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

Andrea Vandelli[1,2], Michele Monti[1,3], Edoardo Milanetti[4,5],
Jakob Rupert [6], Elsa Zacco[3], Elias Bechara[1,3],
Riccardo Delli Ponti [7,*] and Gian Gaetano Tartaglia [1,3,6,8,*]

[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] 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
[6] Department of Biology 'Charles Darwin', Sapienza University of Rome, P.le A. Moro 5, Rome 00185, Italy
[7] School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore, 637551, Singapore
[8] 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

Specific elements of viral genomes regulate interactions within host cells. Here, we calculated the secondary structure content of >2500 coronaviruses and computed >100000 human protein interactions with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). We found that the 3' and 5' are the most structured elements in the viral genome and the 5' has the strongest propensity to associate with human proteins. The domain encompassing nucleotides 23000 – 24000 is highly conserved both at the sequence and structural level, while the region upstream varies significantly. These two sequences code for a domain of the viral protein Spike S that interacts with the human receptor angiotensin-converting enzyme 2 (ACE2) and has the potential to bind sialic acids. Our predictions indicate that the first 1000 nucleotides in the 5' can interact with proteins

36 involved in viral RNA processing such as Double-stranded RNA-specific Editase 1 ADARB1, A-
37 kinase anchor protein 8-like AKAP8L and ATP-dependent RNA Helicase DDX1, in addition to
38 other high-confidence candidate partners. These interactions, previously reported to be implicated
39 in HIV, reveal important information on host-virus interactions. The list of transcriptional and post-
40 transcriptional elements recruited by SARS-CoV-2 genome provides clues on the biological
41 pathways associated with gene expression changes in human cells.
42
43
44 **INTRODUCTION**
45
46 A disease named Covid-19 by the World Health Organization and caused by the severe acute
47 respiratory syndrome coronavirus 2 (SARS-CoV-2) has been recognized as responsible for the
48 pneumonia outbreak that started in December, 2019 in Wuhan City, Hubei, China [1] and spread in
49 February to Milan, Lombardy, Italy [2] becoming pandemic. As of April 2020, the virus infected
50 >2'000'000 people in >200 countries.
51
52 SARS-CoV-2 is a positive-sense single-stranded RNA virus that shares similarities with other beta-
53 coronavirus such as severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East
54 respiratory syndrome coronavirus (MERS-CoV)[3]. Bats have been identified as the primary host for
55 SARS-CoV and SARS-CoV-2 [4,5] but the intermediate host linking SARS-CoV-2 to humans is still
56 unknown, although a recent report indicates that pangolins could be involved [6].
57
58 Coronaviruses use species-specific proteins to mediate the entry in the host cell and the spike S
59 protein activates the infection in human respiratory epithelial cells in SARS-CoV, MERS-CoV and
60 SARS-CoV-2 [7]. Spike S is assembled as a trimer and contains around 1,300 amino acids within
61 each unit [8,9]. The receptor binding domain (RBD) of Spike S, which contains around 300 amino
62 acids, mediates the binding with angiotensin-converting enzyme, (ACE2) attacking respiratory
63 cells. Another region upstream of the RBD, present in MERS-CoV but not in SARS-CoV, is
64 involved in the adhesion to sialic acid and could play a key role in regulating viral infection [7,10].
65
66 At present, very few molecular details are available on SARS-CoV-2 and its interactions with the
67 human host, which are mediated by specific RNA elements [11]. To study the RNA structural content,
68 we used *CROSS* [12] that was previously developed to investigate large transcripts such as the human
69 immunodeficiency virus HIV-1 [13]. *CROSS* predicts the structural profile of RNA molecules (single-

and double-stranded state) at single-nucleotide resolution using sequence information only. Here, 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 provides 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 [14] and promote assembly of specific complexes [15,16], SARS-CoV-2 domains enriched in double-stranded content are expected to establish interactions within host cells that are important to replicate the virus [17]. To investigate the interactions of SARS-CoV-2 RNA with human proteins, we employed *cat*RAPID [18,19]. *cat*RAPID [20] estimates the binding potential of a specific protein for an RNA molecule through van der Waals, hydrogen bonding and secondary structure propensities allowing identification of interaction partners with high confidence [21]. 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 [22], 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 occurring in the spike S protein, which could explain the spread velocity of SARS-CoV-2 compared to SARS-CoV and MERS-CoV [23,24]. 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

Structured elements within RNA molecules attract proteins [14] and reveal regions important for interactions with the host [25]. Indeed, each gene expressed from SARS-CoV-2 is preceded by conserved transcription-regulating sequences that act as signal for the transcription complex during the synthesis of the RNA minus strand to promote a strand transfer to the leader region to resume the synthesis. This process is named discontinuous extension of the minus strand and is a variant of similarity-assisted template switching that operates during viral RNA recombination [17].

104

105 To analyze SARS-CoV-2 structure (reference Wuhan strain MN908947.3), we employed *CROSS* [12]
106 to predict the double- and single-stranded content of RNA genomes such as HIV-1 [13]. We found the
107 highest density of double-stranded regions in the 5' (nucleotides 1-253), membrane M protein
108 (nucleotides 26523-27191), spike S protein (nucleotides 23000-24000), and nucleocapsid N protein
109 (nucleotides 2874-29533; **Fig. 1**) [26]. The lowest density of double-stranded regions were observed
110 at nucleotides 6000-6250 and 20000-21500 and correspond to the regions between the non-
111 structural proteins nsp14 and nsp15 and the upstream region of the spike surface protein S (**Fig. 1**)
112 [26]. In addition to the maximum corresponding to nucleotides 23000-24000, the structural content of
113 spike S protein shows minima at around nucleotides 20500 and 24500 (**Fig. 1**).
114 We used the *Vienna* method [27] to further investigate the RNA secondary structure of specific
115 regions identified with *CROSS* [13]. Employing a 100 nucleotide window centered around *CROSS*
116 maxima and minima, we found good match between *CROSS* scores and Vienna free energies (**Fig.**
117 **1).** Strong agreement is also observed between *CROSS* and *Vienna* positional entropy, indicating
118 that regions with the highest structural content have also the lowest structural diversity.
119
120 Our analysis suggests the presence of structural elements in SARS-CoV-2 that have evolved to
121 interact with specific human proteins [11]. Our observation is based on the assumption that structured
122 regions have an intrinsic propensity to recruit proteins [14], which is supported by the fact that
123 structured transcripts act as scaffolds for protein assembly [15,16].
124
125
126 **Structural comparisons reveal that the spike S region of SARS-CoV-2 is conserved among**
127 **coronaviruses**
128
129 We employed *CROSS*align [13] to study the structural conservation of SARS-CoV-2 in different
130 strains (**Materials and Methods**).
131
132 In our analysis, we compared the Wuhan strain MN908947.3 with around 2800 other coronaviruses
133 (data from NCBI) with human (**Fig. 2**) or other hosts (**Supp. Fig. 1**). When comparing SARS-CoV-
134 2 with human coronaviruses (1387 strains, including SARS-CoV and MERS-CoV), we found that
135 the most conserved region falls inside the spike S genomic locus (**Fig. 2**). More precisely, the
136 conserved region is between nucleotides 23000 - 24000 and exhibits an intricate and stable

137 secondary structure (RNA*fold* minimum free energy= -269 kcal/mol)[27]. High conservation of a
138 structured region suggests a functional activity that might be relevant for host infection.
139
140 The 3' and 5' of SARS-CoV-2 are highly variable in our set. However, the 3' and 5' are more
141 structured in SARS-CoV-2 than other coronaviruses (average structural content for SARS-CoV-2 =
142 0.56, indicating that 56% of the CROSS signal is >0, in the 5' and 0.49 in the 3'; by contrast, other
143 coronaviruses have 0.49 in the 5' and 0.42 in the 3').
144
145
146 **Sequence and structural comparisons among SARS-CoV-2 strains indicate conservation of**
147 **the ACE2 binding site and high variability in a region potentially interacting with sialic acids.**
148
149 To better investigate the sequence conservation of SARS-CoV-2, we compared 62 strains isolated
150 from different countries during the pandemic (including China, USA, Japan, Taiwan, India, Brazil,
151 Sweden, and Australia; data from NCBI and in VIPR www.viprbrc.org; **Materials and Methods**).
152 Our analysis aims to determine the relationship between structural content and sequence
153 conservation.
154
155 Using *Clustal W* for multiple sequence alignments [28], we observed general conservation of the
156 coding regions (**Fig. 3A**). The 5' and 3' show high variability due to experimental procedures of the
157 sequencing and are discarded in this analysis [29] . One highly conserved region is between
158 nucleotides 23000 - 24000 in the spike S genomic locus, while sequences up- and downstream are
159 variable (red bars in **Fig. 3A**). We then used CROSS*align* [13] to compare the structural content
160 (**Materials and Methods**). High variability of structure is observed for both the 5' and 3' and for
161 nucleotides between 21000 - 22000 as well as 24000 - 25000, associated with the S region (red bars
162 in **Fig. 3A**). The rest of the regions are significantly conserved at a structural level (p-value <
163 0.0001; Fisher's test).
164
165 We then compared protein sequences coded by the spike S genomic locus (NCBI reference
166 QHD43416) and found that both sequence (**Fig. 3A**) and structure (**Fig. 2**) of nucleotides 23000 -
167 24000 are highly conserved. The region corresponds to amino acids 330-500 that contact the host
168 receptor angiotensin-converting enzyme 2 (ACE2) [30] promoting infection and provoking lung injury
169 [24,31]. By contrast, the region upstream of the binding site receptor ACE2 and located in
170 correspondence to the minimum of the structural profile at around nucleotides 22500-23000 (**Fig. 1**)

