# Structural analysis of SARS-CoV-2 and predictions of the human interactome

Andrea Vandelli[1,2], Michele Monti[1,3], Edoardo Milanetti[4,5], Riccardo Delli Ponti[6,]*
and Gian Gaetano Tartaglia [1,3,5,7,]*

[1] Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Dr. Aiguader 88, 08003 Barcelona, Spain and Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain
[2] Systems Biology of Infection Lab, Department of Biochemistry and Molecular Biology, Biosciences Faculty, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Spain
[3] RNA System Biology Lab, department of Neuroscience and Brain Technologies, Istituto Italiano di Tecnologia, Via Morego 30, 16163, Genoa, Italy.
[4] Department of Physics, Sapienza University, Piazzale Aldo Moro 5, 00185, Rome, Italy
[5] Center for Life Nanoscience, Istituto Italiano di Tecnologia, Viale Regina Elena 291, 00161, Rome, Italy
[4] Department of Biology 'Charles Darwin', Sapienza University of Rome, P.le A. Moro 5, Rome 00185, Italy
[6] School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore, 637551, Singapore
[7] Institucio Catalana de Recerca i Estudis Avançats (ICREA), 23 Passeig Lluis Companys, 08010 Barcelona, Spain

*to whom correspondence should be addressed to: riccardo.ponti@ntu.edu.sg (RDP) and giangaetano.tartaglia@uniroma1.it or gian.tartaglia@iit.it (GGT)

## ABSTRACT

We calculated the structural properties of >2500 coronaviruses and computed >100000 human protein interactions with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Using the *CROSS* method, we found that the SARS-CoV-2 region encompassing nucleotides 23000 - 24000 is highly conserved at the structural level, while the region upstream varies significantly. These two sequences are important for viral infection as they code for a domain of the viral protein Spike S interacting with the human receptor angiotensin-converting enzyme 2 (ACE2) and, in the close homologue from Middle East respiratory syndrome coronavirus (MERS-CoV), sialic acids. We predict highly structured regions at the 5' and 3' where our calculations indicate strong propensity to bind to human proteins involved in viral replication. Using the *cat*RAPID method, we identified that the 5' interacts with double-stranded RNA-specific editase 1 ADARB1, 2-5Adependent ribonuclease RNASEL, ATP-dependent RNA helicase DDX1 and A-kinase anchor protein 8-like AKAP8L, in addition to >10 high-confidence candidate partners. These interactions, also implicated in HIV replication, should be further investigated for a better understanding of host-virus interaction mechanisms.

## INTRODUCTION

A novel disease named Covid-19 by the World Health Organization and caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has been recognized as responsible for the pneumonia outbreak that started in December, 2019 in Wuhan City, Hubei, China [1] and spread in February to Milan, Lombardy, Italy [2] becoming pandemic. As of April 2020, the virus infected >1’000’000 people in more than 200 countries.

SARS-CoV-2 is a positive-sense single-stranded RNA virus that shares similarities with other beta-coronavirus such as severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV) 3. Bats have been identified as the primary host for SARS-CoV and SARS-CoV-2 4,5 but the intermediate host linking SARS-CoV-2 to humans is still unknown, although a recent report indicates that pangolins could be involved 6.

Coronaviruses use species-specific regions to mediate the entry in the host cell and SARS-CoV, MERS-CoV and SARS-CoV-2, the spike S protein activates the infection in human respiratory epithelial cells [7]. Spike S is assembled as a trimer and contains around 1,300 amino acids within each unit [8]. In the S’ region of the protein, the receptor binding domain (RBD), which contains around 300 amino acids, mediates the binding with angiotensin-converting enzyme, (ACE2) attacking respiratory cells. Another region upstream of the RBD, present in MERS-CoV but not in SARS-CoV, is involved in the adhesion to sialic acid and could play a key role in regulating viral infection [7,9].

At present, very few molecular details are available on SARS-CoV-2 and its interactions with human host, which are mediated by specific RNA elements [10]. To study the RNA structural content, we used *CROSS* [11] that was previously developed to investigate large transcripts such as the human immunodeficiency virus HIV-1 [12]. *CROSS* predicts the structural profile (single- and double-stranded state) at single-nucleotide resolution using sequence information only. We performed sequence and structural alignments among 62 SARS-CoV-2 strains and identified the conservation of specific elements in the spike S region, which provide clues on the evolution of domains involved in the binding to ACE2 and sialic acid.

As highly structured regions of RNA molecules have strong propensity to form stable contacts with proteins [13] and promote assembly of specific complexes [14,15], SARS-CoV-2 domains enriched in double-stranded content are expected to establish interactions within host cells that are important to replicate the virus. To investigate the interactions potential of SARS-CoV-2 RNA with human proteins, we employed *cat*RAPID [16,17]. *cat*RAPID [18] estimates the binding potential of protein and RNA molecules through van der Waals, hydrogen bonding and secondary structure propensities of allowing identification of interaction partners with high confidence [19]. The unbiased analysis of more than 100000 protein interactions with SARS-CoV-2 RNA reveals that the 5' of SARS-CoV-2 has strong propensity to bind to human proteins involved in viral infection and especially reported to be associated with HIV infection. A comparison between SARS-CoV and HIV reveals indeed similarities [20], but the relationship between SARS-CoV-2 and HIV is still unexplored. Interestingly, HIV and SARS-CoV-2, but not SARS-CoV nor MERS-CoV, have a furin-cleavage site occurs in the spike S protein, which could explain the spread velocity of SARS-CoV-2 compared to SARS-CoV and MERS-CoV [21,22]. Yet, many processes related to SARS-CoV-2 replication are unknown and our study aims to suggest relevant protein interactions for further investigation.

We hope that our large-scale calculations of structural properties and binding partners of SARS-CoV-2 will be useful to identify the mechanisms of virus replication within the human host.

## RESULTS

### SARS-CoV-2 contains highly structured elements

Structural elements within RNA molecules attract proteins [13] and reveal regions important for interactions with the host [23].

To analyze SARS-CoV-2 (reference Wuhan strain MN908947), we employed *CROSS* [11] that predicts the double- and single-stranded content of large transcripts such as *Xist* and HIV-1 [12]. We found the highest density of double-stranded regions in the 5' (nucleotides 1-253), membrane M protein (nucleotides 26523-27191), spike S protein (nucleotides 23000-24000), and nucleocapsid N protein (nucleotides 2874-29533; **Fig. 1**) [24]. The lowest density of double-stranded regions were observed at nucleotides 6000-6250 and 20000-21500 and correspond to the regions between the non-structural proteins nsp14 and nsp15 and the upstream region of the spike surface protein S (**Fig. 1**) [24]. In addition to the maximum corresponding to nucleotides 23000-24000, the structural content of spike S protein shows minima at around nucleotides 20500 and 24500 (**Fig. 1**). We used the *Vienna* method [25] to further investigate the RNA secondary structure of specific regions identified with *CROSS* [12]. Employing a 100 nucleotide window centered around *CROSS* maxima and minima, we found good match between *CROSS* scores and Vienna free energies (**Fig. 1).** Strong agreement is also observed between *CROSS* and *Vienna* positional entropy, indicating that regions with the highest structural content have also the lowest structural diversity.

Our analysis suggests presence of structural elements in SARS-CoV-2 that have evolved to interact with specific human proteins [10]. Our observation is based on the assumption that structured regions have an intrinsic propensity to recruit proteins [13], which is supported by the fact that structured transcripts act as scaffolds for protein assembly [14,15].

## Structural comparisons reveal that the spike S region of SARS-CoV-2 is conserved among coronaviruses

We employed *CROSS*align [12] to study the structural conservation of SARS-CoV-2 in different strains (**Materials and Methods**).

