Structural analysis of SARS-CoV-2 and prediction 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 and 24000 is highly conserved at the structural level, while the region 1000 nucleotides up-stream varies significantly. The two sequences code for a domain of the spike S protein that binds to the host receptor angiotensin-converting enzyme 2 (ACE2) that mediates human infection and in the homologue from Middle East respiratory syndrome coronavirus (MERS-CoV) interacts with sialic acids. We predicted structured regions at the 5' and 3' where our calculations indicate strong propensity to bind proteins. Using the *cat*RAPID method, we computed 3500 protein interactions with the 5' and identified Cyclin T1 CCNT1, ATP-dependent RNA helicase DDX1, Zinc Finger

Protein ZNF175 and other 20 high-confidence candidate partners. We propose these proteins, also implicated in HIV replication, to 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 March 2020, the virus infected >300'000 people in more than 190 countries.

SARS-CoV-2 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.

The 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 [6]. Spike S is assembled as a trimer and contains around 1,300 amino acids within each unit [7]. 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 [6,8].

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 [9]. To study the RNA structural content, we used *CROSS* [10] that was previously developed to investigate large transcripts such as the human immunodeficiency virus HIV-1 [11]. *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 [16] and promote assembly of complexes [17,18], SARS-CoV-2 domains containing large amount of double-stranded content are expected to establish specific interactions in host cells. To investigate the interactions with human proteins, we employed *cat*RAPID [12,13]. *cat*RAPID [14] estimates the binding potential or protein and RNA molecules through van der Waals, hydrogen bonding and secondary structure propensities of allowing identification of interaction partners with high confidence [15]. Our analysis revealed that the 5' of SARS-CoV-2 has strong propensity to attract human proteins, especially those associated with viral infection, among which we found a group linked to HIV infection. Intriguingly, a previous study reported similarities of viral proteins in SARS-CoV and HIV-1 [19]. In HIV and SARS-CoV-2, but not SARS-CoV nor MERS-CoV, a furin-cleavage site occurs in the spike S protein [20]. This unique feature could explain the spread velocity of SARS-CoV-2 compared to SARS-CoV and MERS-CoV.

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

RESULTS

SARS-CoV-2 contains highly structured elements

Structural elements within RNA molecules attract proteins [16] and reveal regions important for interactions with the host [21].

To analyze SARS-CoV-2 (reference Wuhan strain MN908947), we employed *CROSS* [10] that predicts the double- and single-stranded content of large transcripts such as *Xist* and HIV-1 [11]. 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**) [22]. 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) [22]. 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 [23] to further investigate the RNA secondary structure of specific regions identified with *CROSS* [11]. 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 [9]. Our observation is based on the assumption that structured regions have an intrinsic propensity to recruit proteins [16], which is supported by the fact that structured transcripts act as scaffolds for protein assembly [17,18].

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

We employed *CROSS*align [11] to study the structural conservation of SARS-CoV-2 in different strains.

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 (**Sup. 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 and 24000 and exhibits an intricate and stable secondary structure (RNA*fold* minimum free energy= -269 kcal/mol)[23]. 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 [9], 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 the region 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). Our analysis aims to determine the relationship between structural content and sequence conservation.

Using *Clustal W* for multiple sequence alignments [24], 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 and 24000 in the spike S genomic locus, while sequences up- and down-stream are variable (**Fig. 3A**). We then used CROSS*align* [11] to compare the structural content. High variability of structure is observed for both the 5' and 3' and for nucleotides between 21000 and 22000 as well as 24000 and 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 the sequence (**Fig. 3A**) and structure (**Fig. 2**) of nucleotides 23000 and 24000 are highly conserved. The region corresponds to amino acids 330-500 that contact the host receptor angiotensin-converting enzyme 2 (ACE2) [25] provoking lung injury [26,27]. 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 (**Fig. 3A**). This part of the spike S region corresponds to amino acids 243-302 that in MERS-CoV bind 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") [8,28,29].

