral RNA are rather large, up to hundreds of nucleotides, hindering a near single-nucleotide analysis, and are limited to the viral UTRs only, which cover a very limited portion of the SARS-CoV-2 genome, while the whole 30kb RNA might be heavily regulated by RBPs. This makes a direct comparison with our method limited.

We mainly identify two sets of high-confidence RBPS putatively binding the SARS-CoV-2 genome: RBPs which were already identified or predicted to bind the viral RNAs by previous studies, and associated to protein-protein interaction networks relevant in the context of SARS-CoV-2 infection (Table 1) and new candidate RBPs, which were not previously reported, and open up new avenues for follow-up experimental studies (Table 2). In the first group we highlight the stress granules-related protein TARDBP, harbouring high-confidence predicted binding sites from both pysster and DeepRIPE with their corresponding learned sequence motifs. This first group also includes the core stress granule components G3BP1 and CAPRIN1, already identified as key regulators of SARS-CoV-2 infected cells [Johnson et al. (2021)] and associated to pathways such as ER stress, Inflammation, cytokine storm and others (Table 1). In particular, protein G3BP3 has been found to interact specifically with SARS-CoV-2 nucleocapsid (N) protein, which is essential for viral replication and localizes at stress granules where it sequesters G3BP3 away from its typical interactions partners [Nabeel-Shah et al. (2022)].Our and previous data (Table1) suggest that pervasive direct binding of G3BP3 to the SARS-CoV-2 RNA could consistute an additional mechanisms used by the virus to interfere with the G3BP3 PPI network and impair stress granule formation. The fact that G3BP3 binding is enriched in correspondence of the sgRNA encoding for protein N (Figre 3c-d) might also suggest a direct regulation of this transcript y this RBP in a sort of feedback loop manner. Other RBPs with predicted binding to SARS-CoV-2, also confirmed by previous studies, include QKI [Sun et al. (2021)], PTBP1 and DD3X3 [Vandelli et al. (2020)].

The differential effects of viral sequences and their role in sequestering host RBPs could imply a deregulation or association of RBPs in COVID-19 patient data. We additionally interrogated gene, protein and genetic associations to severe COVID-19 disease courses for interesting RBPs identified in this study. Several variants in the RBP coding regions are found to be nominally significantly associated to COVID-19 severity as published by the COVID -19 Host Genetics Initiative (A2, release 6, [Initiative and Ganna (2021)]). One of which was MBNL1, an RBP identified for

the first time in our study to interact with the SARS-CoV-2 genome (variant with strongest association has p-value of 0.009), and FXR2 (variant lowest p-value of 0.0055). FXR2 additionally was found to selectively interact with MERS viral proteins but not with viral proteins from SARS-CoV-1 and SARS-CoV-2 [Gordon et al. (2020a)]. This is in agreement with our findings that, in the comparative analysis with other betacoronaviruses, we could observe extensive binding of FXR2 along the 3' UTR of SARS-CoV-1 and MERS, while SARS-CoV-2 showed a complete lack of predicted FXR2 binding sites, owing to its significantly shorter 3' UTR. The FTO protein was identified as key risk factor for obesity, which is also a known risk for COVID-19 severity. FTO coding region harbored also nominal genetic associations to COVID-19 severity (variant lowest p-value of 0.0053). Both FTO and MBNL1 RBPs were additionally found to be significantly regulated on gene level in blood serum of patients admitted to ICU care (FTO with adj. p-value 7.72E-06 and MBNL1 with adj. p-value: 0091496) [Overmyer et al. (2021)].

The application of our trained models enabled the prediction of putative RBP binding at single base resolution. The scores output by the models can be used to quantify how changes in the nucleotide sequence affect the potential binding for RBPs of interest. This is of particular importance to both better understand the efficiency of specific lineages of SARS-CoV-2, as well as provide with a map of mutations of high potential for hijacking important host RBPs, or on the contrary evade binding of anti-viral RBPs. We thus generated mutated versions of the SARS-CoV-2 reference genome by spiking in sets of variants-of-concern identified for the 10 main viral variants identified since the beginning of the pandemic, such as the alpha and delta viral variants. Our analyses first confirmed that the models can be effectively used to pinpoint mutations with high impact potential using the prediction scores. These mutations can be further evaluated through the computation of attribution maps, allowing to spot which are the important nucleotide in a given window of interest, and how the importance of these nucleotides is impacted by the mutation. We could identify that some RBPs presented very large gain or loss of binding across many strains, such as the fat mass and obesity-associated protein (FTO), found as bearing such event in 7 out of 10 strains. FTO is demethylase [Fu et al. (2013)], and while it's been suggested that the virus could hijack the host epigenome [Atlante et al. (2020)], a recent study showed that the viral genome itself was methylated Liu et al. (2021b). The authors showed that methylation was negatively associated with replication efficiency. Although there was no clear trend towards systematic loss of binding of FTO across the viral variants, we were able to point out multiple close-by mutations in the alpha variant that were associated to a significant loss, around the position 28,280 (Figure 4D).

Another RBP of interest is the Serine/arginine-rich splicing factor 7 (SRSF7). It has been suggested that this spliceosome protein could be sequestrated by the viral genome [Srivastava et al. (2020)], the later thus acting as a sponge through these putative binding sites and affecting splicing in the host transcriptome. Nonetheless, SR factors have also been implicated in increasing translation efficiency in the context of HIV infection [Mahiet and Swanson (2016)]. Among the high-impact mutations, two are of particular interest: a G to T transversion at position 210 in the 5' UTR, found both in delta and kappa viral variants, is associated with a delta score of -0.31. Another position is found mutated across multiple strains, with actually different alternative nucleotides: at position 23,604 (S protein gene) a C to A transition is found in alpha and mu variants, while a C to G transversion is found in delta and kappa variants. Both mutations are associated to a gain of binding (+0.39 for C>G, +0.32 for C>A). This position has been suggested to be a major driver of the increased infection efficiency of these viral variant, as a modifier of the S protein sequence (P680R) [Liu et al. (2021c)], although additional studies indicate that other mutations may be required for an actual effect ([Lubinski et al. (2021)], [Zhang et al. (2021)]). The gain of binding we identify here could also suggest that the translation of the S gene into the protein is improved through the recognition of the newly created binding site by SRSF7. Overall, variants of concern have been pinpointed through their potential impact on the sequences of viral proteins that are of major importance for the life cycle of the virus, from entry in host cells to replication. It is nonetheless very interesting to see them as having high impacts on putative RBP binding.

Comparative analysis of RBP-RNA interactions across seven coronaviruses sheds light onto genomic features and factors which confer unique characteristics to SARS-CoV-2 transmission and pathogenicity compared to SARS-CoV-1, MERS and less pathogenic coronaviruses. Both variants of concern and comparative analysis highlight gain or loss of binding affecting host RBP-viral interactions and therefore pinpoint RBPs which can be prioritized for further screening. Because the COVID-19 pandemic is still ongoing, the analysis provided here might also be useful in assisting medical experts in fighting the pandemic's progression. We found RBPs with important roles in other RNA viruses infections and this is important for the task of re-purposing existing drugs. In fact, if similar RBP-RNA interactions are detected in other viruses, the similarity to better understood viruses can help providing clinical testing of RBP anti-viral agents. Until adequate curative treatments, besides vaccination against SAR-CoV-2 infection is found, our lives and our society will be deeply affected. Therefore this pandemic needs both short-term solutions, such as finding drugs to treat massive numbers of sick patients, as well as long-term development of preventive and curative drugs for future Corona outbreaks.