In this analysis, we compared the Wuhan strain MN908947 with around 2800 other coronaviruses (data from NCBI) having as host human (**Fig. 2**) or other species (**Supp. Fig. 1**). When comparing SARS-CoV-2 with human coronaviruses (1387 strains, including SARS-CoV and MERS-CoV), we found that the most conserved region falls inside the spike S genomic locus (**Fig. 2**). More precisely, the conserved region is between nucleotides 23000 - 24000 and exhibits an intricate and stable secondary structure (RNA*fold* minimum free energy= -269 kcal/mol )[25]. High conservation of a structured regions suggests a functional activity that might be relevant for host infection.

While the 3’ and 5’ of SARS-CoV-2 were shown to be relatively conserved in some beta-coronavirus [10], they are highly variable in the entire set. However, the 3’ and 5’ are more structured in SARS-CoV-2 than other coronaviruses (average structural content for SARS-CoV-2 = 0.56 in the 5' and 0.49 in the 3'; other coronaviruses 0.49 in the 5' and 0.42 in the 3').

### Sequence and structural comparisons among SARS-CoV-2 strains indicate conservation of the ACE2 binding site and high variability in a region potentially interacting with sialic acids.

To better investigate the sequence conservation of SARS-CoV-2, we compared 62 strains isolated form different countries during the pandemic (including China, USA, Japan, Taiwan, India, Brazil, Sweden, and Australia; data from NCBI and in VIPR www.viprbrc.org; **Materials and Methods**). Our analysis aims to determine the relationship between structural content and sequence conservation.

Using *Clustal W* for multiple sequence alignments [26], we observed general conservation of the coding regions with several *minima* in correspondence to areas between genes (**Fig. 3A**). One highly conserved region is between nucleotides 23000 - 24000 in the spike S genomic locus, while sequences up- and down-stream are variable (**Fig. 3A**). We then used CROSS*align* [12] to compare the structural content (**Materials and Methods**). High variability of structure is observed for both the 5' and 3' and for nucleotides between 21000 - 22000 as well as 24000 - 25000, associated with the S region (red bars in **Fig. 3A**). The rest of the regions are significantly conserved at a structural level (p-value < 0.0001; Fisher's test).

We then compared  protein sequences coded by the spike S genomic locus (NCBI reference QHD43416) and found that both sequence (**Fig. 3A**) and structure (**Fig. 2**) of nucleotides 23000 - 24000 are highly conserved. The region corresponds to amino acids 330-500 that contact the host receptor angiotensin-converting enzyme 2 (ACE2) [27] promoting infection and provoking lung injury [22,28]. By contrast, the region upstream of the binding site receptor ACE2 and located in correspondence to the minimum of the structural profile at around nucleotides 22500-23000 (**Fig. 1**) is highly variable [29], as calculated with *Tcoffee* multiple sequence alignments [29]  (**Fig. 3A**). This part of the spike S region corresponds to amino acids 243-302 that in MERS-CoV binds to sialic acids regulating infection through cell-cell membrane fusion (**Fig. 3B;** see related manuscript by E. Milanetti *et al.* "In-Silico evidence for two receptors based strategy of SARS-CoV-2") [9,30,31].

Our analysis suggests that the structural region between nucleotides 23000 and 24000 of Spike S region is conserved among coronaviruses (**Fig. 2**) and the binding site for ACE2 has poor variation in human SARS-CoV-2 strains (**Fig. 3B**). By contrast, the region upstream, potentially involved in adhesion to sialic acids, has almost poor structural content and varies significantly in the human population (**Fig. 3B**).

## Analysis of human interactions with SARS-CoV-2 identifies proteins involved in viral replication and HIV infection

In order to obtain insights on how the virus is replicated in human cells, we predicted SARS-CoV-2 interactions with the whole RNA-binding human proteome. Following a protocol to study structural conservation in viruses [12], we first divided the Wuhan sequence in 30 fragments of 1000 nucleotides each moving from the 5' to 3' and then calculated the protein-RNA interactions of each fragment with *cat*RAPID *omics* (3340 canonical and non-canonical RNA-binding proteins, or RBPs, for a total 102000 interactions) [16]. Proteins such as PTBP1 showed the highest interaction propensity (or Z-score; **Materials and Methods**) at the 5' while others such as HNRNPQ showed the highest interaction propensity at the 3', in agreement with previous studies on coronaviruses [32].

For each fragment, we predicted the most significant interactions by filtering according to the Z score. We used three different thresholds in ascending order of stringency: $Z \geq 1.50$, 1.75 and 2 respectively. Importantly, we removed from the list proteins that were predicted to interact promiscuously with different fragments. Fragment 1 corresponds to the 5' and is the most contacted by RBPs (around 120 with $Z \geq 2$ high-confidence interactions; **Fig. 4A**), which is in agreement with the observation that highly structured regions attract a large number of proteins [13]. Indeed, the 5' contains a leader sequence and the untranslated region with multiple stem loop structures that control RNA replication and transcription [33,34].

The interactome of each fragment was then analysed using *clever*GO, a tool for GO enrichment analysis [35]. Proteins interacting with fragments 1, 2 and 29 were associated with annotations related to viral processes (**Fig. 4B; Supp. Table 1**). Considering the three thresholds applied (**Materials and Methods**), we found 22 viral proteins for fragment 1, 2 proteins for fragment 2 and 11 proteins for fragment 29 (**Fig. 4C**).

Among the high-confidence interactors of fragment 1, we discovered RBPs involved in positive regulation of viral processes and viral genome replication, such as double-stranded RNA-specific editase 1 ADARB1 (Uniprot P78563 [36]) and 2-5A-dependent ribonuclease RNASEL (Q05823). We also identified proteins related to the establishment of integrated proviral latency, including X-ray repair cross-complementing protein 5 XRCC5 (P13010) and X-ray repair cross-complementing protein 6 XRCC6 (P12956; **Fig. 4D**).

Importantly, we found proteins related to defence response to viruses, such as ATP-dependent RNA helicase DDX1 (Q92499), are involved in the negative regulation of viral genome replication. Some proteins are listed as DNA binding proteins such as Cyclin-T1 CCNT1 (Uniprot code O60563 [36]), Zinc finger protein 175 ZNF175 (Q9Y473), while Prospero homeobox protein 1 PROX1 (Q92786) were included because they could have potential RNA-binding ability (**Fig. 4D**) [37]. As for fragment 2, we found two canonical RBPs: E3 ubiquitin-protein ligase TRIM32 (Q13049) and E3 ubiquitin-protein ligase TRIM21 (P19474), which are listed as negative regulators of viral release from host cell, negative regulators of viral transcription and positive regulators of viral entry into host cells. Finally, for fragment 29, 10 of the 11 viral proteins found are members of the *Gag polyprotein family*, that perform different tasks during HIV assembly, budding, maturation. More than a simple scaffold protein forming the viral core, Gag proteins are recognized as elements able to select viral and host proteins as they traffic to the cell membrane (**Supp. Table 1**) [38].