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 analysed protein-RNA interactions with SARS-CoV-2 against the whole RNA-binding human proteome. Following a protocol to study structural conservation in viruses [11], we divided the Wuhan sequence in 30 fragments of 1000 nucleotides moving from the 5' to 3' and calculated the protein-RNA interactions of each fragment with the human proteome using *cat*RAPID *omics* (105000 interactions, consisting of around 3500 protein candidates for each of the 30 fragments) [12]. For each fragment, we identified the most significant interactions by filtering according to the Z interaction propensity. We used three different thresholds in ascending order of stringency: Z greater or equal than 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 RNA-binding proteins (around 120 Z>2 high-confidence interactions; **Fig. 4A**), which is in agreement with the observation that highly structured regions attract a large number of proteins [16]. the 5' contains a leader sequence and the untranslated region with multiple stem loop structures that control RNA replication and transcription [30,31].

The interactome of each fragment was then analysed using *clever*GO, a tool for GO enrichment analysis [32]. 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 RNA-binding proteins involved in positive regulation of viral processes and viral genome replication, such as CCNT1 (Uniprot code O60563), ADARB1 (P78563) and RNASEL (Q05823). We also identified proteins related to the establishment of integrated proviral latency, including XRCC5 (P13010) and XRCC6 (P12956; **Fig. 5**). Importantly, we found proteins related to defence response to viruses, such as DDX1 (Q92499) and ZNF175 (Q9Y473), while PROX1 (Q92786) is involved in the negative regulation of viral genome replication. Some of the remaining proteins are listed as DNA binding

proteins and were included because they could have potential RNA-binding ability (**Fig. 5**)[33]. As for fragment 2, we found two viral proteins: TRIM32 (Q13049) and 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 *endogenous retrovirus group K Gag polyprotein family*, that perform different tasks during virus assembly, budding, maturation (**Sup. Table 1**).

Analysis of functional annotations carried out with *GeneMania* [34] 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. 5**) [35,36]. Interestingly, some of the proteins, including CCNT1, DDX1, ZNF175 for fragment 1 and TRIM32 for fragment 2, are reported to be necessary for HIV functions and replications inside the cells. More specifically, in the case of HIV infection, CCNT1 binds to the transactivation domain of the viral nuclear transcriptional activator, Tat, increasing Tat's affinity for the transactivation response RNA element; by doing so, it becomes an essential cofactor for Tat, promoting RNA Pol II activation and allowing transcription of viral genes [37,38]. 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 [39,40]. ZNF175 is reported to interfere with HIV-1 replication by suppressing Tat-induced viral LTR promoter activity [41]. Finally, TRIM32 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* [42].

CONCLUSIONS

Our study is motivated by the need to identify interactions involved in Covid-19 spreading. Using advanced computational approaches, we investigated the structural content of the virus and predicted its binding to human proteins.

We employed *CROSS* [11,43] 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 has evolved to specifically interact

with its host partner ACE2 [25] and mutations increasing the binding affinity are infrequent. As the nucleic acids encoding for this region are enriched in double-stranded content, we speculate that the structure might attract host regulatory proteins, 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 acid in MERS-CoV (see related manuscript by E. Milanetti *et al*. "In-Silico evidence for two receptors based strategy of SARS-CoV-2") [6,8,29] and regulates host cell infection [28]. The fact that the binding region change in the different strains might indicate different binding affinities, which could provide clues on the different levels of contagion in the human population. Interestingly, the sialic acid binding 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 [12,13] we predicted that the highly structured region at the 5' binds is the region with largest number or protein partners, including DDX1, which has been previously reported to be essential for HIV-1 and coronavirus IBV [39,40], the non-canonical RNA-binding proteins CCNT1 [37,38] and ZNF175 [41] involved in polymerase II recruitment. Connections to regulatory pathways involving NOTCH2, MYC and MAX have also been identified.

We hope that our analysis would be useful to the scientific community to identify the interactions to 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 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 (BFU2014-55054-P and BFU2017-86970-P), 'Fundació La Marató de TV3' (PI043296) and the collaboration with Peter St. George-Hyslop financed by the Wellcome Trust. We acknowledge support of the Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa 2013-2017'. We also acknowledge the support of the CERCA Programme/Generalitat de Catalunya and of Spanish Ministry for Science and Competitiveness (MINECO) to the EMBL partnership.

MATERIALS AND METHODS

Structure prediction

We predicted the secondary structure of transcripts using *CROSS* (Computational Recognition of Secondary Structure [11,43]. *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 [23] to further investigate the RNA secondary structure of minima and maxima identified with *CROSS* [11].