References

Atlante, S., Mongelli, A., Barbi, V., Martelli, F., Farsetti, A., and Gaetano, C. (2020). The epigenetic implication in coronavirus infection and therapy. *Clinical Epigenetics*, 12(1):156.

Bartas, M., Brázda, V., Bohálová, N., Cantara, A., Volná, A., Stachurová, T., Malachová, K., Jagelská, E., Porubiaková, O., Červeň, J., and Pecinka, P. (2020). In-depth bioinformatic analyses of nidovirales including human sars-cov-2, sars-cov, mers-cov viruses suggest important roles of non-canonical nucleic acid structures in their lifecycles. 11:1583.

Bojkova, D., Klann, K., Koch, B., Widera, M., Krause, D., Ciesek, S., Cinatl, J., and Münch, C. (2020). Proteomics of sars-cov-2-infected host cells reveals therapy targets. *Nature*, 583(7816):469–472.

Budach, S. and Marsico, A. (2018). Pysster: classification of biological sequences by learning sequence and structure motifs with convolutional neural networks. *Bioinformatics*, 34(17):3035–3037.

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 m6a rna modification pathway blocks sars-cov-2 and hcov-oc43 replication. *Genes & development*, 35(13-14):1005–1019.

Ciccosanti, F., Di Rienzo, M., Romagnoli, A., Colavita, F., Refolo, G., Castilletti, C., Agrati, C., Brai, A., Manetti, F., Botta, L., Capobianchi, M., Ippolito, G., Piacentini, M., and Fimia, G. (2021). Proteomic analysis identifies the rna helicase ddx3x as a host target against sars-cov-2 infection. *Antiviral Res*, 190:105064.

Davey, N. E., Travé, G., and Gibson, T. J. (2011). How viruses hijack cell regulation. *Trends in biochemical sciences*, 36(3):159–169.

Dominguez, D., Freese, P., Alexis, M. S., Su, A., Hochman, M., Palden, T., Bazile, C., Lambert, N. J., Van Nostrand, E. L., Pratt, G. A., et al. (2018). Sequence, structure, and context preferences of human rna binding proteins. *Molecular cell*, 70(5):854–867.

Farkas, C., Mella, A., Turgeon, M., and Haigh, J. J. (2021). A novel sars-cov-2 viral sequence bioinformatic pipeline has found genetic evidence that the viral 3 untranslated region (utr) is evolving and generating increased viral diversity. *Frontiers in microbiology*, 12.

Fernandes, J. D., Hinrichs, A. S., Clawson, H., Gonzalez, J. N., Lee, B. T., Nassar, L. R., Raney, B. J., Rosenbloom, K. R., Nerli, S., Rao, A. A., Schmelter, D., Fyfe, A., Maulding, N., Zweig, A. S., Lowe, T. M., Ares, M., Corbet-Detig, R., Kent, W. J., Haussler, D., and Haeussler, M. (2020). The UCSC SARS-CoV-2 Genome Browser. *Nature Genetics*, 52(10):991–998.

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(9):2394–2411.

Fu, Y., Jia, G., Pang, X., Wang, R. N., Wang, X., Li, C. J., Smemo, S., Dai, Q., Bailey, K. A., Nobrega, M. A., Han, K.-L., Cui, Q., and He, C. (2013). FTO-mediated formation of N6-hydroxymethyladenosine and N6-formyladenosine in mammalian RNA. *Nature Communications*, 4(1).

Fung, T. S. and Liu, D. X. (2019). Human coronavirus: host-pathogen interaction. *Annual review of microbiology*, 73:529–557.

Garcia-Moreno, M., Järvelin, A. I., and Castello, A. (2018). Unconventional rna-binding proteins step into the virus–host battlefront. *Wiley Interdisciplinary Reviews: RNA*, 9(6):e1498.

Ghanbari, M. and Ohler, U. (2020). Deep neural networks for interpreting rna-binding protein target preferences. *Genome Research*, 30(2):214–226.

Gordon, D. E., Hiatt, J., Bouhaddou, M., Rezelj, V. V., Ulferts, S., et al. (2020a). Comparative host-coronavirus protein interaction networks reveal pan-viral disease mechanisms. *Science*, 370(6521):eabe9403.

Gordon, D. E., Jang, G. M., Bouhaddou, M., Xu, J., Obernier, K., White, K. M., O'Meara, M. J., Rezelj, V. V., Guo, J. Z., Swaney, D. L., et al. (2020b). A sars-cov-2 protein interaction map reveals targets for drug repurposing. *Nature*, 583(7816):459–468.

Hafner, M., Katsantoni, M., Köster, T., Marks, J., Mukherjee, J., Staiger, D., Ule, J., and Zavolan, M. (2021). Clip and complementary methods. *Nature Reviews Methods Primers*, 1(1):1–23.

Hentze, M. W., Castello, A., Schwarzl, T., and Preiss, T. (2018). A brave new world of rna-binding proteins. *Nature reviews Molecular cell biology*, 19(5):327–341.

Initiative, C.-. H. G. and Ganna, A. (2021). Mapping the human genetic architecture of covid-19: an update. *medRxiv*.

Johnson, N andNeufeldt, C., Cortese, M., Knight, M., Lilley, K., Martinez, J., Davis, I., Bartenschlager, R., Mohammed, S., and Castello, A. (2021). Global analysis of protein-rna interactions in sars-cov-2-infected cells reveals key regulators of infection. *Mol Cell*, 81(13):2851.

Kim, D., Lee, J., Yang, J., Kim, J., Kim, V., and Chang, H. (2020). The architecture of sars-cov-2 transcriptome. *Cell*, 181(4):914–921.

Kim, D.-K., Weller, B., Lin, C.-W., Sheykhkarimli, D., Knapp, J. J., Kishore, N., Sauer, M., Rayhan, A., Young, V., Marin-de la Rosa, N., Pogoutse, O., et al. (2021). A map of binary sars-cov-2 protein interactions implicates host immune regulation and ubiquitination. *bioRxiv*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kuo, P.-H., Chiang, C.-H., Wang, Y.-T., Doudeva, L. G., and Yuan, H. S. (2014). The crystal structure of tdp-43 rrm1-dna complex reveals the specific recognition for ug-and tg-rich nucleic acids. *Nucleic acids research*, 42(7):4712–4722.

Labeau, A., Lefevre-Utile, A., Bonnet-Madin, L., Fery-Simonian, L., Soumelis, V., Lotteau, V., Vidalain, P.-O., Amara, A., and Meertens, L. (2021). Characterization and functional interrogation of sars-cov-2 rna interactome. *BioRxiv*.