Analysis of functional annotations carried out with *GeneMania* [39] reveals that proteins interacting with the 5' of SARS-CoV-2 RNA are associated with regulatory pathways involving NOTCH2, MYC and MAX that have been previously connected to viral infection processes (**Fig. 4B**) [40,41]. Interestingly, some of the proteins, including DDX1, ZNF175 and CCNT1 for fragment 1 and TRIM32 for fragment 2, are reported to be necessary for HIV functions and replication inside the cell. The roles of these proteins in the replication of a retrovirus such as HIV are expected to be different from those associated with SARS-CoV-2, yet it has been reported that SARS-CoV-2 represses host gene expression through a number of unknown mechanisms, which could also involve sequestration of transcriptional elements such as Cyclin-T1 CCNT1 [42]. DDX1 is required for HIV-1 Rev function as well as for HIV-1 and coronavirus IBV replication and it binds to the RRE sequence of HIV-1 RNAs [43,44]. ZNF175 is relatively uncharacterized reported to interfere with HIV-1 replication by suppressing Tat-induced viral LTR promoter activity [45]. Finally, CCNT1 is 7SK snRNA binding and regulates transactivation domain of the viral nuclear transcriptional activator, Tat [46,47]. In addition, TRIM32 (fragment 2) is a well-defined Tat binding protein and, more specifically, it binds to the activation domain of HIV-1 Tat and can also interact with the HIV-2 and EIAV Tat proteins *in vivo* [48].

**Analysis of interactions with SARS-CoV-2 Open Reading Frames identifies additional interactions involved in HIV infection**

Recently, Gordon *et al*. reported a list of human proteins binding to Open Reading Frames (ORFs) translated from SARS-CoV-2 [49]. Identified through affinity purification followed by mass spectrometry quantification, 332 proteins from HEK-293T cells interact with viral ORF peptides. By selecting 274 proteins binding at the 5' with Z score ≥1.5 (**Supp. Table 1**), of which 140 are exclusively interacting with fragment 1 (**Fig. 4B**), we found that 8 are also reported in the list by Gordon *et al*. [49], which indicates significant enrichment (representation factor of 2.5; p-value of 0.02; hypergeometric test with human proteome in background). The fact that our list of protein-RNA binding partners contains elements identified also in the protein-protein network analysis is not surprising, as ribonucleoprotein complexes evolve together [13] and their components sustain each other activities through different types of interactions [15].

We note that out of 332 interactions, 60 are RBPs (as reported in Uniprot [36]), which represents a considerable fraction (20%), considering that there are around 1500 RBPs in the human proteome (6%) and fully justified by the fact that they involve association with viral RNAs. Comparing the RBPs present in Gordon *et al.* [49] and those present in our list (79 as reported in Uniprot), we found an overlap of 6 proteins (representation factor = 26.5; p-value < $10^{-8}$; hypergeometric test), including: Janus kinase and microtubule-interacting protein 1 JAKMIP1 (Q96N16), A-kinase anchor protein 8 AKAP8 (O43823) and A-kinase anchor protein 8-like AKAP8L (Q9ULX6), which in case of HIV-1 infection is involved as a DEAD/H-box RNA helicase binding [50], signal recognition particle subunit SRP72 (O76094), binding to the 7S RNA in presence of SRP68, La-related protein 7, LARP7 (Q4G0J3) and La-related protein 4B LARP4B (Q92615), which are part of a system for transcriptional regulation acting by means of the 7SK RNP system [51] (**Fig. 4E; Supp. Table 2**). We speculate that sequestration of elements binding to the 7S RNA is part of a viral program aiming to host genes [42].

Moreover, by analysing the RNA interaction potential of all the 332 proteins by Gordon *et al.* [49], *cat*RAPID identified 38 putative binders at the 5' (Z score ≥ 1.5; 27 occurring exclusively in the 5' and not in other regions of the RNA) [16], including Serine/threonine-protein kinase TBK1

(Q9UHD2), among which 10 RBPs (as reported in Uniprot) such as: Splicing elements U3 small nucleolar ribonucleoprotein protein MPP10 (O00566) and Pre-mRNA-splicing factor SLU7 (O95391), snRNA methylphosphate capping enzyme MEPCE involved in negative regulation of transcription by RNA polymerase II 7SK (Q7L2J0) [52], Nucleolar protein 10 NOL10 (Q9BSC4) and protein kinase A Radixin RDX (P35241; in addition to those mentioned above; **Supp. Table 2**).

## HIV-related RBPs are significantly enriched in the 5’ interactions

In the list of 274 proteins predicted to bind at the 5’ (fragment 1) with Z score ≥1.5, we found 10 hits reported to be involved in HIV (**Supp. Table 3**), which is a highly significant enrichment (p-value=0.0004; Fisher’s exact test), considering that the total number of HIV-related proteins is 35 in the whole *cat*RAPID library (3340 elements). The complete list of proteins includes ATP-dependent RNA helicase DDX1 (Q92499 also involved in Coronavirus replication [43,44]), ATP-dependent RNA helicase DDX3X (O00571 also involved in Dengue and Zika Viruses), Tyrosine-protein kinase HCK (P08631, nucleotide binding), Arf-GAP domain and FG repeat-containing protein 1 (P52594), Double-stranded RNA-specific editase 1 ADARB1 (P78563), Insulin-like growth factor 2 mRNA-binding protein 1 IGF2BP1 (Q9NZI8), A-kinase anchor protein 8-like AKAP8L (Q9ULX6; its partner AKAP8 is also found in Gordon *et al.* [49]) Cyclin-T1 CCNT1 (O60563; DNA-binding) and Forkhead box protein K2 FOXK2 (Q01167; DNA-binding; **Supp. Table 3**).

Smaller enrichments were found for proteins related to Hepatitis B virus (HBV; p-value=0.01; 3 hits out of 7 in the whole *cat*RAPID library; Fisher’s exact test), Nuclear receptor subfamily 5 group A member 2 NR5A2 (DNA-binding; O00482), Interferon-induced, double-stranded RNA-activated protein kinase EIF2AK2 (P19525), and SRSF protein kinase 1 SRPK1 (Q96SB4) as well as Influenza (p-value=0.03; 2 hits out of 4; Fisher’s exact test), Synaptic functional regulator FMR1 (Q06787) and RNA polymerase-associated protein RTF1 homologue (Q92541; **Supp. Table 3**). By contrast, no significant enrichments were found for other viruses such as for instance Ebola.

Interestingly, specific drugs are reported in ChEMBL[53] for HIV-related proteins ATP-dependent RNA helicase DDX1 (CHEMBL2011807), ATP-dependent RNA helicase DDX3X (CHEMBL2011808), Cyclin-T1 CCNT1 (CHEMBL2348842), and Tyrosine-protein kinase HCK (CHEMBL2408778)[53], as well as HVB-related proteins Nuclear receptor subfamily 5 group A member 2 NR5A2 (CHEMBL3544), Interferon-induced, double-stranded RNA-activated protein kinase EIF2AK2 (CHEMBL5785) and SRSF protein kinase 1 SRPK1 (CHEMBL4375), which could be a starting point for further investigations.

## CONCLUSIONS

Our study is motivated by the need to identify molecular mechanisms involved in Covid-19 spreading. Using advanced computational approaches, we investigated the structural content of SARS-CoV-2 RNA and predicted human proteins that bind it.

We employed *CROSS* [12,54] to compare the structural properties of 2800 coronaviruses and identified elements conserved in SARS-CoV-2 strains. The regions containing the highest amount of structure are the 5' as well as glycoproteins spike S and membrane M.

We found that the spike S protein domain encompassing amino acids 330-500 is highly conserved across SARS-CoV-2 strains. This result suggests that spike S must have evolved to specifically interact with its host partner ACE2 [27] and mutations increasing the binding affinity are highly infrequent. As the nucleic acids encoding for this region are enriched in double-stranded content, we speculate that the structure might attract host regulatory elements, which further constrains its variability. The fact that the ACE2 receptor binding site is conserved among the SARS-CoV-2 strains suggests that a specific drug can be designed to prevent host interaction and thus infection, which could work for a large number of coronaviruses.