Structural conservation

We used *CROSS*align [11,43] an algorithm based on Dynamic Time Warping (DTW), to check and evaluate the structural conservation between different viral genomes [11]. 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 [11].

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* [12,13] , an algorithm that estimates the binding propensity of protein-RNA

pairs by combining secondary structure, hydrogen bonding and van der Waals contributions. 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. We tried 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 [32], 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* [34].

RNA and protein alignments

The alignments were performed using *Clustal W* [24] on 62 SARS-CoV-2 strains and related Spike S proteins. The variability in the spike S region was measured by computing Shannon entropy on translated RNA sequences. The Shannon entropy is computed as follows:

$$S(a) = - \text{Sum}_i P(a,i) \log P(a,i)$$

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. 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).
7. Walls, A. C. *et al.* Cryo-electron microscopy structure of a coronavirus spike glycoprotein trimer. *Nature* **531**, 114–117 (2016).
8. 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).
9. Yang, D. & Leibowitz, J. L. The Structure and Functions of Coronavirus Genomic 3' and 5' Ends. *Virus Res.* **206**, 120–133 (2015).
10. 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).
11. 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).

12. Agostini, F. *et al.* catRAPID omics: a web server for large-scale prediction of protein-RNA interactions. *Bioinforma. Oxf. Engl.* **29**, 2928–2930 (2013).
13. Cirillo, D. *et al.* Quantitative predictions of protein interactions with long noncoding RNAs. *Nat. Methods* **14**, 5–6 (2017).
14. Bellucci, M., Agostini, F., Masin, M. & Tartaglia, G. G. Predicting protein associations with long noncoding RNAs. *Nat. Methods* **8**, 444–445 (2011).
15. 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.
16. Sanchez de Groot, N. *et al.* RNA structure drives interaction with proteins. *Nat. Commun.* **10**, 3246 (2019).
17. 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).
18. Cerase, A. *et al.* Phase separation drives X-chromosome inactivation: a hypothesis. *Nat. Struct. Mol. Biol.* **26**, 331 (2019).
19. Kliger, Y. & Levanon, E. Y. Cloaked similarity between HIV-1 and SARS-CoV suggests an anti-SARS strategy. *BMC Microbiol.* **3**, 20 (2003).
20. Hallenberger, S. *et al.* Inhibition of furin-mediated cleavage activation of HIV-1 glycoprotein gp160. *Nature* **360**, 358–361 (1992).
21. 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).
22. Wu, A. *et al.* Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. *Cell Host Microbe* **27**, 325–328 (2020).
23. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).

24. Madeira, F. *et al.* The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* **47**, W636–W641 (2019).
25. 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.
26. 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).
27. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
28. Qing, E., Hantak, M., Perlman, S. & Gallagher, T. Distinct Roles for Sialoside and Protein Receptors in Coronavirus Infection. *mBio* **11**, (2020).
29. Milanetti, E. *et al.* In-Silico evidence for two receptors based strategy of SARS-CoV-2. *ArXiv200311107 Phys. Q-Bio* (2020).
30. Lu, K., Heng, X. & Summers, M. F. Structural determinants and mechanism of HIV-1 genome packaging. *J. Mol. Biol.* **410**, 609–633 (2011).
31. Fehr, A. R. & Perlman, S. Coronaviruses: An Overview of Their Replication and Pathogenesis. *Methods Mol. Biol. Clifton NJ* **1282**, 1–23 (2015).
32. 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).
33. Castello, A. *et al.* Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* **149**, 1393–1406 (2012).
34. 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).
35. Hayward, S. D. Viral interactions with the Notch pathway. *Semin. Cancer Biol.* **14**, 387–396 (2004).

36. 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).
37. 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).
38. 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).
39. Fang, J. *et al.* A DEAD box protein facilitates HIV-1 replication as a cellular co-factor of Rev. *Virology* **330**, 471–480 (2004).
40. 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).
41. Carlson, K. A. *et al.* Molecular characterization of a putative antiretroviral transcriptional factor, OTK18. *J. Immunol. Baltim. Md 1950* **172**, 381–391 (2004).
42. 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).
43. 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).
44. Arabi, Y. M. *et al.* Treatment of Middle East respiratory syndrome with a combination of lopinavir/ritonavir and interferon-β1b (MIRACLE trial): statistical analysis plan for a recursive two-stage group sequential randomized controlled trial. *Trials* **21**, 8 (2020).