171 is highly variable [32], as indicated by *T-coffee* multiple sequence alignments [32] (**Fig. 3A**). This part of
172 the spike S region corresponds to amino acids 243-302 that in MERS-CoV binds to sialic acids
173 regulating infection through cell-cell membrane fusion (**Fig. 3B;** see related manuscript by E.
174 Milanetti *et al*.) [10,33,34].
175
176 Our analysis suggests that the structural region between nucleotides 23000 and 24000 of Spike S
177 region is conserved among coronaviruses (**Fig. 2**) and that the binding site for ACE2 has poor
178 variation in human SARS-CoV-2 strains (**Fig. 3B**). By contrast, the region upstream, which has
179 propensity to bind sialic acids [10,33,34], showed poor structural content and high variability (**Fig. 3B**).
180
181 **Analysis of human interactions with SARS-CoV-2 identifies proteins involved in viral**
182 **replication**
183
184 In order to obtain insights on how the virus replicates in human cells, we predicted SARS-CoV-2
185 interactions with the whole RNA-binding human proteome. Following a protocol to study structural
186 conservation in viruses [13], we first divided the Wuhan sequence in 30 fragments of 1000 nucleotides
187 each moving from the 5' to 3' and then calculated the protein-RNA interactions of each fragment
188 with *cat*RAPID *omics* (3340 canonical and putative RNA-binding proteins, or RBPs, for a total
189 102000 interactions) [18]. Proteins such as Polypyrimidine tract-binding protein 1 PTBP1 (Uniprot
190 P26599) showed the highest interaction propensity (or Z-score; **Materials and Methods**) at the 5'
191 while others such as Heterogeneous nuclear ribonucleoprotein Q HNRNPQ (O60506) showed the
192 highest interaction propensity at the 3', in agreement with previous studies on coronaviruses (**Fig.**
193 **4A**) [35].
194
195 For each fragment, we predicted the most significant interactions by filtering according to the Z
196 score. We used three different thresholds in ascending order of stringency: $Z \geq 1.50$, 1.75 and 2
197 respectively and we removed from the list the proteins that were predicted to interact promiscuously
198 with more than one fragment. Fragment 1 corresponds to the 5' and is the most contacted by RBPs
199 (around 120 with $Z \geq 2$ high-confidence interactions; **Fig. 4B**), which is in agreement with the
200 observation that highly structured regions attract a large number of proteins [14]. Indeed, the 5'
201 contains multiple stem loop structures that control RNA replication and transcription [36,37]. By
202 contrast, the 3' and fragment 23 (Spike S), which are still structured but to a lesser extent, attract
203 fewer proteins (10 and 5, respectively), while fragment 20 (between orf1ab and Spike S) that is
204 predicted to be unstructured, does not have binding partners.

The interactome of each fragment was then analysed using *clever*GO, a tool for Gene Ontology (GO) enrichment analysis [38]. Proteins interacting with fragments 1, 2 and 29 were associated with annotations related to viral processes (**Fig. 4C; 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. 4D**).

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 [39]) 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. 4B,E**). Nucleolin NCL (P19338), a protein known to be involved in coronavirus processing, was predicted to bind tightly to the 5' (**Supp. Table 1**) [40].

Importantly, we found proteins related to defence response to viruses, such as ATP-dependent RNA helicase DDX1 (Q92499), that are involved in negative regulation of viral genome replication. Some DNA-binding proteins such as Cyclin-T1 CCNT1 (O60563), Zinc finger protein 175 ZNF175 (Q9Y473) and Prospero homeobox protein 1 PROX1 (Q92786) were included because they could have potential RNA-binding ability (**Fig. 4E**) [41]. 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, 10 of the 11 viral proteins detected for fragment 29 are members of the Gag polyprotein family, that perform different tasks during HIV assembly, budding, maturation. More than simple scaffold elements, Gag proteins are versatile elements that bind to viral and host proteins as they traffic to the cell membrane (**Supp. Table 1**) [42].

Analysis of functional annotations carried out with *GeneMania* [43] revealed 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. 4E**) [44,45]. Interestingly, some proteins, including DDX1, CCNT1 and ZNF175 for fragment 1 and TRIM32 for fragment 2, have been shown to be necessary for HIV functions and replication inside the cell. The roles of these proteins in retroviral replication are expected to be different in SARS-CoV-2. As

239 SARS-CoV-2 represses host gene expression through a number of unknown mechanisms [46],
240 sequestration of elements such as DDX1 and CCNT1 could be exploited to alter biological
241 pathways in the host cell. Supporting this hypothesis, DXX1 and CCNT1 have been shown to
242 condense in large ribonucleoprotein assemblies such as stress granules [47–49] that are the primary
243 target of RNA viruses [50]. Regarding the function of these proteins in disease, DDX1 is required for
244 HIV-1 Rev as well as for avian coronavirus IBV replication and it binds to the RRE sequence of
245 HIV-1 RNAs [1–3], while CCNT1 binds to 7SK snRNA and regulates transactivation domain of the
246 viral nuclear transcriptional activator, Tat [53,54].
247
248 **Analysis of SARS-CoV-2 proteins reveals common interactions**
249
250 Recently, Gordon *et al*. reported a list of human proteins binding to Open Reading Frames (ORFs)
251 translated from SARS-CoV-2 [57]. Identified through affinity purification followed by mass
252 spectrometry quantification, 332 proteins from HEK-293T cells interact with viral ORF peptides.
253 By selecting 274 proteins binding at the 5' with Z score ≥1.5 (**Supp. Table 1**), of which 140 are
254 exclusively interacting with fragment 1 (**Fig. 4B**), we found that 8 are also reported in the list by
255 Gordon *et al*. [57], which indicates significant enrichment (representation factor of 2.5; p-value of
256 0.02; hypergeometric test with human proteome in background). The fact that our list of protein-
257 RNA binding partners contains elements identified also in the protein-protein network analysis is
258 not surprising, as ribonucleoprotein complexes evolve together [14] and their components sustain each
259 other through different types of interactions [16].
260
261 We note that out of 332 interactions, 60 are RBPs (as reported in Uniprot [39]), which represents a
262 considerable fraction (i.e., 20%), considering that there are around 1500 RBPs in the human
263 proteome (i.e., 6%) and fully justified by the fact that they involve association with viral RNA.
264 Comparing the RBPs present in Gordon *et al.* [57] and those present in our list (79 as in Uniprot), we
265 found an overlap of 6 proteins (representation factor = 26.5; p-value < 10^{-8}; hypergeometric test),
266 including: Janus kinase and microtubule-interacting protein 1 JAKMIP1 (Q96N16), A-kinase
267 anchor protein 8 AKAP8 (O43823) and A-kinase anchor protein 8-like AKAP8L (Q9ULX6), which
268 in case of HIV-1 infection is involved as a DEAD/H-box RNA helicase binding [58], signal
269 recognition particle subunit SRP72 (O76094), binding to the 7S RNA in presence of SRP68, La-
270 related protein 7, LARP7 (Q4G0J3) and La-related protein 4B LARP4B (Q92615), which are part
271 of a system for transcriptional regulation acting by means of the 7SK RNP system [59] (**Fig. 4F;**
272 **Supp. Table 2**). Interestingly, SRP72, LARP7 and LARP4B proteins assemble in stress granules