Lee, S., Lee, Y.-s., 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*.

Li, Z. and Nagy, P. (2011a). Diverse roles of host rna binding proteins in rna virus replication. *RNA Biol*, 8(2):305–15.

Li, Z. and Nagy, P. D. (2011b). Diverse roles of host rna binding proteins in rna virus replication. *RNA biology*, 8(2):305–315.

Licata, L., Lo Surdo, P., Iannuccelli, M., Palma, A., Micarelli, E., Perfetto, L., Peluso, D., Calderone, A., Castagnoli, L., and Cesareni, G. (2020). Signor 2.0, the signaling network open resource 2.0: 2019 update. *Nucleic Acids Res*, 8(48(D1)):D504–D510.

Liu, D. X., Liang, J. Q., and Fung, T. S. (2021a). Human coronavirus-229e,-oc43,-nl63, and-hku1 (coronaviridae). *Encyclopedia of Virology*, page 428.

Liu, J., Xu, Y.-P., Li, K., Ye, Q., Zhou, H.-Y., Sun, H., Li, X., Yu, L., Deng, Y.-Q., Li, R.-T., Cheng, M.-L., He, B., Zhou, J., Li, X.-F., Wu, A., Yi, C., and Qin, C.-F. (2021b). The m6A methylome of SARS-CoV-2 in host cells. *Cell Research*, 31(4).

Liu, Y., Liu, J., Johnson, B. A., Xia, H., Ku, Z., Schindewolf, C., Widen, S. G., An, Z., Weaver, S. C., Menachery, V. D., Xie, X., and Shi, P.-Y. (2021c). Delta spike P681R mutation enhances SARS-CoV-2 fitness over Alpha variant. *bioRxiv*, page 2021.08.12.456173.

Lubinski, B., Frazier, L. E., Phan, M. V. T., Bugembe, D. L., Tang, T., Daniel, S., Cotten, M., Jaimes, J. A., and Whittaker, G. R. (2021). Spike protein cleavage-activation mediated by the SARS-CoV-2 P681R mutation: a case-study from its first appearance in variant of interest (VOI) A.23.1 identified in Uganda. Technical report.

Luo, H., Chen, Q., Chen, J., Chen, K., Shen, X., and Jiang, H. (2005). The nucleocapsid protein of sars coronavirus has a high binding affinity to the human cellular heterogeneous nuclear ribonucleoprotein a1. *FEBS letters*, 579(12):2623–2628.

Mahiet, C. and Swanson, C. M. (2016). Control of HIV-1 gene expression by SR proteins. *Biochemical Society Transactions*, 44(5):1417–1425.

Manokaran, G., Finol, E., Wang, C., Gunaratne, J., Bahl, J., Ong, E., Tan, H., Sessions, O., Ward, A., Gubler, D., Harris, E., Garcia-Blanco, M., and Ooi, E. (2015). Dengue subgenomic rna binds trim25 to inhibit interferon expression for epidemiological fitness. *Science*, 9(350):6257.

Meyer, K. D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C. E., and Jaffrey, S. R. (2012). Comprehensive analysis of mrna methylation reveals enrichment in 3 utrs and near stop codons. *Cell*, 149(7):1635–1646.

Molleston, J. M. and Cherry, S. (2017). Attacked from all sides: Rna decay in antiviral defense. *Viruses*, 9(1):2.

Mukherjee, M. and Goswami, S. (2020). Global cataloguing of variations in untranslated regions of viral genome and prediction of key host rna binding protein-microrna interactions modulating genome stability in sars-cov-2. *PloS one*, 15(8):e0237559.

Mukherjee, N., Wessels, H., Lebedeva, S., Sajek, M., Ghanbari, M., Garzia, A., Munteanu, A., Yusuf, D., Farazi, T., Hoell, J., Akat, K., Akalin, A., Tuschl, and Ohler, U. (2019). Deciphering human ribonucleoprotein regulatory networks. *Nucleic Acids Res*, 47(2):570–581.

Nabeel-Shah, S., Lee, H., Ahmed, N., Burke, G., Farhangmehr, S., Ashraf, K., Pu, S., Braunschweig, U., Zhong, G., Wei, H., Tang, H., Yang, J., Marcon, E., Blencowe, B., Zhang, Z., and Greenblatt, J. (2022). Sars-cov-2 nucleocapsid protein binds host mrnas and attenuates stress granules to impair host stress response. *iScience*, 25(1):103562.

Overmyer, K., Shishkova, E., Miller, I., Balnis, J., Bernstein, M., Peters-Clarke, T., Meyer, J., Quan, Q., Muehlbauer, L., Trujillo, E., and He, Y. e. a. (2021). Large-scale multi-omic analysis of covid-19 severity. *Cell Syst*, 12(1):23–40.

Pan, X., Yang, Y., Xia, C.-Q., Mirza, A. H., and Shen, H.-B. (2019). Recent methodology progress of deep learning for rna–protein interaction prediction. *Wiley Interdisciplinary Reviews: RNA*, 10(6):e1544.

Sayers, E. W., Beck, J., Bolton, E. E., Bourexis, D., Brister, J. R., Canese, K., Comeau, D. C., Funk, K., Kim, S., Klimke, W., et al. (2021). Database resources of the national center for biotechnology information. *Nucleic acids research*, 49(D1):D10.

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. *Nature Microbiology*, 6(3):339–353.

Schneider, W. M., Luna, J. M., Hoffmann, H.-H., Sánchez-Rivera, F. J., Leal, A. A., Ashbrook, A. W., Le Pen, J., Ricardo-Lax, I., Michailidis, E., Peace, A., Stenzel, A. F., Lowe, S. W., MacDonald, M. R., Rice, C. M., and Poirier, J. T. (2021). Genome-scale identification of sars-cov-2 and pan-coronavirus host factor networks. *Cell*, 184(1):120–132.e14.

Sievers, F. and Higgins, D. G. (2014). Clustal omega, accurate alignment of very large numbers of sequences. In *Multiple sequence alignment methods*, pages 105–116. Springer.

Sola, I., Almazán, F., Zúñiga, S., and Enjuanes, L. (2015). Continuous and discontinuous rna synthesis in coronaviruses. *Annu Rev Virol*, 1(3):265–88.

Soto-Acosta, R., Xie, X., Shan, C., Baker, C. K., Shi, P.-Y., Rossi, S. L., Garcia-Blanco, M. A., and Bradrick, S. (2018). Fragile x mental retardation protein is a zika virus restriction factor that is antagonized by subgenomic flaviviral rna. *Elife*, 7:e39023.

Spagnolo, J. F. and Hogue, B. G. (2000). Host protein interactions with the 3 end of bovine coronavirus rna and the requirement of the poly (a) tail for coronavirus defective genome replication. *Journal of virology*, 74(11):5053–5065.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Srivastava, R., Daulatabad, S. V., Srivastava, M., and Janga, S. C. (2020). Role of SARS-CoV-2 in altering the RNA binding protein and miRNA directed post-transcriptional regulatory networks in humans. *bioRxiv*, page 2020.07.06.190348.