FIGURES LEGENDS

Fig. 1 *Using the CROSS approach, 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 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) is observed in coding regions, including the region between nucleotides 23000 and 24000 of spike S protein. High structural variability (pink bars) 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) 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 protein interactions for different SARS-CoV-2 regions (colours indicate different 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*

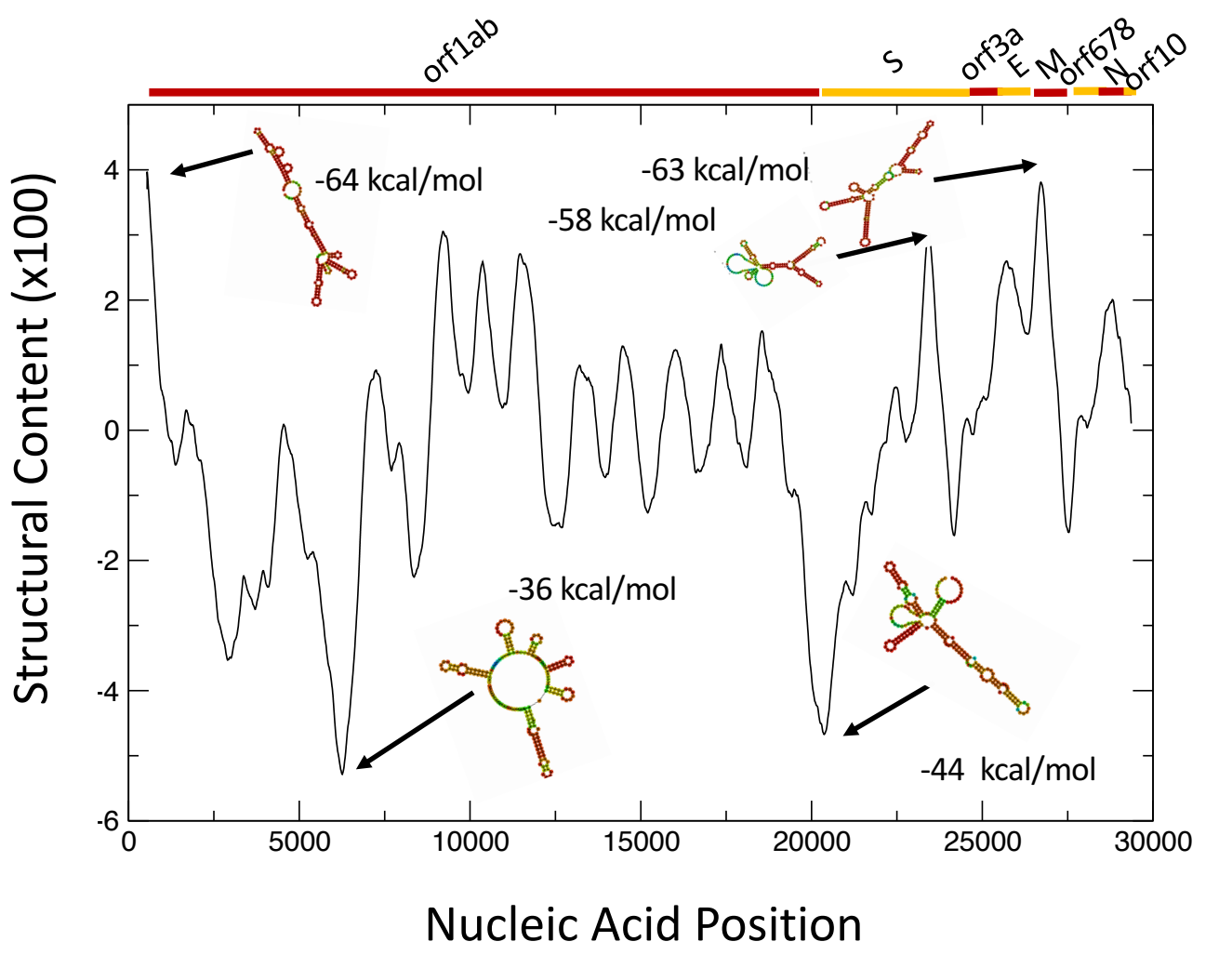
*applying Bonferroni correction); (**C**) Viral processes are the third largest cluster identified in our analysis.*