273 [60–62] that are the targeted by RNA viruses [50]. We speculate that sequestration of these elements is
274 orchestrated by a viral program aiming to recruit host genes [46].
275
276 Moreover, by directly analysing the RNA interaction potential of all the 332 proteins by Gordon *et*
277 *al.* [57], *cat*RAPID identified 38 putative binders at the 5' (Z score ≥ 1.5; 27 occurring exclusively in
278 the 5' and not in other regions of the RNA) [18], including Serine/threonine-protein kinase TBK1
279 (Q9UHD2), among which 10 RBPs (as reported in Uniprot) such as: Splicing elements U3 small
280 nucleolar ribonucleoprotein protein MPP10 (O00566) and Pre-mRNA-splicing factor SLU7
281 (O95391), snRNA methylphosphate capping enzyme MEPCE involved in negative regulation of
282 transcription by RNA polymerase II 7SK (Q7L2J0) [63], Nucleolar protein 10 NOL10 (Q9BSC4) and
283 protein kinase A Radixin RDX (P35241; in addition to those mentioned above; **Supp. Table 2**).
284
285 **HIV-related RBPs are significantly enriched in the 5' interactions**
286
287 In the list of 274 proteins binding to the 5' (fragment 1) with Z score ≥1.5, we found 10 hits
288 associated with HIV (**Supp. Table 3**), which represents a significant enrichment (p-value=0.0004;
289 Fisher's exact test), considering that the total number of HIV-related proteins is 35 in the whole
290 *cat*RAPID library (3340 elements). The complete list of proteins includes ATP-dependent RNA
291 helicase DDX1 (Q92499 also involved in Coronavirus replication [51,52]), ATP-dependent RNA
292 helicase DDX3X (O00571 also involved in Dengue and Zika Viruses), Tyrosine-protein kinase
293 HCK (P08631, nucleotide binding), Arf-GAP domain and FG repeat-containing protein 1 (P52594),
294 Double-stranded RNA-specific editase 1 ADARB1 (P78563), Insulin-like growth factor 2 mRNA-
295 binding protein 1 IGF2BP1 (Q9NZI8), A-kinase anchor protein 8-like AKAP8L (Q9ULX6; its
296 partner AKAP8 is also found in Gordon *et al.* [57]) Cyclin-T1 CCNT1 (O60563; DNA-binding) and
297 Forkhead box protein K2 FOXK2 (Q01167; DNA-binding; **Fig. 4B,E Supp. Table 3**).
298
299
300 Smaller enrichments were found for proteins related to Hepatitis B virus (HBV; p-value=0.01; 3
301 hits out of 7 in the whole *cat*RAPID library; Fisher's exact test), Nuclear receptor subfamily 5
302 group A member 2 NR5A2 (DNA-binding; O00482), Interferon-induced, double-stranded RNA-
303 activated protein kinase EIF2AK2 (P19525), and SRSF protein kinase 1 SRPK1 (Q96SB4) as well
304 as Influenza (p-value=0.03; 2 hits out of 4; Fisher's exact test), Synaptic functional regulator FMR1
305 (Q06787) and RNA polymerase-associated protein RTF1 homologue (Q92541; **Supp. Table 3**). By
306 contrast, no significant enrichments were found for other viruses such as for instance Ebola.
307

308 Very importantly, specific chemical compounds have been developed to interact with HIV- and
309 HVB-related proteins. The list of HIV-related targets reported in ChEMBL [64] includes ATP-
310 dependent RNA helicase DDX1 (CHEMBL2011807), ATP-dependent RNA helicase DDX3X
311 (CHEMBL2011808), Cyclin-T1 CCNT1 (CHEMBL2348842), and Tyrosine-protein kinase HCK
312 (CHEMBL2408778)[64], among other targets. In addition, HVB-related targets are Nuclear receptor
313 subfamily 5 group A member 2 NR5A2 (CHEMBL3544), Interferon-induced, double-stranded
314 RNA-activated protein kinase EIF2AK2 (CHEMBL5785) and SRSF protein kinase 1 SRPK1
315 (CHEMBL4375). We hope that this list can be the starting point for further pharmaceutical studies.
316
317 **Phase-separating proteins are enriched in the 5' interactions**
318
319 A number of proteins identified in our *cat*RAPID calculations have been previously reported to
320 coalesce in large ribonucleoprotein assemblies such as stress granules. Among these proteins, we
321 found double-stranded RNA-activated protein kinase EIF2AK2 (P19525), Nucleolin NCL
322 (P19338), ATP-dependent RNA helicase DDX1 (Q92499), Cyclin-T1 CCNT1 (O60563), signal
323 recognition particle subunit SRP72 (O76094), LARP7 (Q4G0J3) and La-related protein 4B
324 LARP4B (Q92615) as well as Polypyrimidine tract-binding protein 1 PTBP1 (P26599) and
325 Heterogeneous nuclear ribonucleoprotein Q HNRNPQ (O60506) [62]. To further investigate the
326 propensity of these proteins to phase separate in stress granules, we used the *cat*GRANULE
327 algorithm (**Materials and Methods**) [65] . We found that the 274 proteins binding to the 5' (fragment
328 1) with Z score ≥1.5 are highly prone to accumulate in stress-granules (274 proteins with the lowest
329 Z score are used in the comparison; p-value<0.0001; Kolmogorov-Smirnoff; **Fig. 4G; Supp. Table**
330 **4**).
331
332 This finding is particularly relevant because RNA viruses are known to antagonize stress granules
333 formation [50]. Indeed, the role of stress granules and processing bodies in translation suppression and
334 RNA decay have impact on virus replication [66].
335

336 **CONCLUSIONS**

337

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

341

342 We employed *CROSS* [13,67] to compare the structural properties of 2800 coronaviruses and identified
343 elements conserved in SARS-CoV-2 strains. The regions containing the highest amount of structure
344 are the 5' as well as glycoproteins spike S and membrane M.

345

346 We found that the spike S protein domain encompassing amino acids 330-500 is highly conserved
347 across SARS-CoV-2 strains. This result suggests that spike S must have evolved to specifically
348 interact with its host partner ACE2 [30] and mutations increasing the binding affinity are highly
349 infrequent. As the nucleic acids encoding for this region are enriched in double-stranded content,
350 we speculate that the structure might attract host regulatory elements, thus further constraining the
351 variability. The fact that the Spike S region is highly conserved among all the analysed SARS-CoV-
352 2 strains suggests that a specific drug can be designed against it to prevent interactions within the
353 host.

354

355 By contrast, the highly variable region at amino acids 243-302 in spike S protein corresponds to the
356 binding site of sialic acids in MERS-CoV (see manuscript by E. Milanetti *et al.*) [7,10,34] and could
357 play a role in infection [33]. The fact that the binding region changes in the different strains might
358 indicate a variety of binding affinities for sialic acids, which could provide clues on the specific
359 responses in the human population. Interestingly, the sialic acid binding site is absent in SARS-CoV
360 but present in MERS-CoV, which represents an important difference between the diseases.

361

362 Both our sequence and structural analyses of spike S protein indicate high conservation among
363 coronaviruses and suggest that human engineering of SARS-CoV-2 is highly unlikely.

364

365 Using *cat*RAPID [18,19] we computed >100000 protein interactions with SARS-CoV-2 and found
366 previously reported interactions such as Polypyrimidine tract-binding protein 1 PTBP1,
367 Heterogeneous nuclear ribonucleoprotein Q HNRNPQ and Nucleolin NCL [40]. In addition, we
368 discovered that the highly structured region at the 5' has the largest number of protein partners
369 including ATP-dependent RNA helicase DDX1, which was previously reported to be essential for

370 HIV-1 and coronavirus IBV replication [51,52], double-stranded RNA-specific editase 1 ADARB1,
371 which catalyses the hydrolytic deamination of adenosine to inosine and might take part in the
372 chemical modification of SARS-CoV-2 RNA [46] . Other relevant interactions are XRCC5 and
373 XRCC6 members of the HDP-RNP complex interacting with ATP-dependent RNA helicase DHX9
374 [68] and 2-5A-dependent ribonuclease RNASEL that has antiviral effects through a combination of
375 cleavage of single-stranded viral RNAs, inhibition of protein synthesis, induction of apoptosis, and
376 induction of antiviral genes [69].
377
378 A significant overlap exists with the list of protein interactions reported by Gordon *et al.* [57], and
379 among the candidate partners we identified AKAP8L, involved as a DEAD/H-box RNA helicase
380 binding protein involved in HIV infection [58]. In general, proteins associated with retroviral
381 replication are expected to play different roles in SARS-CoV-2. As SARS-CoV-2 massively
382 represses host gene expression [46], we hypothesize that the virus hijacks host pathways by recruiting
383 transcriptional and post-transcriptional elements interacting with polymerase II genes and splicing
384 factors such as for instance La-related protein 7 LARP7 and A-kinase anchor protein 8-like
385 AKAP8L. The link to proteins previously studied in the context of HIV and other viruses, if
386 confirmed experimentally, is particularly relevant for the repurposing of existing drugs [64].
387
388 The idea that SARS-CoV-2 sequesters different elements of the transcriptional machinery is
389 particularly intriguing and is supported by the fact that a large number of proteins identified in our
390 screening are found in stress granules [62]. Indeed, stress granules protect the host innate immunity
391 and are hijacked by viruses to favour their own replication [66]. Moreover, as coronaviruses
392 transcription uses discontinuous RNA synthesis that involves high-frequency recombination [40], it is
393 possible that pieces of the viruses resulting from a mechanism called defective interfering RNAs [70]
394 could act as scaffold to attract host proteins [14,15].
395
396

Acknowledgements

The authors would like to thank Dr. Mattia Miotto, Dr Lorenzo Di Rienzo, Dr. Alexandros Armaos, Dr. Alessandro Dasti and Dr. Claudia Giambartolomei for discussions. We are particularly grateful to Prof. Annalisa Pastore for critical reading, Dr. Gilles Mirambeau for the RT vs RdRP analysis, Dr. Andrea Cerase for the discussing on stress granules and Dr. Roberto Giambruno for pointing to the analysis of PTBP1 and HNRNPQ experiments.