By contrast, the highly variable region at amino acids 243-302 in spike S protein corresponds to the binding site of sialic acids in MERS-CoV (see manuscript by E. Milanetti *et al*. “In-Silico evidence for two receptors based strategy of SARS-CoV-2” ) [7,9,31] and could play a role in infection [30]. The fact that the binding region changes in the different strains might indicate a variety of binding affinities for sialic acids, which could provide clues on the specific responses in the human population. Interestingly, the sialic acid binding site is absent in SARS-CoV but present in MERS-CoV, which indicates that it must have evolved recently.

Both our sequence and structural analyses of spike S protein indicate that human engineering of SARS-CoV-2 is highly unlikely.

Using *cat*RAPID [16,17] we computed >100000 protein interactions with SARS-CoV-2 and found that the highly structured region at the 5' has the largest number of protein partners including ATP-dependent RNA helicase DDX1,  which has been previously reported to be essential for HIV-1 and coronavirus IBV replication [43,44], double-stranded RNA-specific editase 1 ADARB1, which catalyses the hydrolytic deamination of adenosine to inosine and  might take part in the chemical modification of SARS-CoV-2 RNA [42] . Other relevant interactions are XRCC5 and XRCC6 members of the HDP-RNP complex interacting with ATP-dependent RNA helicase  DHX9 [55] and 2-5A-dependent ribonuclease RNASEL that has antiviral effects through a combination of cleavage of single-stranded viral RNAs, inhibition of protein synthesis, induction of apoptosis, and induction of antiviral genes [56].

A significant overlap exists with the list of protein interactions reported by Gordon *et al.* [49], and among the candidate partners we identified AKAP8L, involved as a DEAD/H-box RNA helicase binding in HIV infection [50].  In general, proteins involved in the replication of retroviruses such as HIV are expected to play a different role in mechanisms related to SARS-CoV-2 that uses its own RNA-dependent RNA polymerase, yet it must be considered that  SARS-CoV-2 represses host gene expression through a number of  unknown mechanisms, which could imply sequestration of transcriptional components, such as specific polymerase II genes and splicing factors [42]. Thus, the link to HIV and other viruses such as HBV and Influenza could be key to identify targets for the repurposing of drugs for treatment of SARS-CoV-2 infection [53] .

In conclusion, we hope that our analysis would be useful to the scientific community to identify virus-host interactions and block SARS-CoV-2 spreading.

## Acknowledgements

The authors would like to thank Jakob Rupert, Dr. Mattia Miotto, Dr Lorenzo Di Rienzo, Dr. Alexandros Armaos, Dr. Alessandro Dasti, Dr. Elias Bechara, Dr. Claudia Giambartolomei and Dr. Elsa Zacco for discussions.

The research leading to these results has been supported by European Research Council (RIBOMYLOME_309545 to GGT, ASTRA_855923 to GGT), Spanish Ministry of Economy and Competitiveness BFU2017-86970-P, the H2020 projects INFORE_825080 and IASIS_727658, as well as the collaboration with Peter St. George-Hyslop financed by the Wellcome Trust.

**Contributions.** GGT and RDP conceived the study. AV carried out *cat*RAPID analysis of protein interactions, RDP calculated *CROSS* structures of coronaviruses, GGT, MM and EM performed and analysed sequence alignments. AV, RDP and GGT wrote the paper.

## MATERIALS AND METHODS

### *Structure prediction*

We predicted the secondary structure of transcripts using *CROSS* (Computational Recognition of Secondary Structure [12,54]. *CROSS* was developed to perform high-throughput RNA profiling. The algorithm predicts the structural profile (single- and double-stranded state) at single-nucleotide resolution using sequence information only and without sequence length restrictions (scores > 0 indicate double stranded regions). We used the *Vienna* method [25] to further investigate the RNA secondary structure of minima and maxima identified with *CROSS* [12].

### *Structural conservation*

We used *CROSS*align [12,54] an algorithm based on Dynamic Time Warping (DTW), to check and evaluate the structural conservation between different viral genomes [12]. CROSS*align* was previously employed to study the structural conservation of ~5000 HIV genomes. SARS-CoV-2 fragments (1000 nt, not overlapping) were searched inside other complete genomes using the OBE (open begin and end) module, in order to search a small profile inside a larger one. The lower the structural distance, the higher the structural similarities (with a minimum of 0 for almost identical secondary structure profiles). The significance is assessed as in the original publication [12].

### *Sequence collection*

The FASTA sequences of the complete genomes of SARS-CoV-2  were downloaded from Virus Pathogen Resource (VIPR; www.viprbrc.org), for a total of 62 strains. Regarding the overall coronaviruses, the sequences were downloaded from NCBI selecting only complete genomes, for a total of 2862 genomes. The reference Wuhan sequence with available annotation (EPI_ISL_402119) was downloaded from Global Initiative on Sharing All Influenza Data. (GISAID https://www.gisaid.org/).

### *Protein-RNA interaction prediction*

Interactions between each fragment of target sequence and the human proteome were predicted using *cat*RAPID *omics* [16,17] , an algorithm that estimates the binding propensity of protein-RNA pairs by combining secondary structure, hydrogen bonding and van der Waals contributions. As reported in a recent analysis of about half a million of experimentally validated interactions [31], the algorithm is able to separate interacting vs non-interacting pairs with an area under the ROC curve of 0.78.

The complete list of interactions between the 30 fragments and the human proteome is available at http://crg-webservice.s3.amazonaws.com/submissions/2020-03/252523/output/index.html?unlock=f6ca306af0. The output then is filtered according to the Z-score column, which is the interaction propensity normalised by the mean and standard deviation calculated over the reference RBP set (http://s.tartaglialab.com/static_files/shared/faqs.html#4). We used three different thresholds in ascending order of stringency: Z greater or equal than 1.50, 1.75 and 2 respectively and for each threshold we then selected the proteins that were unique for each fragment for each threshold.

### *GO terms analysis*

*clever*GO [35], an algorithm for the analysis of Gene Ontology annotations, was used to determine which fragments present enrichment in GO terms related to viral processes. Analysis of functional annotations was performed in parallel with *GeneMania* [39].

### *RNA and protein alignments*

We sued *Clustal W* [26] for 62 SARS-CoV-2 strains alignments and *Tcoffee* [29] for spike S proteins alignments. The variability in the spike S region was measured by computing Shannon entropy on translated RNA sequences. The Shannon entropy is computed as follows:

S(a) = - Sum_i P(a,i) log P(a,i)

Where *a* correspond to the amino acid at the position *i* and P(a,i) is the frequency of a certain amino-acid *a* at position *i* of the sequence. Low entropy indicates poorly variability: if P(a,x) = 1 for one *a* and 0 for the rest, then S(x) =0. By contrast, if the frequencies of all amino acids are equally distributed, the entropy reaches its maximum possible value.