Stukalov, A., Girault, V., Grass, V., Karayel, O., Bergant, V., Urban, C., Haas, D. A., Huang, Y., Oubraham, L., Wang, A., Hamad, M. S., et al. (2021). Multilevel proteomics reveals host perturbations by sars-cov-2 and sars-cov. *Nature*, 594(7862):246–252.

Sun, L., Li, P., Ju, X., Rao, J., Huang, W., Ren, L., Zhang, S., Xiong, T., Xu, K., Zhou, X., Gong, M., Miska, E., Ding, Q., Wang, J., and Zhang, Q. C. (2021). In vivo structural characterization of the sars-cov-2 rna genome identifies host proteins vulnerable to repurposed drugs. *Cell*, 184(7):1865–1883.e20.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR.

Tan, Y. W., Hong, W., and Liu, D. X. (2012). Binding of the 5-untranslated region of coronavirus rna to zinc finger cchc-type and rna-binding motif 1 enhances viral replication and transcription. *Nucleic acids research*, 40(11):5065–5077.

Tirumuru, N., Zhao, B., Lu, W., Lu, Z., C, H., and Wu, L. (2016). N(6)-methyladenosine of hiv-1 rna regulates viral infection and hiv-1 gag protein expression. *Elife*, 5:e15528.

Van Nostrand, E., Freese, P., Pratt, G., Wang, X., Wei, X., Xiao, R., Blue, S., Chen, J., Cody, N., and Dominguez, D. (2020a). A large-scale binding and functional map of human rna-binding proteins. *Nature*, 583(7818):711–719.

Van Nostrand, E. L., Freese, P., Pratt, G. A., Wang, X., Wei, X., Xiao, R., Blue, S. M., Chen, J.-Y., Cody, N. A., Dominguez, D., et al. (2020b). A large-scale binding and functional map of human rna-binding proteins. *Nature*, 583(7818):711–719.

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(20):11270–11283.

Vashist, S., Urena, L., Chaudhry, Y., and Goodfellow, I. (2012). Identification of rna-protein interaction networks involved in the norovirus life cycle. *J Virol*, 86(22):11977–90.

V'kovski, P., Kratzel, A., Steiner, S., Stalder, H., and Thiel, V. (2021). Coronavirus biology and replication: implications for sars-cov-2. *Nature Reviews Microbiology*, 19(3):155–170.

Wei, J., Alfajaro, M. M., DeWeirdt, P. C., Hanna, R. E., Lu-Culligan, W. J., Cai, W. L., Strine, M. S., Zhang, S.-M., Graziano, V. R., Schmitz, C. O., et al. (2021). Genome-wide crispr screens reveal host factors critical for sars-cov-2 infection. *Cell*, 184(1):76–91.e13.

Wu, C.-H., Chen, P.-J., and Yeh, S.-H. (2014). Nucleocapsid phosphorylation and rna helicase ddx1 recruitment enables coronavirus transition from discontinuous to continuous transcription. *Cell host & microbe*, 16(4):462–472.

Xu, Y.-P., Li, K., Ye, Q., Zhou, H.-Y., Sun, H., Li, X., Yu, L., Deng, Y.-Q., Li, R.-T., Cheng, M.-L., et al. (2021). The m 6 a methylome of sars-cov-2 in host cells. *Cell research*, 31(4):404–414.

Zannella, C., Rinaldi, L., Boccia, G., Chianese, A., Sasso, F. C., De Caro, F., Franci, G., and Galdiero, M. (2021). Regulation of m6a methylation as a new therapeutic option against covid-19. *Pharmaceuticals*, 14(11):1135.

Zhang, J., Xiao, T., Cai, Y., Lavine, C. L., Peng, H., Zhu, H., Anand, K., Tong, P., Gautam, A., Mayer, M. L., Walsh, R. M., Rits-Volloch, S., Wesemann, D. R., Yang, W., Seaman, M. S., Lu, J., and Chen, B. (2021). Membrane fusion and immune evasion by the spike protein of SARS-CoV-2 Delta variant. Technical report.

Zhang, X., Hao, H., Ma, L., Zhang, Y., Hu, X., Chen, Z., Liu, D., Yuan, J., Hu, Z., and Guan, W. (2020). Methyltransferase-like 3 modulates severe acute respiratory syndrome coronavirus-2 rna n6-methyladenosine modification and replication. *bioRxiv*.