The research leading to these results has been supported by European Research Council (RIBOMYLOME_309545 and ASTRA_855923), the H2020 projects IASIS_727658 and INFORE_825080, the Spanish Ministry of Economy and Competitiveness BFU2017-86970-P 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, JR, EZ and EB analysed the prediction results. AV, RDP and GGT wrote the paper.

The authors do not have conflicts of interests.

417 **MATERIALS AND METHODS**

418

419 ***Structure prediction***

420

421 We predicted the secondary structure of transcripts using *CROSS* (Computational Recognition of
422 Secondary Structure [13,67]. *CROSS* was developed to perform high-throughput RNA profiling. The
423 algorithm predicts the structural profile (single- and double-stranded state) at single-nucleotide
424 resolution using sequence information only and without sequence length restrictions (scores > 0
425 indicate double stranded regions). We used the *Vienna* method [27] to further investigate the RNA
426 secondary structure of minima and maxima identified with *CROSS* [13].

427

428 ***Structural conservation***

429

430 We used *CROSS*align [13,67] an algorithm based on Dynamic Time Warping (DTW), to check and
431 evaluate the structural conservation between different viral genomes [13]. CROSS*align* was
432 previously employed to study the structural conservation of ~5000 HIV genomes. SARS-CoV-2
433 fragments (1000 nt, not overlapping) were searched inside other complete genomes using the OBE
434 (open begin and end) module, in order to search a small profile inside a larger one. The lower the
435 structural distance, the higher the structural similarities (with a minimum of 0 for almost identical
436 secondary structure profiles). The significance is assessed as in the original publication [13].

437

438 ***Sequence collection***

439

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

446

447

448

449

450

451 ***Protein-RNA interaction prediction***

452

453 Interactions between each fragment of target sequence and the human proteome were predicted
454 using *cat*RAPID *omics* [18,19] , an algorithm that estimates the binding propensity of protein-RNA
455 pairs by combining secondary structure, hydrogen bonding and van der Waals contributions. As
456 reported in a recent analysis of about half a million of experimentally validated interactions [31], the
457 algorithm is able to separate interacting vs non-interacting pairs with an area under the ROC curve
458 of 0.78. The complete list of interactions between the 30 fragments and the human proteome is
459 available at http://crg-webservice.s3.amazonaws.com/submissions/2020-
460 03/252523/output/index.html?unlock=f6ca306af0. The output then is filtered according to the Z-
461 score column, which is the interaction propensity normalised by the mean and standard deviation
462 calculated over the reference RBP set (http://s.tartaglialab.com/static_files/shared/faqs.html#4). We
463 used three different thresholds in ascending order of stringency: Z greater or equal than 1.50, 1.75
464 and 2 respectively and for each threshold we then selected the proteins that were unique for each
465 fragment for each threshold.

466

467 ***GO terms analysis***

468

469 *clever*GO [38], an algorithm for the analysis of Gene Ontology annotations, was used to determine
470 which fragments present enrichment in GO terms related to viral processes. Analysis of functional
471 annotations was performed in parallel with *GeneMania* [43].

472

473 ***RNA and protein alignments***

474

475 We used *Clustal W* [28] for 62 SARS-CoV-2 strains alignments and *T-Coffee* [32] for spike S proteins
476 alignments. The variability in the spike S region was measured by computing Shannon entropy on
477 translated RNA sequences. The Shannon entropy is computed as follows:

478

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

480

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

485 ***Predictions of phase separation***

486

487 *cat*GRANULE [65] was employed to identify proteins assembling into biological condensates. Scores
488 > 0 indicate that a protein is prone to phase separate. Structural disorder, nucleic acid binding
489 propensity and amino acid patterns such as arginine-glycine and phenylalanine-glycine are key
490 features combined in this computational approach [65].

491

492

493

494 **References**

495

496 1. Zhu, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J.*
497 *Med.* **382**, 727–733 (2020).
498 2. D'Antiga, L. Coronaviruses and immunosuppressed patients. The facts during the third
499 epidemic. *Liver Transplant. Off. Publ. Am. Assoc. Study Liver Dis. Int. Liver Transplant. Soc.*
500 (2020) doi:10.1002/lt.25756.
501 3. Cascella, M., Rajnik, M., Cuomo, A., Dulebohn, S. C. & Di Napoli, R. Features, Evaluation and
502 Treatment Coronavirus (COVID-19). in *StatPearls* (StatPearls Publishing, 2020).
503 4. Ge, X.-Y. *et al.* Isolation and characterization of a bat SARS-like coronavirus that uses the
504 ACE2 receptor. *Nature* **503**, 535–538 (2013).
505 5. Follis, K. E., York, J. & Nunberg, J. H. Furin cleavage of the SARS coronavirus spike
506 glycoprotein enhances cell-cell fusion but does not affect virion entry. *Virology* **350**, 358–369
507 (2006).
508 6. Xiao, K. *et al.* Isolation and Characterization of 2019-nCoV-like Coronavirus from Malayan
509 Pangolins. *bioRxiv* 2020.02.17.951335 (2020) doi:10.1101/2020.02.17.951335.
510 7. Park, Y.-J. *et al.* Structures of MERS-CoV spike glycoprotein in complex with sialoside
511 attachment receptors. *Nat. Struct. Mol. Biol.* **26**, 1151–1157 (2019).
512 8. Walls, A. C. *et al.* Cryo-electron microscopy structure of a coronavirus spike glycoprotein
513 trimer. *Nature* **531**, 114–117 (2016).

514 9. Ou, X. *et al.* Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its
515 immune cross-reactivity with SARS-CoV. *Nat. Commun.* **11**, 1–12 (2020).
516 10. Li, W. *et al.* Identification of sialic acid-binding function for the Middle East respiratory
517 syndrome coronavirus spike glycoprotein. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E8508–E8517
518 (2017).
519 11. Yang, D. & Leibowitz, J. L. The Structure and Functions of Coronavirus Genomic 3' and 5'
520 Ends. *Virus Res.* **206**, 120–133 (2015).
521 12. Delli Ponti, R., Marti, S., Armaos, A. & Tartaglia, G. G. A high-throughput approach to profile
522 RNA structure. *Nucleic Acids Res.* **45**, e35–e35 (2017).
523 13. Delli Ponti, R., Armaos, A., Marti, S. & Gian Gaetano Tartaglia. A Method for RNA Structure
524 Prediction Shows Evidence for Structure in lncRNAs. *Front. Mol. Biosci.* **5**, 111 (2018).
525 14. Sanchez de Groot, N. *et al.* RNA structure drives interaction with proteins. *Nat. Commun.* **10**,
526 3246 (2019).
527 15. Cid-Samper, F. *et al.* An Integrative Study of Protein-RNA Condensates Identifies Scaffolding
528 RNAs and Reveals Players in Fragile X-Associated Tremor/Ataxia Syndrome. *Cell Rep.* **25**,
529 3422-3434.e7 (2018).
530 16. Cerase, A. *et al.* Phase separation drives X-chromosome inactivation: a hypothesis. *Nat. Struct.*
531 *Mol. Biol.* **26**, 331 (2019).
532 17. Moreno, J. L., Zúñiga, S., Enjuanes, L. & Sola, I. Identification of a Coronavirus Transcription
533 Enhancer. *J. Virol.* **82**, 3882–3893 (2008).
534 18. Agostini, F. *et al.* catRAPID omics: a web server for large-scale prediction of protein-RNA
535 interactions. *Bioinforma. Oxf. Engl.* **29**, 2928–2930 (2013).
536 19. Cirillo, D. *et al.* Quantitative predictions of protein interactions with long noncoding RNAs.
537 *Nat. Methods* **14**, 5–6 (2017).
538 20. Bellucci, M., Agostini, F., Masin, M. & Tartaglia, G. G. Predicting protein associations with
539 long noncoding RNAs. *Nat. Methods* **8**, 444–445 (2011).