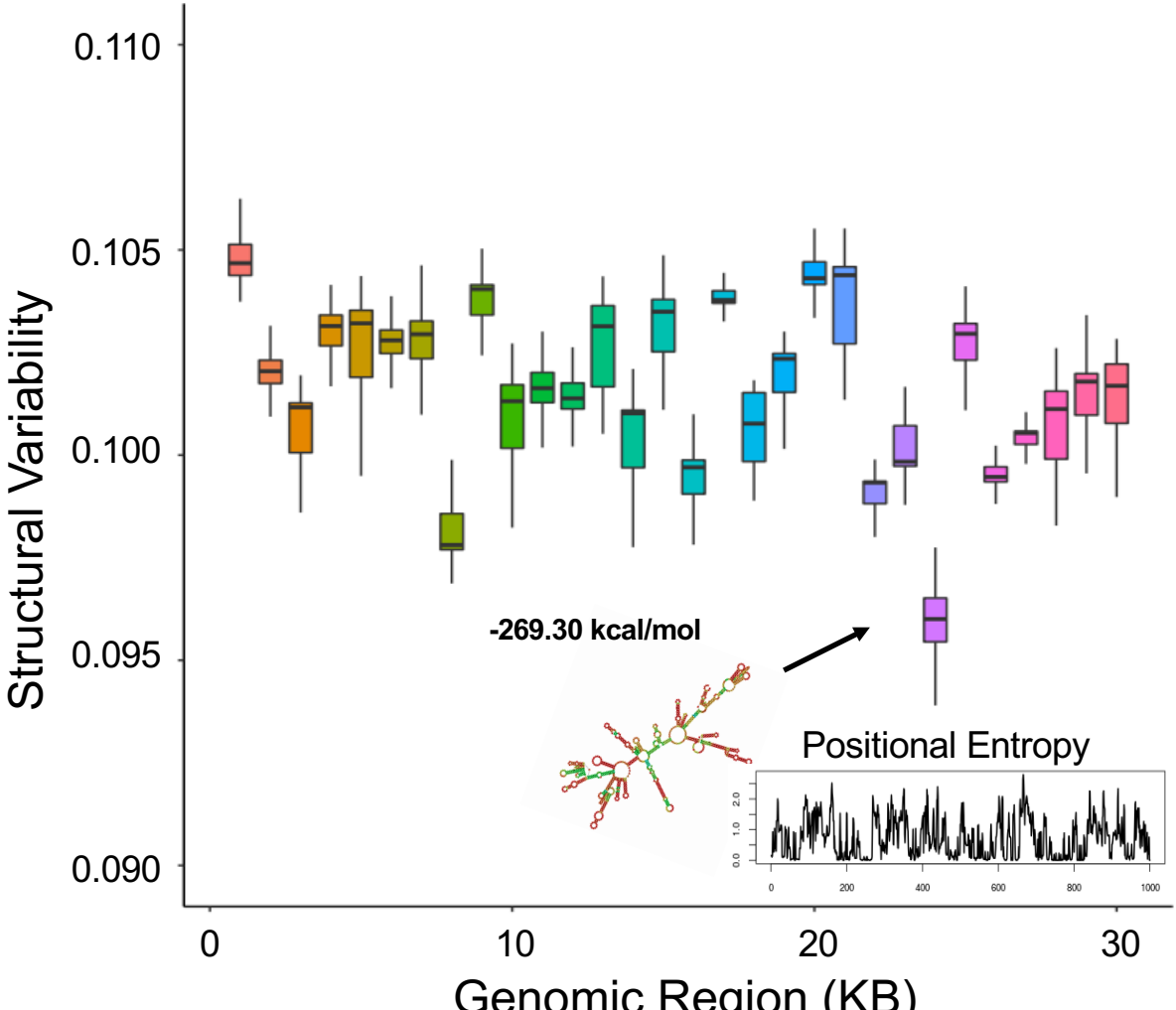
Fig. 5. *Protein interactions with the 5' of SARS-CoV-2 RNA (inner circle) (**Fig. 4**) and associations with other human genes retrieved from literature (green: genetic associations; pink: physical associations).*