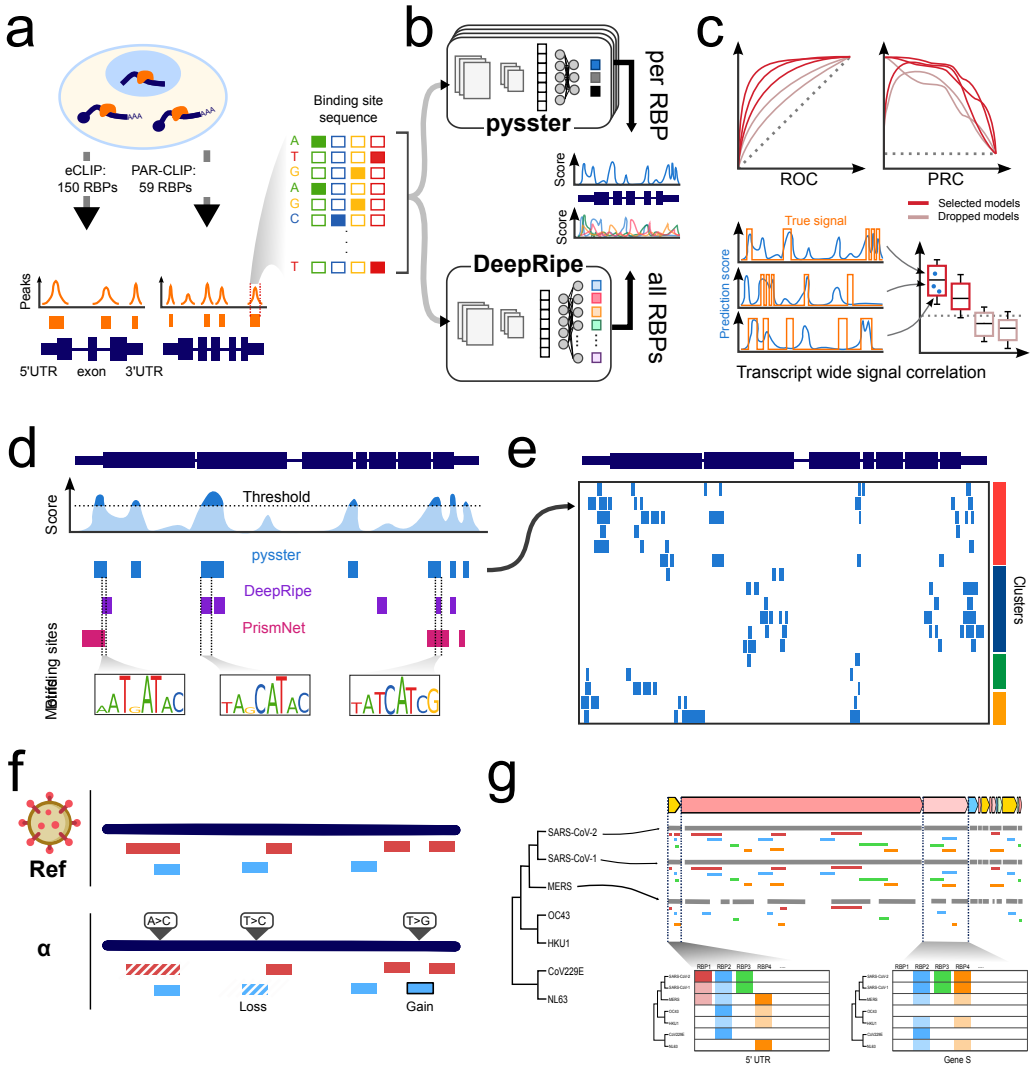


Figure 1: Illustrated pipeline of the computational mapping of the human - SARS-CoV-2 protein-RNA interactome. **a**. Local interactions between RNA-binding proteins (RBPs) and transcripts can be experimentally measured notably through eCLIP and PAR-CLIP protocols, enabling the quantification of locally accumulated reads, and the calling of peaks. Such peaks were obtained from eCLIP data measured by Yeo et al 2016 for 150 RBPs, and from PAR-CLIP data measured by Mukherjee et al 2019 for 59 RBPs. **b**. Sequences from these peaks were used to train two deep learning models, composed of convolutional neural networks enabling the detection of complex sequence motifs. These models can then be applied to predict for a given sequence its potential for binding by an RBP. Pysster models are generated per RBP and exploit different types of negative control sequences for accurate binding site prediction, while DeepRipe makes use of all available RBP narrow peaks, to predict for a new sequence the binding profile of all input RBPs, in a single pass. **c**. A selection of high-performance models was established through complete evaluation of performance of the models, from overall performance metrics to in-practice, sequence-wide evaluation. **d**. All retained models were applied to scan the entire genome of SARS-CoV-2, and binding sites were predicted from consecutive, high-prediction scores positions. These were compared across the different model architectures, and against published data. Moreover, sequence preference was evaluated for models with reported motifs in databases. **e**. These predictions are compiled in the first in-silico map of host-protein - viral RNA interactome. **f**. The prediction models were applied to evaluate the impact of variants of concerns, as well as to evaluate the evolutionary trajectory of affinity of host RBPs to the corona viruses genomes **g**. All in all, the use of prediction models enable the rapid identification of host RBPs likely to be involved in the biological life cycle of the virus.

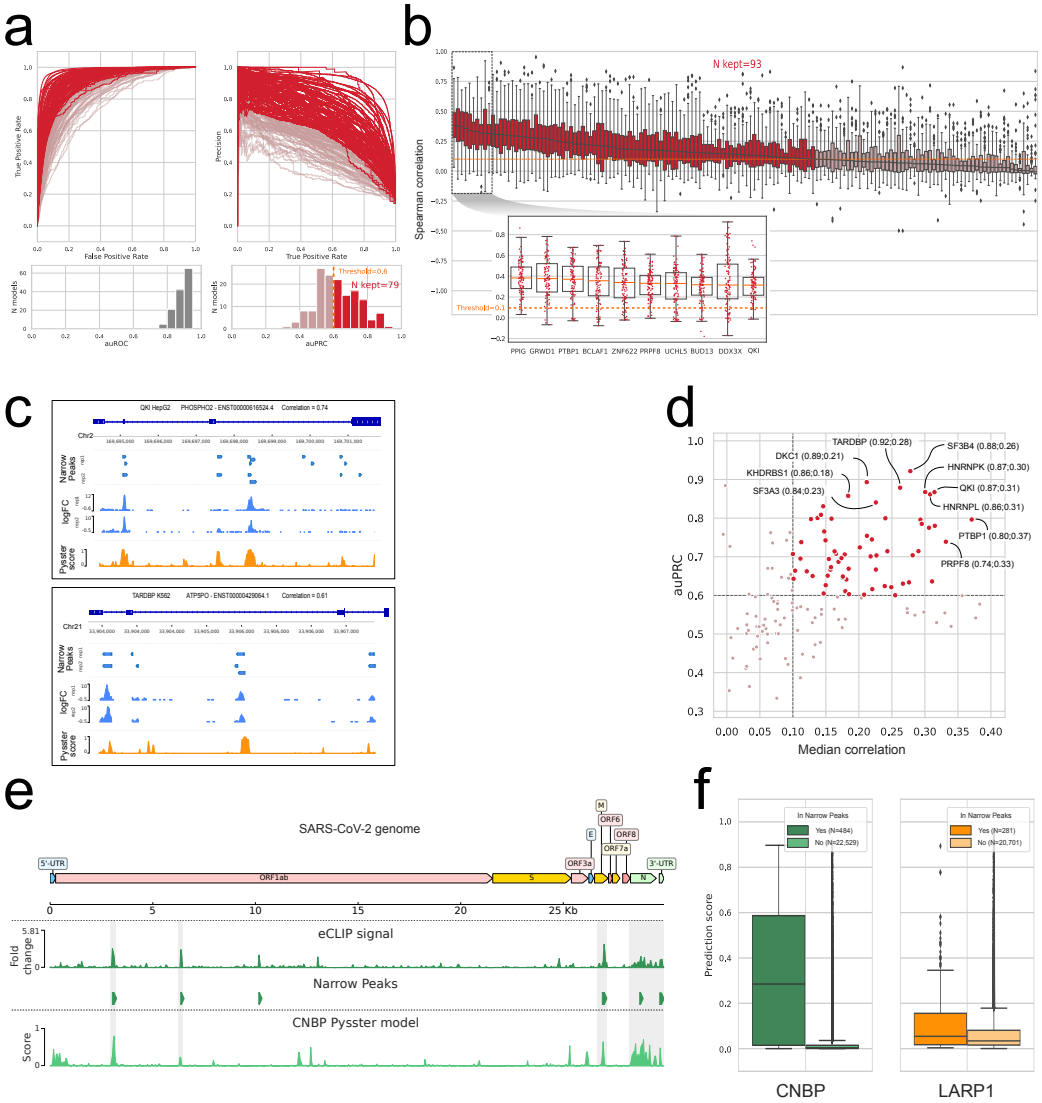


Figure 2: Evaluation of Pysster models' performance and high-quality model selection. **a**. Receiver Operating Curve (ROC) and Precision Recall Curve (PRC) for all 150 Pysster models trained from ENCODE eCLIP datasets. A first threshold was set on the area under the PRCs (AUPRC), leading to a subset of 79 models passing the threshold. **b**. Boxplots of correlations between eCLIP and prediction scores from 100 left-out transcripts per RBP model. This correlation highlights the performance of models in a realistic context full-length scan of a sequence. A second threshold was set on median correlation coefficient, leading to a subset of 93 models passing the threshold. The 10 models with highest median correlation are displayed in a detailed sub-plot. **c**. Genome-browser view illustrating the comparison between eCLIP signals and model prediction scores over full-length transcripts. Two of the best models are presented, with signal from left-out transcripts with high correlation of between eCLIP log-fold-change signals and prediction scores from the Pysster models. **d**. Scatterplot of the AUPRC and median correlation values for each model, highlighting the finale subset of high-quality models. The top 10 models are labeled. **e**. Comparison of genome-wide eCLIP signal and Pysster prediction scores from the CNBP eCLIP datasets generated over the SARS-CoV-2 genome by Schmidt et al. (2021). **f**. Boxplot of Pysster prediction scores from position within or without overlap from called narrow peaks, for the CNBP model and the LARP1 model.