540 21. Lang, B., Armaos, A. & Tartaglia, G. G. RNAct: Protein–RNA interaction predictions for
541 model organisms with supporting experimental data. *Nucleic Acids Res.*
542 doi:10.1093/nar/gky967.
543 22. Kliger, Y. & Levanon, E. Y. Cloaked similarity between HIV-1 and SARS-CoV suggests an
544 anti-SARS strategy. *BMC Microbiol.* **3**, 20 (2003).
545 23. Hallenberger, S. *et al.* Inhibition of furin-mediated cleavage activation of HIV-1 glycoprotein
546 gp160. *Nature* **360**, 358–361 (1992).
547 24. Glowacka, I. *et al.* Differential downregulation of ACE2 by the spike proteins of severe acute
548 respiratory syndrome coronavirus and human coronavirus NL63. *J. Virol.* **84**, 1198–1205
549 (2010).
550 25. Gultyaev, A. P., Richard, M., Spronken, M. I., Olsthoorn, R. C. L. & Fouchier, R. A. M.
551 Conserved structural RNA domains in regions coding for cleavage site motifs in hemagglutinin
552 genes of influenza viruses. *Virus Evol.* **5**, (2019).
553 26. Wu, A. *et al.* Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV)
554 Originating in China. *Cell Host Microbe* **27**, 325–328 (2020).
555 27. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
556 28. Madeira, F. *et al.* The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic*
557 *Acids Res.* **47**, W636–W641 (2019).
558 29. Hrdlickova, R., Toloue, M. & Tian, B. RNA-Seq methods for transcriptome analysis. *Wiley*
559 *Interdiscip. Rev. RNA* **8**, (2017).
560 30. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin
561 of SARS-CoV-2. *Nat. Med.* 1–3 (2020) doi:10.1038/s41591-020-0820-9.
562 31. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin.
563 *Nature* **579**, 270–273 (2020).

564 32. Di Tommaso, P. *et al.* T-Coffee: a web server for the multiple sequence alignment of protein
565 and RNA sequences using structural information and homology extension. *Nucleic Acids Res.*
566 **39**, W13–W17 (2011).
567 33. Qing, E., Hantak, M., Perlman, S. & Gallagher, T. Distinct Roles for Sialoside and Protein
568 Receptors in Coronavirus Infection. *mBio* **11**, (2020).
569 34. Milanetti, E. *et al.* In-Silico evidence for two receptors based strategy of SARS-CoV-2.
570 *ArXiv200311107 Phys. Q-Bio* (2020).
571 35. Galán, C. *et al.* Host cell proteins interacting with the 3' end of TGEV coronavirus genome
572 influence virus replication. *Virology* **391**, 304–314 (2009).
573 36. Lu, K., Heng, X. & Summers, M. F. Structural determinants and mechanism of HIV-1 genome
574 packaging. *J. Mol. Biol.* **410**, 609–633 (2011).
575 37. Fehr, A. R. & Perlman, S. Coronaviruses: An Overview of Their Replication and Pathogenesis.
576 *Methods Mol. Biol. Clifton NJ* **1282**, 1–23 (2015).
577 38. Klus, P., Ponti, R. D., Livi, C. M. & Tartaglia, G. G. Protein aggregation, structural disorder
578 and RNA-binding ability: a new approach for physico-chemical and gene ontology
579 classification of multiple datasets. *BMC Genomics* **16**, 1071 (2015).
580 39. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
581 40. Sola, I., Mateos-Gomez, P. A., Almazan, F., Zuñiga, S. & Enjuanes, L. RNA-RNA and RNA-
582 protein interactions in coronavirus replication and transcription. *RNA Biol.* **8**, 237–248 (2011).
583 41. Castello, A. *et al.* Insights into RNA biology from an atlas of mammalian mRNA-binding
584 proteins. *Cell* **149**, 1393–1406 (2012).
585 42. Bell, N. M. & Lever, A. M. L. HIV Gag polyprotein: processing and early viral particle
586 assembly. *Trends Microbiol.* **21**, 136–144 (2013).
587 43. Warde-Farley, D. *et al.* The GeneMANIA prediction server: biological network integration for
588 gene prioritization and predicting gene function. *Nucleic Acids Res.* **38**, W214–W220 (2010).

589 44. Hayward, S. D. Viral interactions with the Notch pathway. *Semin. Cancer Biol.* **14**, 387–396
590 (2004).
591 45. Dudley, J. P., Mertz, J. A., Rajan, L., Lozano, M. & Broussard, D. R. What retroviruses teach
592 us about the involvement of c- Myc in leukemias and lymphomas. *Leukemia* **16**, 1086–1098
593 (2002).
594 46. Kim, D. *et al.* The architecture of SARS-CoV-2 transcriptome. *bioRxiv* 2020.03.12.988865
595 (2020) doi:10.1101/2020.03.12.988865.
596 47. Ning, W. *et al.* DrLLPS: a data resource of liquid–liquid phase separation in eukaryotes.
597 *Nucleic Acids Res.* **48**, D288–D295 (2020).
598 48. Shin, Y. *et al.* Liquid Nuclear Condensates Mechanically Sense and Restructure the Genome.
599 *Cell* **175**, 1481-1491.e13 (2018).
600 49. Su, X. *et al.* Phase separation of signaling molecules promotes T cell receptor signal
601 transduction. *Science* **352**, 595–599 (2016).
602 50. White, J. P. & Lloyd, R. E. Regulation of stress granules in virus systems. *Trends Microbiol.*
603 **20**, 175–183 (2012).
604 51. Fang, J. *et al.* A DEAD box protein facilitates HIV-1 replication as a cellular co-factor of Rev.
605 *Virology* **330**, 471–480 (2004).
606 52. Xu, L. *et al.* The cellular RNA helicase DDX1 interacts with coronavirus nonstructural protein
607 14 and enhances viral replication. *J. Virol.* **84**, 8571–8583 (2010).
608 53. Ivanov, D. *et al.* Cyclin T1 domains involved in complex formation with Tat and TAR RNA are
609 critical for tat-activation. *J. Mol. Biol.* **288**, 41–56 (1999).
610 54. Kwak, Y. T., Ivanov, D., Guo, J., Nee, E. & Gaynor, R. B. Role of the human and murine cyclin
611 T proteins in regulating HIV-1 tat-activation. *J. Mol. Biol.* **288**, 57–69 (1999).
612 55. Carlson, K. A. *et al.* Molecular characterization of a putative antiretroviral transcriptional
613 factor, OTK18. *J. Immunol. Baltim. Md 1950* **172**, 381–391 (2004).

614 56. Locke, M., Tinsley, C. L., Benson, M. A. & Blake, D. J. TRIM32 is an E3 ubiquitin ligase for
615 dysbindin. *Hum. Mol. Genet.* **18**, 2344–2358 (2009).
616 57. Gordon, D. E. *et al.* A SARS-CoV-2-Human Protein-Protein Interaction Map Reveals Drug
617 Targets and Potential Drug-Repurposing. *bioRxiv* 2020.03.22.002386 (2020)
618 doi:10.1101/2020.03.22.002386.
619 58. Xing, L., Zhao, X., Guo, F. & Kleiman, L. The role of A-kinase anchoring protein 95-like
620 protein in annealing of tRNALys3 to HIV-1 RNA. *Retrovirology* **11**, 58 (2014).
621 59. Markert, A. *et al.* The La-related protein LARP7 is a component of the 7SK ribonucleoprotein
622 and affects transcription of cellular and viral polymerase II genes. *EMBO Rep.* **9**, 569–575
623 (2008).
624 60. Schäffler, K. *et al.* A stimulatory role for the La-related protein 4B in translation. *RNA* **16**,
625 1488–1499 (2010).
626 61. Küspert, M. *et al.* LARP4B is an AU-rich sequence associated factor that promotes mRNA
627 accumulation and translation. *RNA* **21**, 1294–1305 (2015).
628 62. Markmiller, S. *et al.* Context-Dependent and Disease-Specific Diversity in Protein Interactions
629 within Stress Granules. *Cell* **172**, 590-604.e13 (2018).
630 63. C, J. *et al.* Systematic Analysis of the Protein Interaction Network for the Human Transcription
631 Machinery Reveals the Identity of the 7SK Capping Enzyme. *Molecular cell* vol. 27
632 https://pubmed.ncbi.nlm.nih.gov/17643375/ (2007).
633 64. Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**,
634 D930–D940 (2019).
635 65. Bolognesi, B. *et al.* A Concentration-Dependent Liquid Phase Separation Can Cause Toxicity
636 upon Increased Protein Expression. *Cell Rep.* **16**, 222–231 (2016).
637 66. Zhang, Q., Sharma, N. R., Zheng, Z.-M. & Chen, M. Viral Regulation of RNA Granules in
638 Infected Cells. *Virol. Sin.* **34**, 175–191 (2019).