1. Zhu, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* **382**, 727–733 (2020).
2. D'Antiga, L. Coronaviruses and immunosuppressed patients. The facts during the third epidemic. *Liver Transplant. Off. Publ. Am. Assoc. Study Liver Dis. Int. Liver Transplant. Soc.* (2020) doi:10.1002/lt.25756.
3. Cascella, M., Rajnik, M., Cuomo, A., Dulebohn, S. C. & Di Napoli, R. Features, Evaluation and Treatment Coronavirus (COVID-19). in *StatPearls* (StatPearls Publishing, 2020).
4. Ge, X.-Y. *et al.* Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503**, 535–538 (2013).
5. Follis, K. E., York, J. & Nunberg, J. H. Furin cleavage of the SARS coronavirus spike glycoprotein enhances cell-cell fusion but does not affect virion entry. *Virology* **350**, 358–369 (2006).
6. Xiao, K. *et al.* Isolation and Characterization of 2019-nCoV-like Coronavirus from Malayan Pangolins. *bioRxiv* 2020.02.17.951335 (2020) doi:10.1101/2020.02.17.951335.
7. Park, Y.-J. *et al.* Structures of MERS-CoV spike glycoprotein in complex with sialoside attachment receptors. *Nat. Struct. Mol. Biol.* **26**, 1151–1157 (2019).
8. Walls, A. C. *et al.* Cryo-electron microscopy structure of a coronavirus spike glycoprotein trimer. *Nature* **531**, 114–117 (2016).
9. Li, W. *et al.* Identification of sialic acid-binding function for the Middle East respiratory syndrome coronavirus spike glycoprotein. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E8508–E8517 (2017).
10. Yang, D. & Leibowitz, J. L. The Structure and Functions of Coronavirus Genomic 3' and 5' Ends. *Virus Res.* **206**, 120–133 (2015).
11. Delli Ponti, R., Marti, S., Armaos, A. & Tartaglia, G. G. A high-throughput approach to profile RNA structure. *Nucleic Acids Res.* **45**, e35–e35 (2017).

12. Delli Ponti, R., Armaos, A., Marti, S. & Gian Gaetano Tartaglia. A Method for RNA Structure Prediction Shows Evidence for Structure in lncRNAs. *Front. Mol. Biosci.* **5**, 111 (2018).
13. Sanchez de Groot, N. *et al.* RNA structure drives interaction with proteins. *Nat. Commun.* **10**, 3246 (2019).
14. Cid-Samper, F. *et al.* An Integrative Study of Protein-RNA Condensates Identifies Scaffolding RNAs and Reveals Players in Fragile X-Associated Tremor/Ataxia Syndrome. *Cell Rep.* **25**, 3422-3434.e7 (2018).
15. Cerase, A. *et al.* Phase separation drives X-chromosome inactivation: a hypothesis. *Nat. Struct. Mol. Biol.* **26**, 331 (2019).
16. Agostini, F. *et al.* catRAPID omics: a web server for large-scale prediction of protein-RNA interactions. *Bioinforma. Oxf. Engl.* **29**, 2928–2930 (2013).
17. Cirillo, D. *et al.* Quantitative predictions of protein interactions with long noncoding RNAs. *Nat. Methods* **14**, 5–6 (2017).
18. Bellucci, M., Agostini, F., Masin, M. & Tartaglia, G. G. Predicting protein associations with long noncoding RNAs. *Nat. Methods* **8**, 444–445 (2011).
19. Lang, B., Armaos, A. & Tartaglia, G. G. RNAct: Protein–RNA interaction predictions for model organisms with supporting experimental data. *Nucleic Acids Res.* doi:10.1093/nar/gky967.
20. Kliger, Y. & Levanon, E. Y. Cloaked similarity between HIV-1 and SARS-CoV suggests an anti-SARS strategy. *BMC Microbiol.* **3**, 20 (2003).
21. Hallenberger, S. *et al.* Inhibition of furin-mediated cleavage activation of HIV-1 glycoprotein gp160. *Nature* **360**, 358–361 (1992).
22. Glowacka, I. *et al.* Differential downregulation of ACE2 by the spike proteins of severe acute respiratory syndrome coronavirus and human coronavirus NL63. *J. Virol.* **84**, 1198–1205 (2010).

23. Gultyaev, A. P., Richard, M., Spronken, M. I., Olsthoorn, R. C. L. & Fouchier, R. A. M. Conserved structural RNA domains in regions coding for cleavage site motifs in hemagglutinin genes of influenza viruses. *Virus Evol.* **5**, (2019).

24. Wu, A. *et al.* Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. *Cell Host Microbe* **27**, 325–328 (2020).

25. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).

26. Madeira, F. *et al.* The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* **47**, W636–W641 (2019).

27. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nat. Med.* 1–3 (2020) doi:10.1038/s41591-020-0820-9.

28. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).

29. Di Tommaso, P. *et al.* T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res.* **39**, W13–W17 (2011).

30. Qing, E., Hantak, M., Perlman, S. & Gallagher, T. Distinct Roles for Sialoside and Protein Receptors in Coronavirus Infection. *mBio* **11**, (2020).

31. Milanetti, E. *et al.* In-Silico evidence for two receptors based strategy of SARS-CoV-2. *ArXiv200311107 Phys. Q-Bio* (2020).

32. Galán, C. *et al.* Host cell proteins interacting with the 3’ end of TGEV coronavirus genome influence virus replication. *Virology* **391**, 304–314 (2009).

33. Lu, K., Heng, X. & Summers, M. F. Structural determinants and mechanism of HIV-1 genome packaging. *J. Mol. Biol.* **410**, 609–633 (2011).

34. Fehr, A. R. & Perlman, S. Coronaviruses: An Overview of Their Replication and Pathogenesis. *Methods Mol. Biol. Clifton NJ* **1282**, 1–23 (2015).

35. Klus, P., Ponti, R. D., Livi, C. M. & Tartaglia, G. G. Protein aggregation, structural disorder and RNA-binding ability: a new approach for physico-chemical and gene ontology classification of multiple datasets. *BMC Genomics* **16**, 1071 (2015).

36. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

37. Castello, A. *et al.* Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* **149**, 1393–1406 (2012).

38. Bell, N. M. & Lever, A. M. L. HIV Gag polyprotein: processing and early viral particle assembly. *Trends Microbiol.* **21**, 136–144 (2013).

39. Warde-Farley, D. *et al.* The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* **38**, W214–W220 (2010).

40. Hayward, S. D. Viral interactions with the Notch pathway. *Semin. Cancer Biol.* **14**, 387–396 (2004).

41. Dudley, J. P., Mertz, J. A., Rajan, L., Lozano, M. & Broussard, D. R. What retroviruses teach us about the involvement of c- Myc in leukemias and lymphomas. *Leukemia* **16**, 1086–1098 (2002).

42. Kim, D. *et al.* The architecture of SARS-CoV-2 transcriptome. *bioRxiv* 2020.03.12.988865 (2020) doi:10.1101/2020.03.12.988865.

43. Fang, J. *et al.* A DEAD box protein facilitates HIV-1 replication as a cellular co-factor of Rev. *Virology* **330**, 471–480 (2004).

44. Xu, L. *et al.* The cellular RNA helicase DDX1 interacts with coronavirus nonstructural protein 14 and enhances viral replication. *J. Virol.* **84**, 8571–8583 (2010).

45. Carlson, K. A. *et al.* Molecular characterization of a putative antiretroviral transcriptional factor, OTK18. *J. Immunol. Baltim. Md 1950* **172**, 381–391 (2004).

46. Ivanov, D. *et al.* Cyclin T1 domains involved in complex formation with Tat and TAR RNA are critical for tat-activation. *J. Mol. Biol.* **288**, 41–56 (1999).