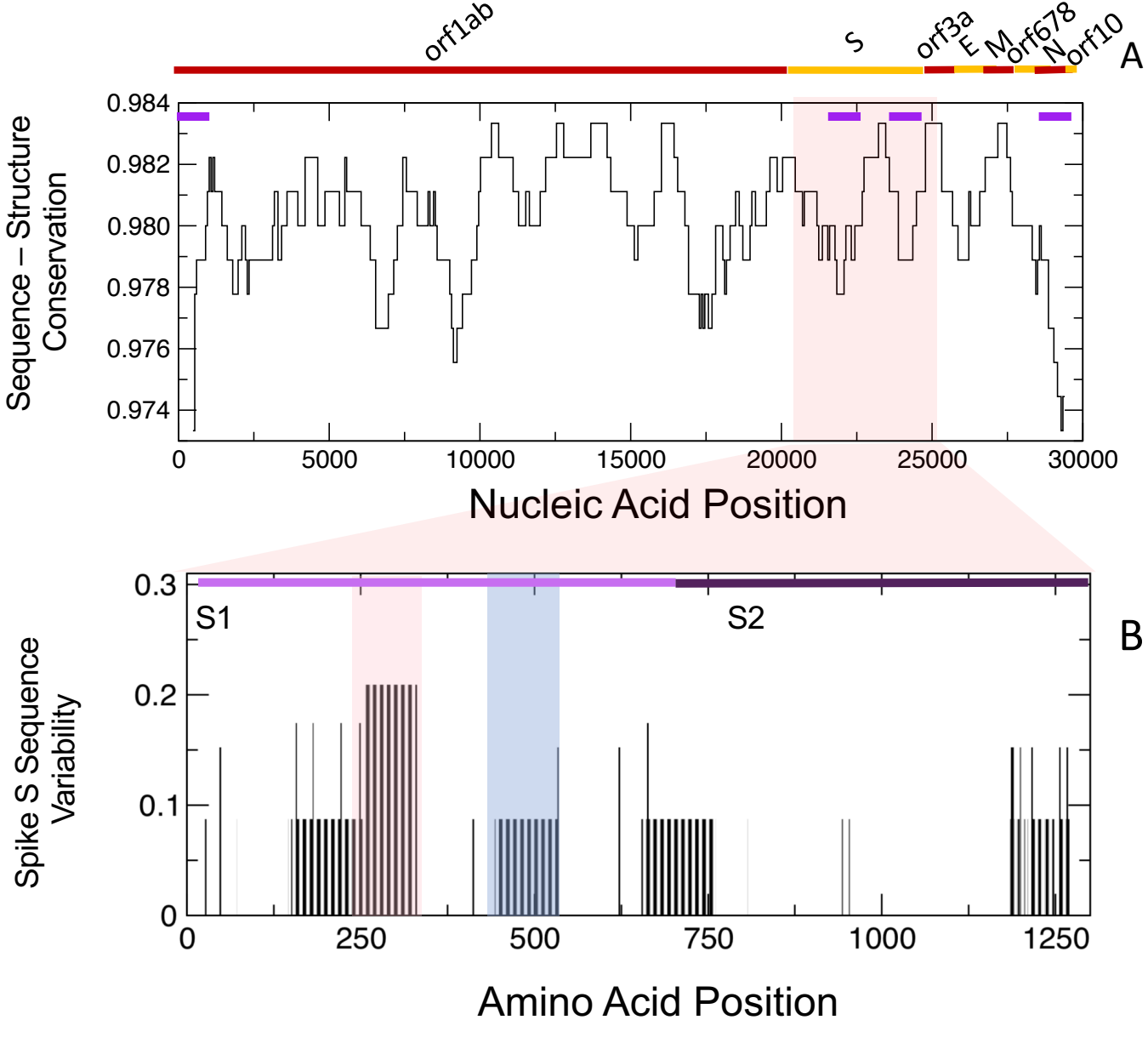
SUPPLEMENTARY MATERIAL

Sup. Table 1. *1) catRAPID score for interactions with fragment 1; 2) GO annotations of viral proteins interacting with fragment 1; 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;*