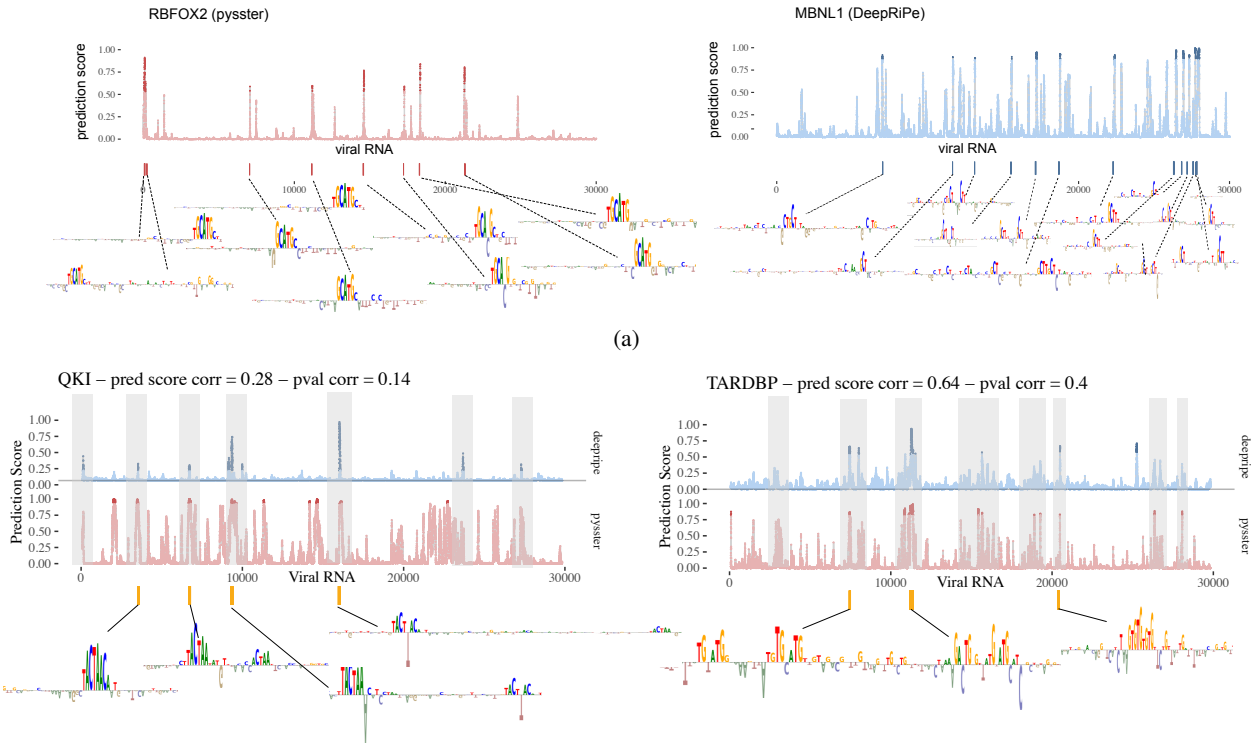
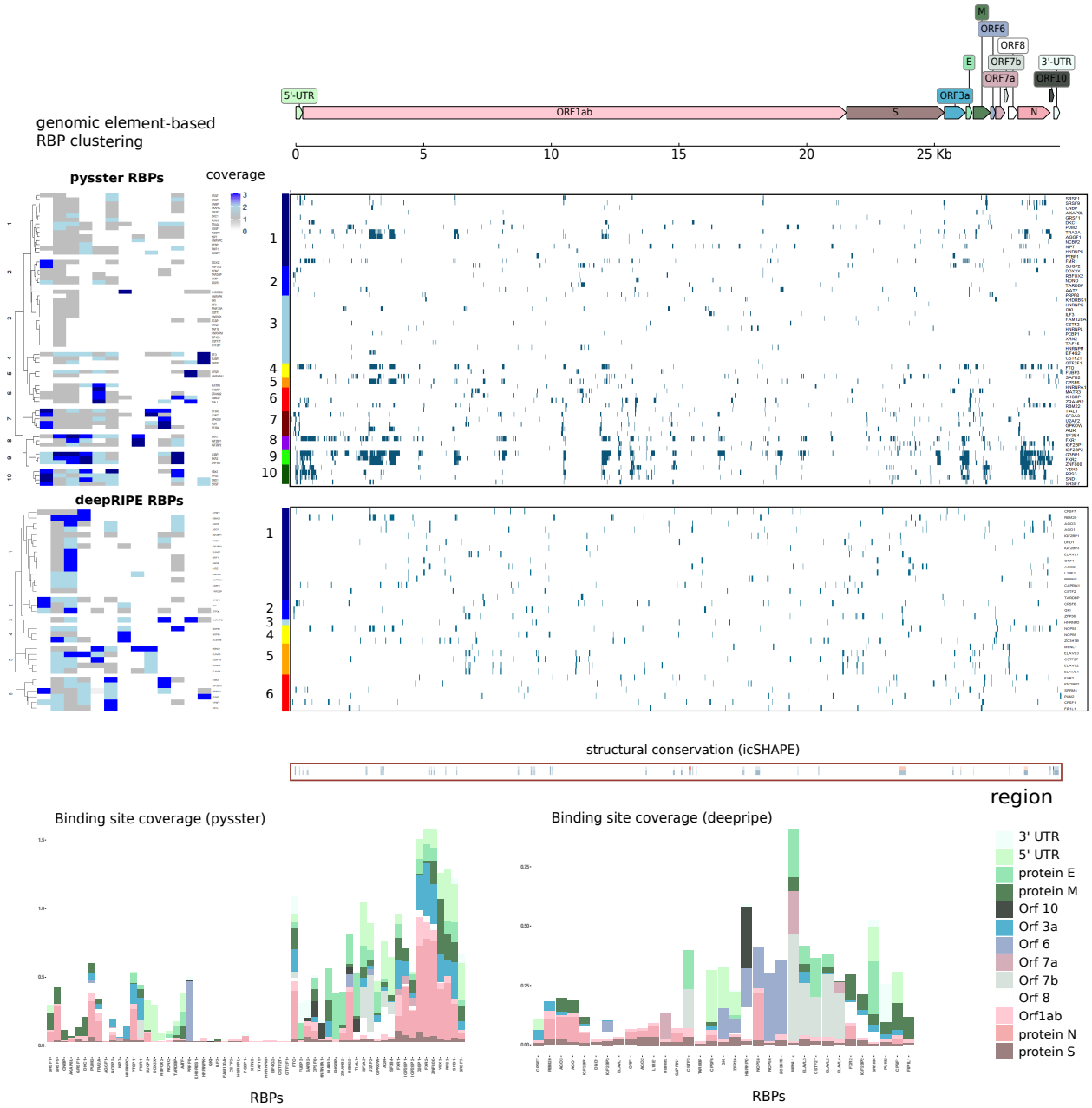