47. Kwak, Y. T., Ivanov, D., Guo, J., Nee, E. & Gaynor, R. B. Role of the human and murine cyclin T proteins in regulating HIV-1 tat-activation. *J. Mol. Biol.* **288**, 57–69 (1999).
48. Locke, M., Tinsley, C. L., Benson, M. A. & Blake, D. J. TRIM32 is an E3 ubiquitin ligase for dysbindin. *Hum. Mol. Genet.* **18**, 2344–2358 (2009).
49. Gordon, D. E. *et al.* A SARS-CoV-2-Human Protein-Protein Interaction Map Reveals Drug Targets and Potential Drug-Repurposing. *bioRxiv* 2020.03.22.002386 (2020) doi:10.1101/2020.03.22.002386.
50. Xing, L., Zhao, X., Guo, F. & Kleiman, L. The role of A-kinase anchoring protein 95-like protein in annealing of tRNALys3 to HIV-1 RNA. *Retrovirology* **11**, 58 (2014).
51. Markert, A. *et al.* The La-related protein LARP7 is a component of the 7SK ribonucleoprotein and affects transcription of cellular and viral polymerase II genes. *EMBO Rep.* **9**, 569–575 (2008).
52. C, J. *et al.* Systematic Analysis of the Protein Interaction Network for the Human Transcription Machinery Reveals the Identity of the 7SK Capping Enzyme. *Molecular cell* vol. 27 https://pubmed.ncbi.nlm.nih.gov/17643375/ (2007).
53. Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940 (2019).
54. Delli Ponti, R., Marti, S., Armaos, A. & Tartaglia, G. G. A high-throughput approach to profile RNA structure. *Nucleic Acids Res.* **45**, e35–e35 (2017).
55. Zhang, S., Schlott, B., Görlach, M. & Grosse, F. DNA-dependent protein kinase (DNA-PK) phosphorylates nuclear DNA helicase II/RNA helicase A and hnRNP proteins in an RNA-dependent manner. *Nucleic Acids Res.* **32**, 1–10 (2004).
56. Siddiqui, M. A., Mukherjee, S., Manivannan, P. & Malathi, K. RNase L Cleavage Products Promote Switch from Autophagy to Apoptosis by Caspase-Mediated Cleavage of Beclin-1. *Int. J. Mol. Sci.* **16**, 17611–17636 (2015).

## FIGURES LEGENDS

***Fig. 1.*** *Using the CROSS approach* [12,54]*, we studied the structural content of SARS-CoV-2. We found the highest density of double-stranded regions in the 5' (nucleotides 1-253), membrane M protein (nucleotides 26523-27191), and the spike S protein (nucleotides 23000-24000). Strong match is observed between CROSS and Vienna analyses (centroid structures shown, indicating that regions with the highest structural content have the lowest free energies.*

***Fig. 2.*** *We employed the CROSSalign approach* [12,54] *to compare the Wuhan strain MN908947 with other coronaviruses (1387 strains, including SARS-CoV and MERS-CoV) indicates that the most conserved region falls inside the spike S genomic locus. The inset shows thermodynamic structural variability (positional entropy) within regions encompassing nucleotides 23000-24000 along with the centroid structure and free energy.*

***Fig. 3.*** *Sequence and structural comparison of human SARS-CoV-2 strains. (**A**) Strong sequence conservation (Clustal W multiple sequence alignments* [35]*) is observed in coding regions, including the region between nucleotides 23000 and 24000 of spike S protein. High structural variability (red bars on top) is observed for both the UTRs and for nucleotides between 21000 and 22000 as well as 24000 and 25000, associated with the S region. The rest of the regions are significantly conserved at a structural level. (**B**) The sequence variability (Shannon entropy computed on Tcoffee multiple sequence alignments* [29]*) in the spike S protein indicate conservation between amino-acids 460 and 520 (blue box) binding to the host receptor angiotensin-converting enzyme 2 ACE2. The region encompassing amino-acids 243 and 302 is highly variable and is implicated in sialic acids in MERS-CoV (red box). The S1 and S2 domains of Spike S protein are displayed.*

***Fig. 4.*** *Characterization of protein interactions with SARS-CoV-2 RNA, (**A**) Number of RBP interactions for different SARS-CoV-2 regions (colours indicate different catRAPID* [16,17] *confidence levels: Z=1.5 or low Z=1.75 or medium and Z=2.0 or high; regions with scores lower than Z=1.5 are omitted); (**B**) Enrichment of viral processes in the 5' of SARS-CoV-2 (precision = term precision calculated from the GO graph structure lvl = depth of the term; go_term = GO term identifier, with link to term description at AmiGO website ; description = Textual label for the term; e/d =* e *signifies enrichment of the term,* d *signifies depletion compared to the population; %_set = coverage on the provided set - how much of the set is annotated with the GO?; %_pop = coverage of the same term on the population; p_bonf = p-value of the enrichment. To correct for multiple*

*testing bias, we are applying Bonferroni correction)* [35]*; (**C**) Viral processes are the third largest cluster identified in our analysis; (**D**) Protein interactions with the 5' of SARS-CoV-2 RNA (inner circle) and associations with other human genes retrieved from literature (green: genetic associations; pink: physical associations); (**E**) Number of RBP interactions identified by Gordon et al.* [49] *for different SARS-CoV-2 regions (see panel A for reference).*