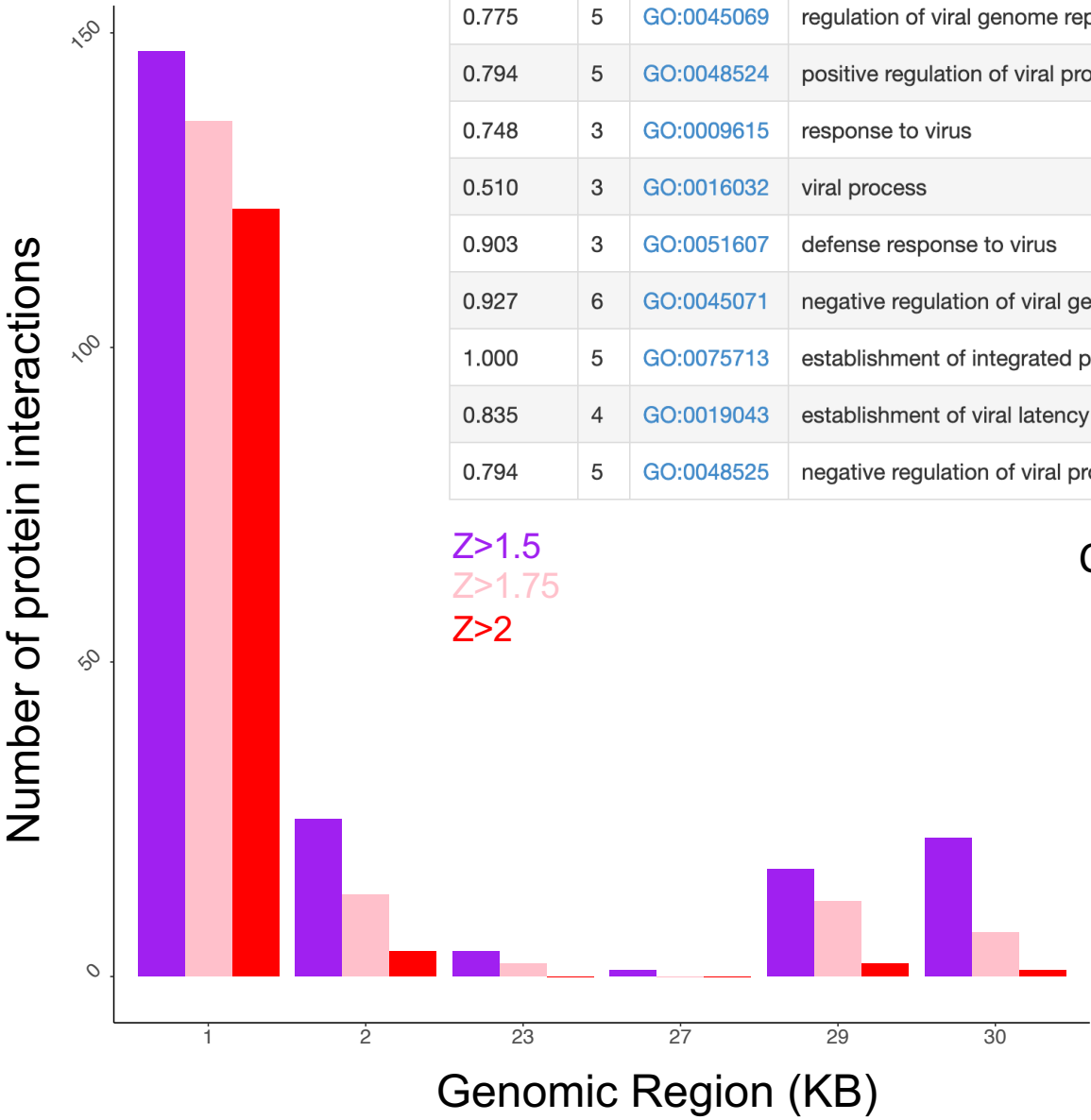
Sup. Figure 1. *CROSSalign was employed to compare the Wuhan strain MN908947 with other coronaviruses (2800 strains, including SARS-CoV, MERS-CoV and coronaviruses having as host other species, such as bats). The result highlights that the most conserved region falls inside the spike S genomic locus.*


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


Structural Variability
0.110
0.105
0.100
0.095
0.090
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


orf1ab
S
orf3a
E
M
orf678
N
orf10
A
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
S1
S2
Spike S Sequence Variability
0.3
0.2
0.1
0
0
250
500
750
1000
1250
Amino Acid Position