639 67. Delli Ponti, R., Marti, S., Armaos, A. & Tartaglia, G. G. A high-throughput approach to profile
640 RNA structure. *Nucleic Acids Res.* **45**, e35–e35 (2017).
641 68. Zhang, S., Schlott, B., Görlach, M. & Grosse, F. DNA-dependent protein kinase (DNA-PK)
642 phosphorylates nuclear DNA helicase II/RNA helicase A and hnRNP proteins in an RNA-
643 dependent manner. *Nucleic Acids Res.* **32**, 1–10 (2004).
644 69. Siddiqui, M. A., Mukherjee, S., Manivannan, P. & Malathi, K. RNase L Cleavage Products
645 Promote Switch from Autophagy to Apoptosis by Caspase-Mediated Cleavage of Beclin-1. *Int.*
646 *J. Mol. Sci.* **16**, 17611–17636 (2015).
647 70. Pathak, K. B. & Nagy, P. D. Defective Interfering RNAs: Foes of Viruses and Friends of
648 Virologists. *Viruses* **1**, 895–919 (2009).
649
650
651

652 **FIGURES LEGENDS**

653

654 ***Fig. 1.*** *Using the CROSS approach* [13,67]*, we studied the structural content of SARS-CoV-2. We*
655 *found the highest density of double-stranded regions in the 5' (nucleotides 1-253), membrane M*
656 *protein (nucleotides 26523-27191), and the spike S protein (nucleotides 23000-24000). Strong*
657 *match is observed between CROSS and Vienna analyses (centroid structures shown), indicating*
658 *that regions with the highest structural content have the lowest free energies.*

659

660 ***Fig. 2.*** *We employed the CROSSalign approach* [13,67] *to compare the Wuhan strain MN908947 with*
661 *other coronaviruses (1387 strains, including SARS-CoV and MERS-CoV) indicates that the most*
662 *conserved region falls inside the spike S genomic locus. The inset shows thermodynamic structural*
663 *variability (positional entropy) within regions encompassing nucleotides 23000-24000 along with*
664 *the centroid structure and free energy.*

665

666 ***Fig. 3.*** *Sequence and structural comparison of human SARS-CoV-2 strains. (**A**) Strong sequence*
667 *conservation (Clustal W multiple sequence alignments* [38]*) is observed in coding regions, including*
668 *the region between nucleotides 23000 and 24000 of spike S protein. High structural variability (red*
669 *bars on top) is observed for both the UTRs and for nucleotides between 21000 and 22000 as well as*
670 *24000 and 25000, associated with the S region. The rest of the regions are significantly conserved*
671 *at a structural level. (**B**) The sequence variability (Shannon entropy computed on T-Coffee multiple*
672 *sequence alignments* [32]*) in the spike S protein indicate conservation between amino-acids 460 and*
673 *520 (blue box) binding to the host receptor angiotensin-converting enzyme 2 ACE2. The region*
674 *encompassing amino-acids 243 and 302 is highly variable and is implicated in sialic acids in*
675 *MERS-CoV (red box). The S1 and S2 domains of Spike S protein are displayed.*

676

677 ***Fig. 4.*** *Characterization of protein interactions with SARS-CoV-2 RNA. (**A**) In agreement with*
678 *studies on coronaviruses* [35]*, PTBP1 shows the highest interaction propensity at the 5' and*
679 *HNRNPQ at 3' (indicated by red bars). (**B**) Number of RBP interactions for different SARS-CoV-2*
680 *regions (colours indicate catRAPID* [18,19] *confidence levels: Z=1.5 or low Z=1.75 or medium and*
681 *Z=2.0 or high; regions with scores lower than Z=1.5 are omitted); (**C**) Enrichment of viral*
682 *processes in the 5' of SARS-CoV-2 (precision = term precision calculated from the GO graph*
683 *structure lvl = depth of the term; go_term = GO term identifier, with link to term description at*
684 *AmiGO website; description = label for the term; e/d = enrichment / depletion compared to the*
685 *population; %_set = coverage on the provided set; %_pop = coverage of the same term on the*

686 *population; p_bonf = p-value of the enrichment. To correct for multiple testing bias, use Bonferroni*
687 *correction)* [38]*; (**D**) Viral processes are the third largest cluster identified in our analysis; (**E**)*
688 *Protein interactions with the 5' of SARS-CoV-2 RNA (inner circle) and associations with other*
689 *human genes retrieved from literature (green: genetic associations; pink: physical associations);*
690 *(**F**) Number of RBP interactions identified by Gordon et al.* [57] *for different SARS-CoV-2 regions*
691 *(see panel A for reference). (**G**) Proteins binding to the 5' with Z score ≥ 1.5 show high propensity*
692 *to accumulate in stress-granules (same number of proteins with Z score* < -1.5 *are used in the*
693 *comparison; *** p-value<0.0001; Kolmogorov-Smirnoff).*

694 **SUPPLEMENTARY MATERIAL**

695

696

697 ***Supp. Figure 1.*** *We employed CROSSalign* [13,67] *was to compare the Wuhan strain MN908947 with*
698 *other coronaviruses (2800 strains, including SARS-CoV, MERS-CoV and coronaviruses having as*
699 *host other species, such as bats). The result highlights that the most conserved region falls inside*
700 *the spike S genomic locus.*

701

702 ***Supp. Table 1.*** *1) catRAPID* [18,19] *score for interactions with fragment 1; 2) GO* [38] *and Uniprot*
703 *annotations of viral proteins interacting with fragment 1 and ; 3) catRAPID score for interactions*
704 *with fragment 2; 4) GO annotations of viral proteins interacting with fragment 2; 5) catRAPID*
705 *score for interactions with fragment 29; 6) GO annotations of viral proteins interacting with*
706 *fragment 29;*

707

708 ***Supp. Table 2.*** *RBP interactions from Gordon et al.* [57] *classified according to catRAPID scores.*
709 *GO* [38] *and Uniprot* [39] *annotations are reported.*