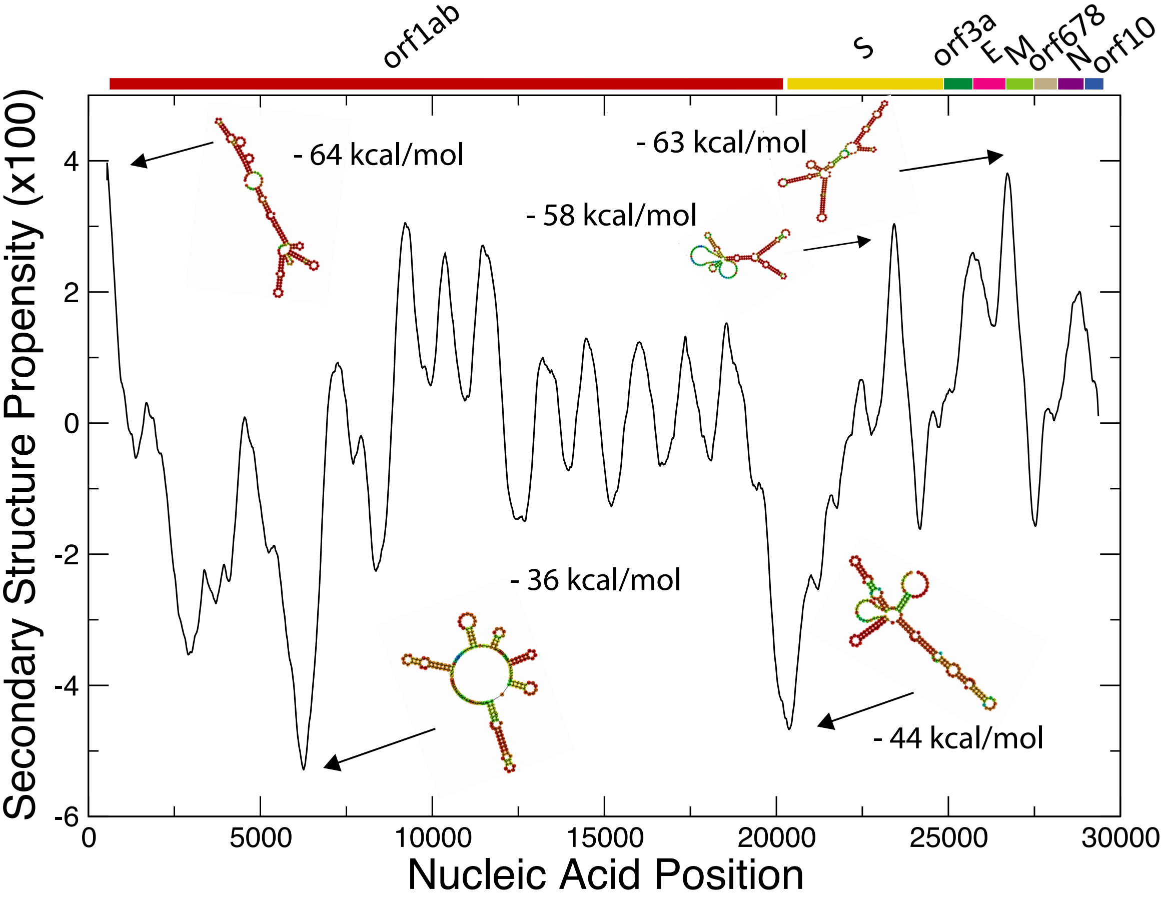
## SUPPLEMENTARY MATERIAL

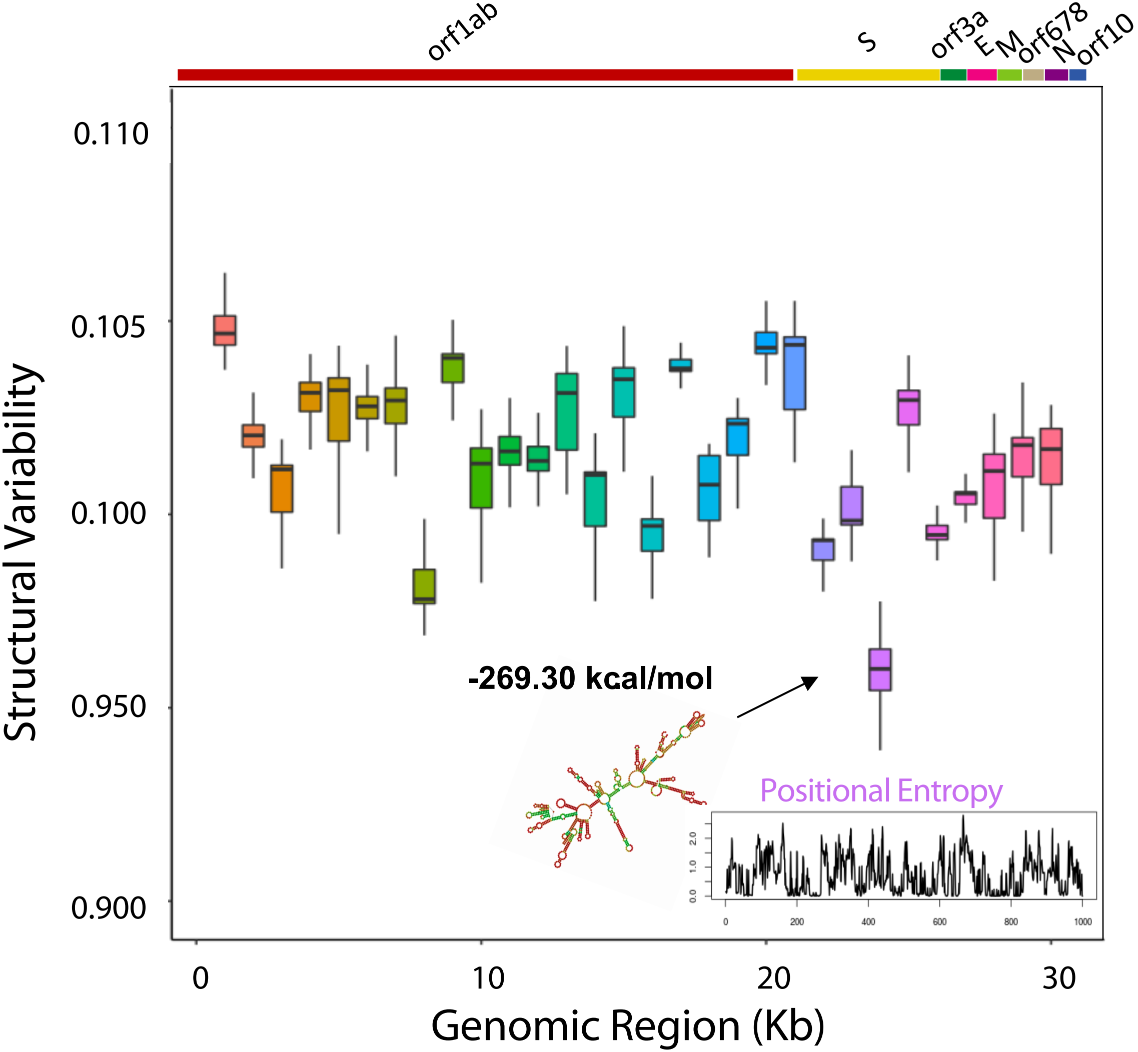
***Supp. Figure 1.*** *We employed CROSSalign* [12,54] *was to compare the Wuhan strain MN908947 with other coronaviruses (2800 strains, including SARS-CoV, MERS-CoV and coronaviruses having as host other species, such as bats). The result highlights that the most conserved region falls inside the spike S genomic locus.*

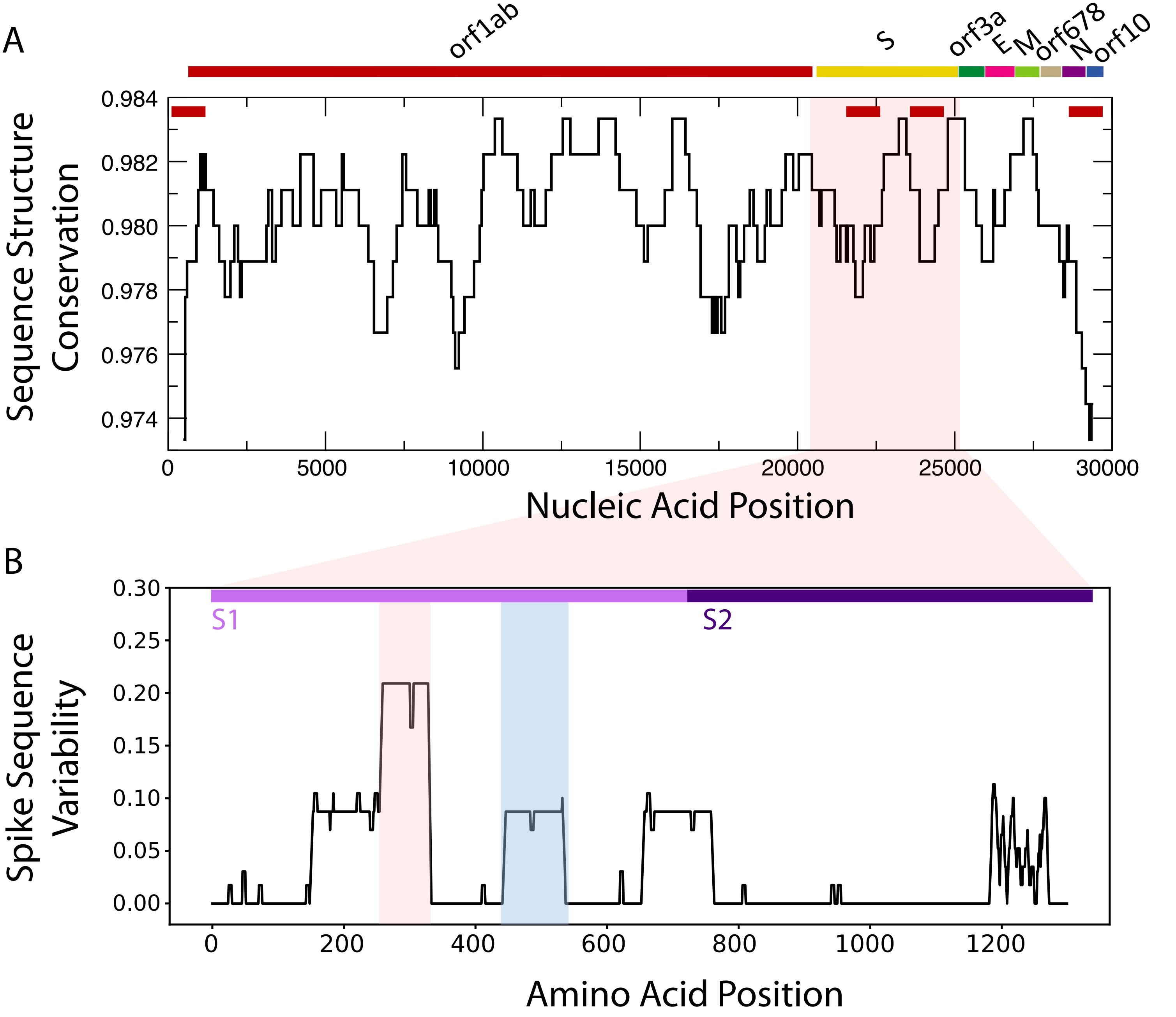
***Supp. Table 1.*** *1) catRAPID* [16,17] *score for interactions with fragment 1; 2) GO* [35] *and Uniprot annotations of viral proteins interacting with fragment 1 and ; 3**)** catRAPID score for interactions with fragment 2; 4) GO annotations of viral proteins interacting with fragment 2; 5**)** catRAPID score for interactions with fragment 29; 6) GO annotations of viral proteins interacting with fragment 29;*