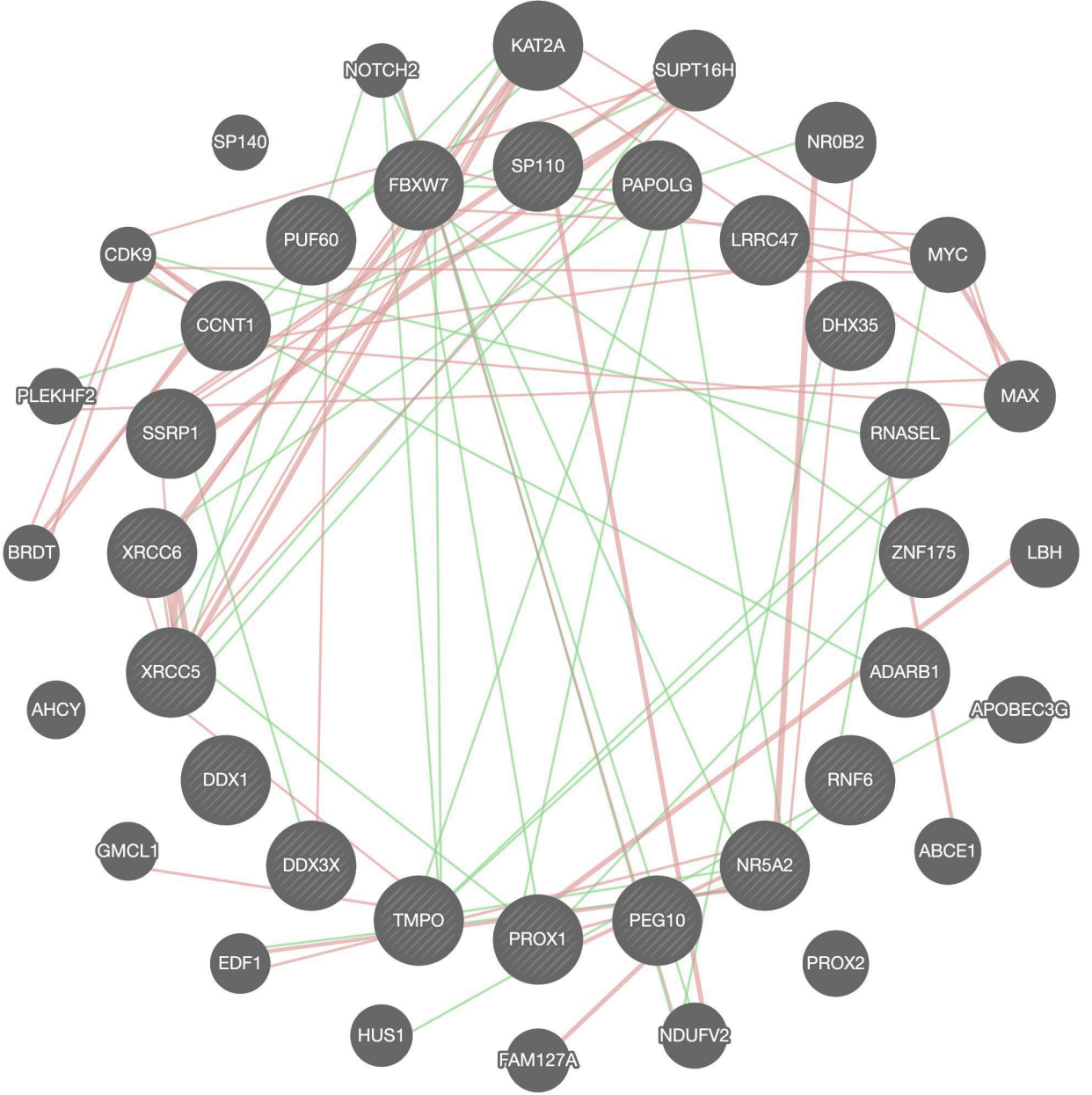
A



B

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

C


KAT2A
NOTCH2
SUPT16H
SP140
NR0B2
SP110
FBXW7
PAPOLG
PUF60
LRRC47
CDK9
MYC
CCNT1
DHX35
PLEKHF2
MAX
SSRP1
RNASEL
BRDT
XRCC6
ZNF175
LBH
XRCC5
ADARB1
AHCY
APOBEC3G
DDX1
RNF6
GMCL1
DDX3X
NR5A2
ABCE1
TMPO
PROX1
PEG10
EDF1
PROX2
HUS1
NDUFV2
FAM127A