Figure 3: Computational map of RBP binding on SARS-CoV-2. **a** Single-nucleotide probability score for RBFOX2 RBP binding as computed by the corresponding pysster model across the whole SARS-CoV-2 genome (left panel). The higher the score, the higher the likelihood that the corresponding position belongs to an RBFOX2 binding site. Points highlighted in bold correspond to significantly predictions, i.e. with bound probability significantly higher than random (empirical p-value < 0.01, see Methods). Wider RBFOX2 binding sites, encompassing more than one significant position are shown as vertical bars underneath, together with their corresponding binding motifs as extracted by means of attribution maps (see Methods). single-nucleotide probability score for MBNL1 RBP binding as computed by the corresponding DeepRIPE model (right panel). Significant positions (empirical p-value < 0.01) are highlighted in bold, and computed binding sites together with their corresponding motifs are shown underneath. **b** Comparison of single-nucleotide probability scores between pysster and DeepRIPE for two RBPs, QKI (left panel) and TARDBP (right panel). Highlighted with grey boxes are those regions of high agreement between prediction scores from the two models. significant binding sites, commonly predicted by both methods are shown underneath the probability plots together with their corresponding learnt motifs from the attribution maps. **c** Clustering of RBPs based on binding site coverage of different regions on SARS-CoV-2 for both pysster and DeepRIPE RBPs separately (left panel). In-silico RBP binding map, at single-nucleotide resolution, for both pysster and DeepRIPE RBPs (right panel). Structurally conserved regions across several coronaviruses are shown below and are determined based on experimental icSHAPE data (citation). **d** Stacked bars indicate the percentage of each SARS-CoV-2 genomic region which is covered by significant binding sites for each of the pysster and DeeRIPE RBPs.


genomic element-based RBP clustering
pysster RBPs
coverage
deepRIPE RBPs
5'-UTR
ORF1ab
S
ORF3a
E
M
ORF6
ORF8
ORF7b
ORF7a
ORF10
3'-UTR
N
0
5
10
15
20
25 Kb
structural conservation (icSHAPE)
region
Binding site coverage (pysster)
Binding site coverage (deepripe)
RBPs
RBPs
3' UTR
5' UTR
protein E
protein M
Orf 10
Orf 3a
Orf 6
Orf 7a
Orf 7b
Orf 8
Orf1ab
protein N
protein S

(c)

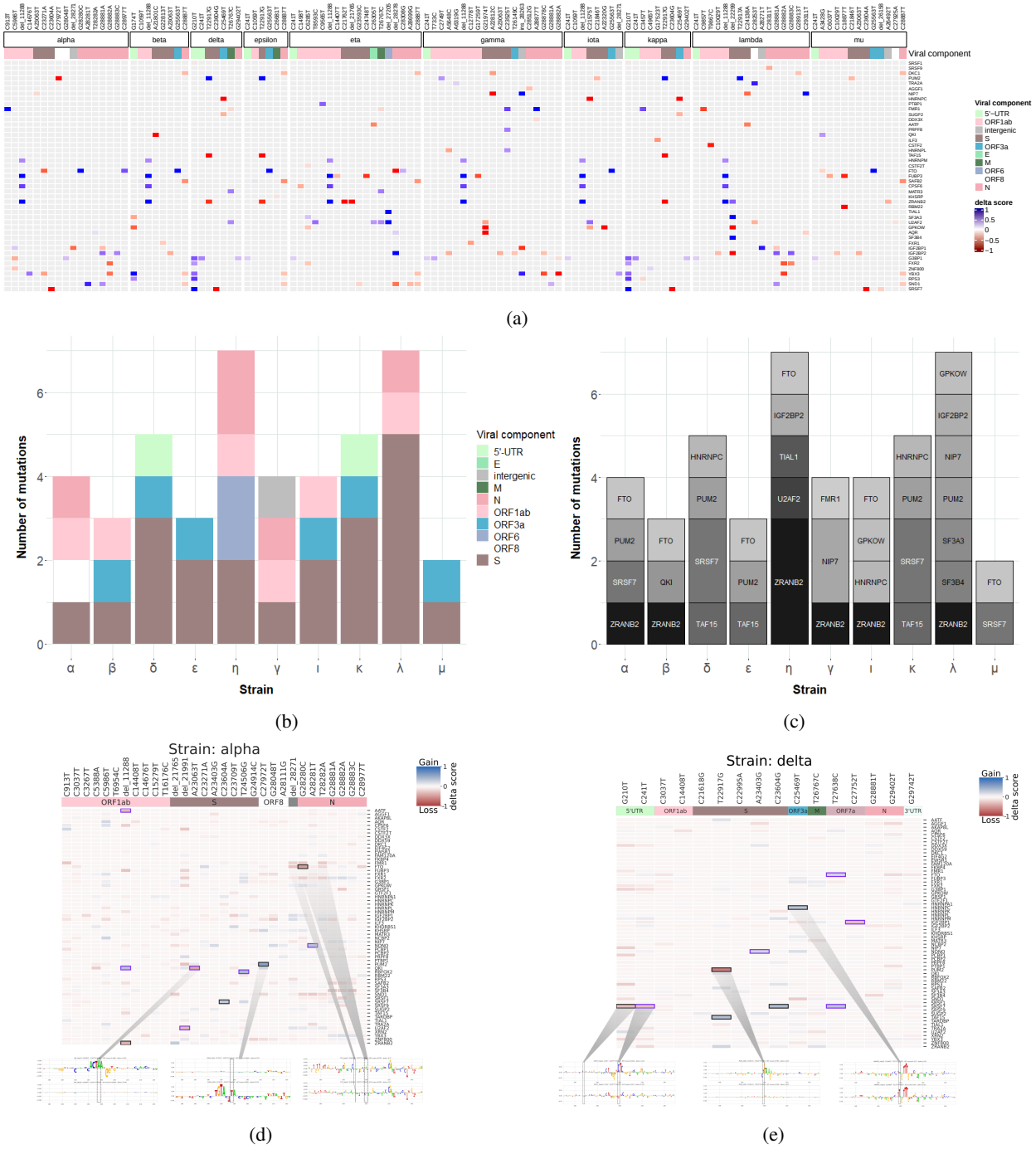


Figure 4: Impact of variants of concern on predicted binding sites. **a**. Joint heatmap of delta scores in Pysster predictions from the 150 variants of concern identified in the different SARS-CoV-2 lineages. Delta-scores represent the difference in prediction score between alternative and reference sequences centered on each mutated viral component, for a given prediction model. A subset of 45 RBPs with at least one impacted binding site are presented, sorted according to the clustering of their binding site patterns (figure 3). Viral components where variants occur are color coded above each strain's heatmap. **b**. Accumulation of high-impact variants of concern in viral components for each lineage. The subset of mutations considered here was established from taking the top 20% across all mutation-RBP combinations. **c**. Accumulation of impacted RBP sites for each lineage. The same subset as in (b) was used here. **d**. Complete heatmap of delta scores in delta scores in Pysster predictions from 31 variants of concern identified in the lineage *alpha*. The top 10 variants of concern with highest absolute delta scores are displayed with line colors (purple: mutated site does not pass the selection thresholds ; black: mutated site passes the selection thresholds). Some sites are further investigated through integrated gradients, where the sequence motifs identified by Pysster models were compared against known motifs from mCrossBase. **e**. Complete heatmap of delta scores in Pysster predictions from 31 variants of concern identified in the lineage *delta*.