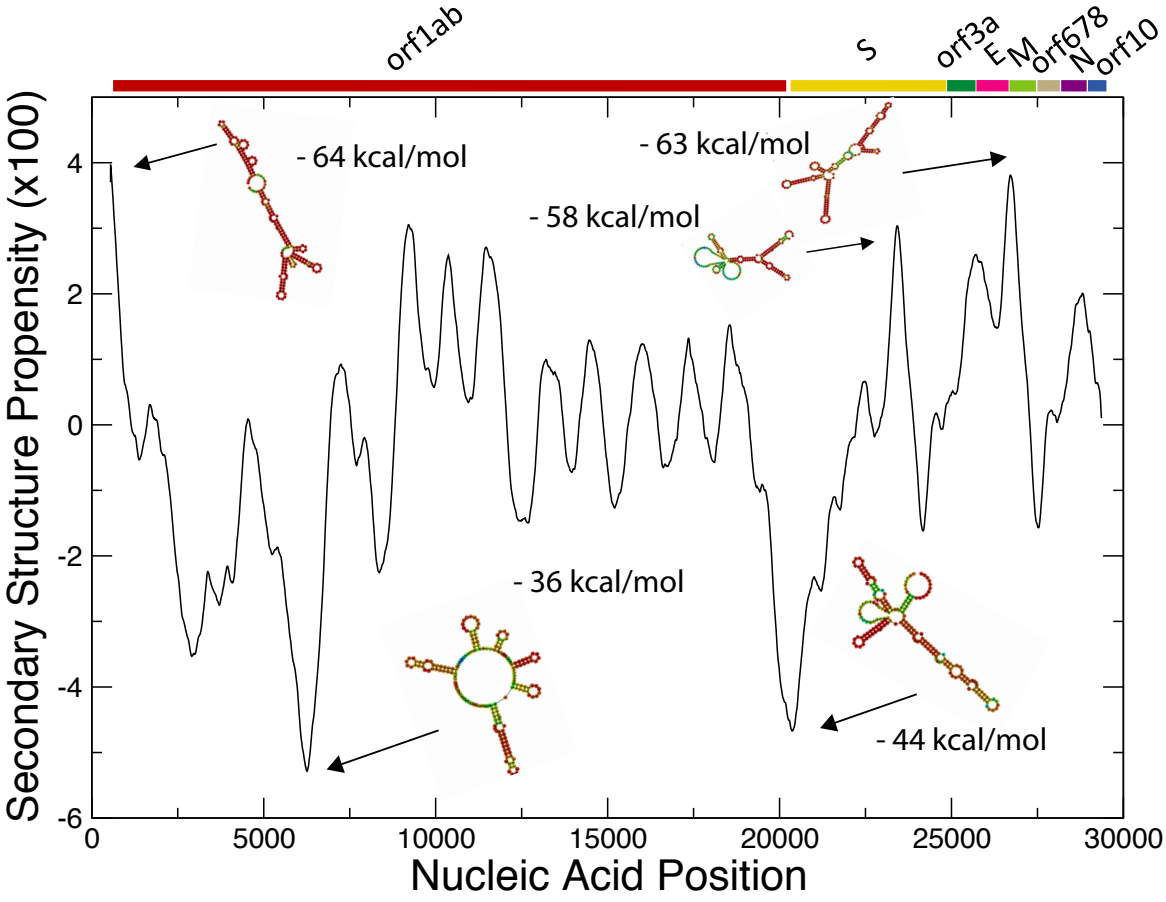
710

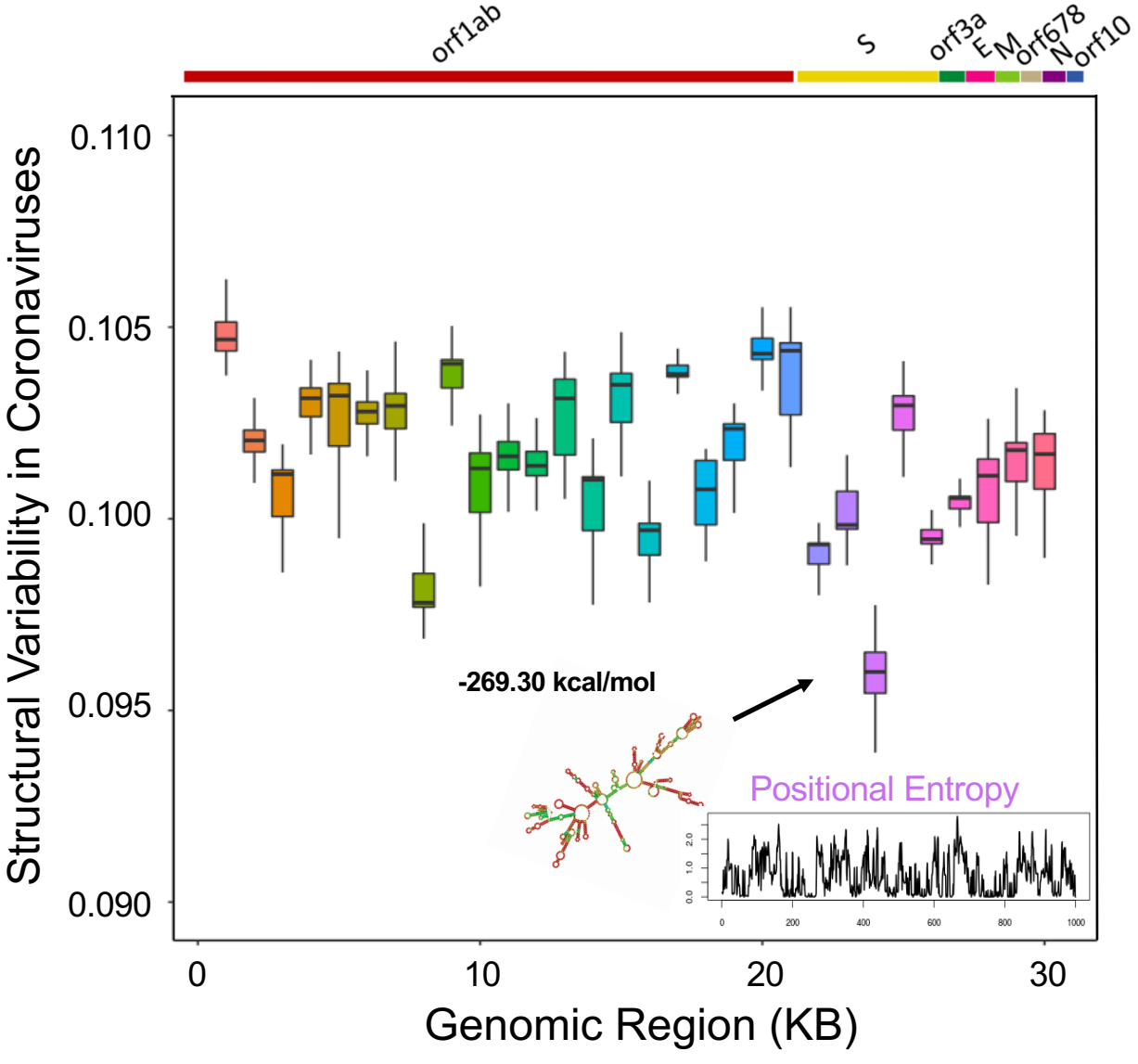
711 ***Supp. Table 3****. RBPs significantly enriched in the 5' with GO* [38] *and Uniprot* [39] *annotations for*
712 *HIV, HBV and Influenza.*