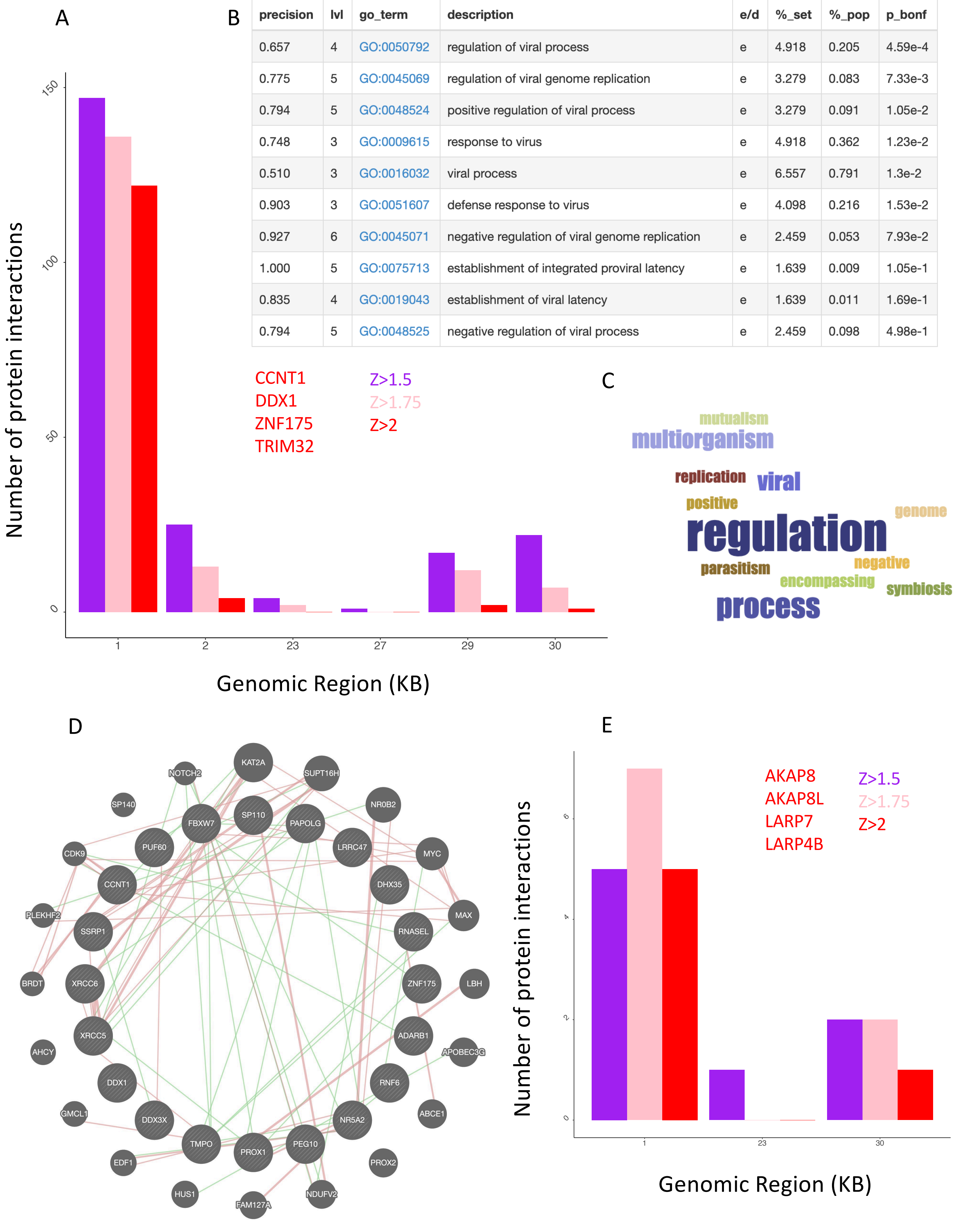
***Supp. Table 2.*** *RBP interactions from Gordon et al.* [49] *classified according to catRAPID scores. GO* [35] *and Uniprot* [36] *annotations are reported.*

***Supp. Table 3**. RBPs significantly enriched in the 5' interactions and HIV, HBV and Influenza*


orf1ab
S
orf3a
E
M
orf678
N
orf10
- 64 kcal/mol
- 63 kcal/mol
- 58 kcal/mol
- 36 kcal/mol
- 44 kcal/mol
Secondary Structure Propensity (x100)
4
2
0
-2
-4
-6
0
5000
10000
15000
20000
25000
30000
Nucleic Acid Position


orf1ab
S
orf3a
E
M
orf678
N
orf10
0.110
0.105
0.100
0.950
0.900
Structural Variability
0
10
20
30
Genomic Region (Kb)
-269.30 kcal/mol
Positional Entropy
0.0
1.0
2.0
0
200
400
600
800
1000


A
orf1ab
S
orf3a
E
M
orf678
N
orf10
Sequence Structure Conservation
0.984
0.982
0.980
0.978
0.976
0.974
0
5000
10000
15000
20000
25000
30000
Nucleic Acid Position
B
0.30
0.25
0.20
0.15
0.10
0.05
0.00
S1
S2
Spike Sequence Variability
0
200
400
600
800
1000
1200
Amino Acid Position



| precision | lvl | go_term | description | e/d | %_set | %_pop | p_bonf |
|---|---|---|---|---|---|---|---|
| 0.657 | 4 | GO:0050792 | regulation of viral process | e | 4.918 | 0.205 | 4.59e-4 |
| 0.775 | 5 | GO:0045069 | regulation of viral genome replication | e | 3.279 | 0.083 | 7.33e-3 |
| 0.794 | 5 | GO:0048524 | positive regulation of viral process | e | 3.279 | 0.091 | 1.05e-2 |
| 0.748 | 3 | GO:0009615 | response to virus | e | 4.918 | 0.362 | 1.23e-2 |
| 0.510 | 3 | GO:0016032 | viral process | e | 6.557 | 0.791 | 1.3e-2 |
| 0.903 | 3 | GO:0051607 | defense response to virus | e | 4.098 | 0.216 | 1.53e-2 |
| 0.927 | 6 | GO:0045071 | negative regulation of viral genome replication | e | 2.459 | 0.053 | 7.93e-2 |
| 1.000 | 5 | GO:0075713 | establishment of integrated proviral latency | e | 1.639 | 0.009 | 1.05e-1 |
| 0.835 | 4 | GO:0019043 | establishment of viral latency | e | 1.639 | 0.011 | 1.69e-1 |
| 0.794 | 5 | GO:0048525 | negative regulation of viral process | e | 2.459 | 0.098 | 4.98e-1 |