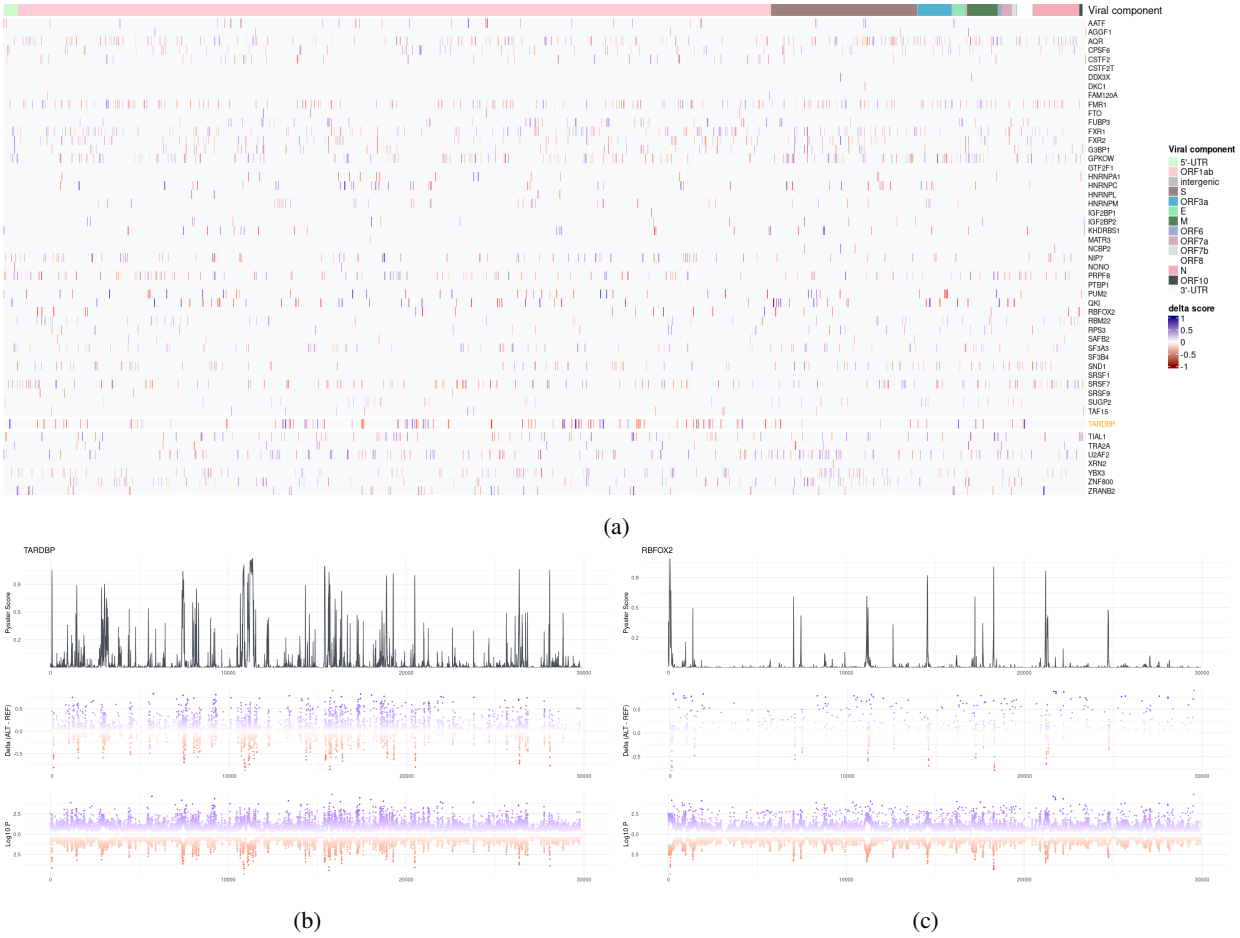


Figure 5: In-silico perturbation analysis of SARS-CoV-2. Nucleotides across the viral genome were perturbed towards the three alternative bases and the alternative base with resulting the highest delta score was considered for downstream analysis. **a)** Delta score heatmap across positions with at least one gain-or loss-of-binding event across all RBPs. A total of 13771 position whose perturbation might affect binding of at least one of the 63 tested RBPs were identified. RBPs such as FTO or XRN2 exhibit strong robustness against single-nucleotide perturbations, with only few positions leading to gain-or loss-of-binding, while others (e.g. AQR and FMR1) are highly sensitive towards perturbations. **b)** and **c)** demonstrate the position-wise delta scores for TARDBP and RBFOX2, respectively.

3' UTR
5' UTR
SARS-CoV-1
MERS
SARS-CoV-2
HCoV-HKU1
HCoV-NL63
HCoV-229E
HCoV-OC43
(a)
(b)
(c)
FTO
FXR2
TARDBP
RBFOX2
Binding Probability
position
(d)
(e)
(f)
(g)

Figure 6: Comparison of SARS-CoV-2 and 6 other betacoronaviruses. RBP binding patterns appear to be conserved between related virus across the viral RNA, with the highly pathogenic viruses SARS-CoV-1, SARS-CoV-2 and MERS exhibiting more similar binding compared to other betacoronaviruses (6a), however, binding across viral elements shows a greater variability and is less conserved (**?**). While the highly pathogenic viruses show conserved binding in their 5' UTR (6c), SARS-CoV-2 deviates from SARS-CoV-1 and MERS in terms of RBP binding in the 3' UTR, likely due to a shorted 3' UTR region (6b). SARS-CoV-1 and SARS-CoV-2 harbor a conserved FTO binding site at the 5' end of the 3' UTR (6d). Meanwhile, FXR2 binding is highly conserved between SARS-CoV-1 and MERS (6e), while SARS-CoV-2 show a loss of FXR2 binding to it's 3' UTR. In the 5' UTR, SARS-CoV-2 shows acquisition of a novel TARDBP binding site, while SARS-CoV-1 and MERS lack TARDBP binding (6f), while binding of RBFOX2 to the 5' UTR of SARS-CoV-1 and SARS-CoV-2 is conserved.