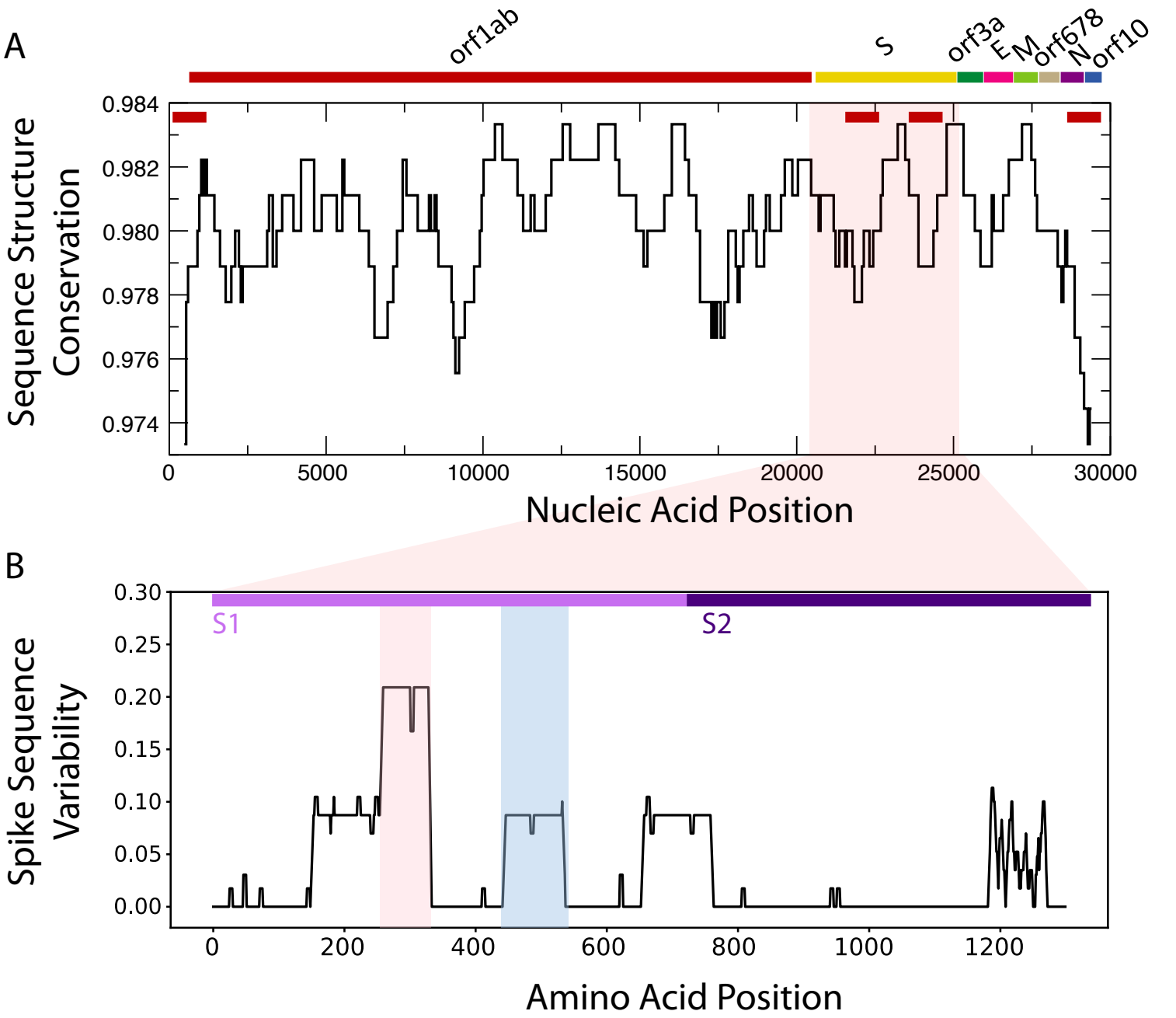
713

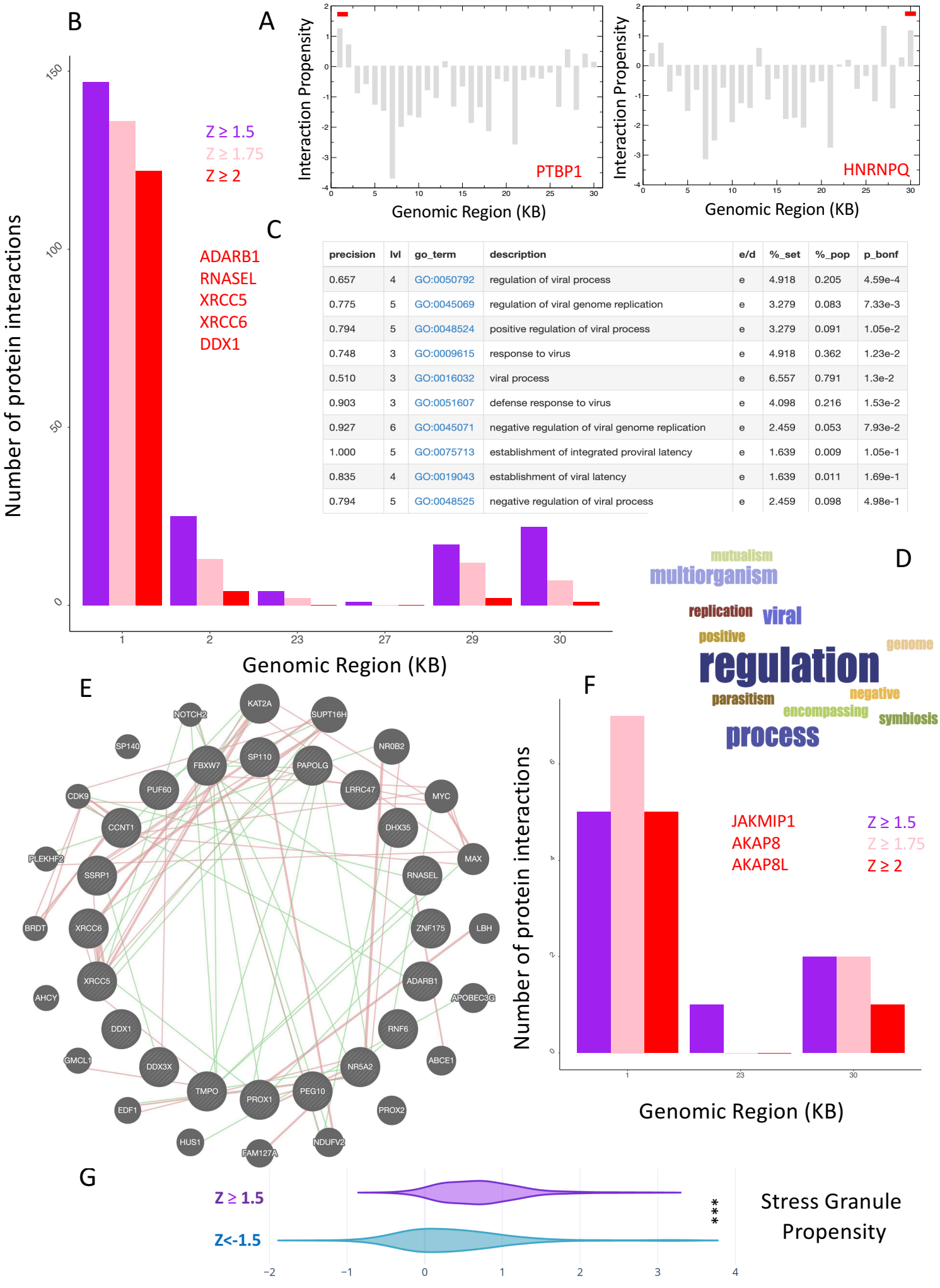
714 ***Supp. Table 4****. Stress granule propensity computed with catGRANULE* [65] *for proteins with*
715 *interactions score* $Z \geq 1.5$ *(top-ranked) and* $Z<-1.5$ *(bottom-ranked). The two groups of proteins*
716 *have the same number of elements.*

717

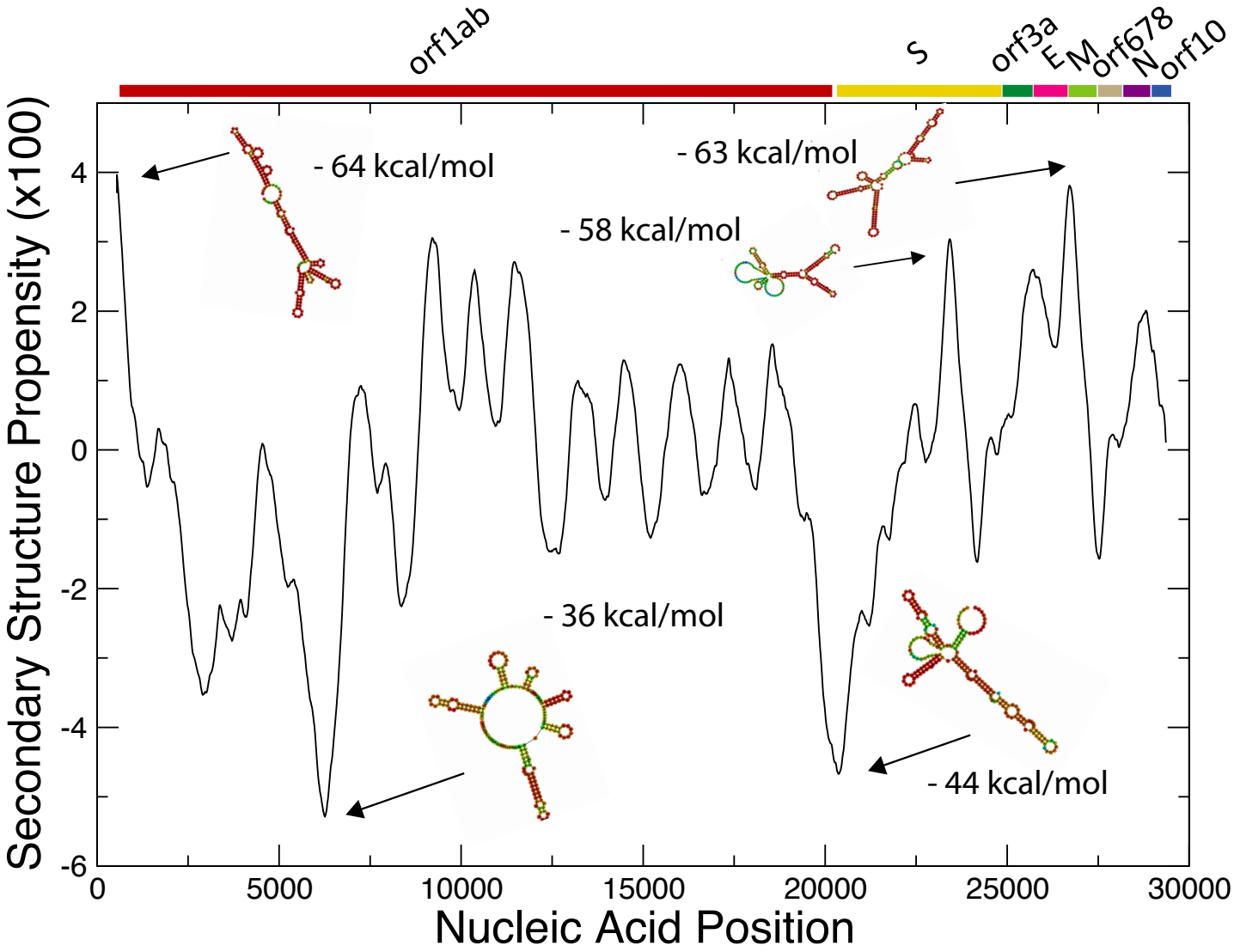

orf1ab
S
orf3a
E
M
orf678
N
orf10
- 64 kcal/mol
- 63 kcal/mol
- 58 kcal/mol
- 36 kcal/mol
- 44 kcal/mol
Secondary Structure Propensity (x100)
4
2
0
-2
-4
-6
0
5000
10000
15000
20000
25000
30000
Nucleic Acid Position

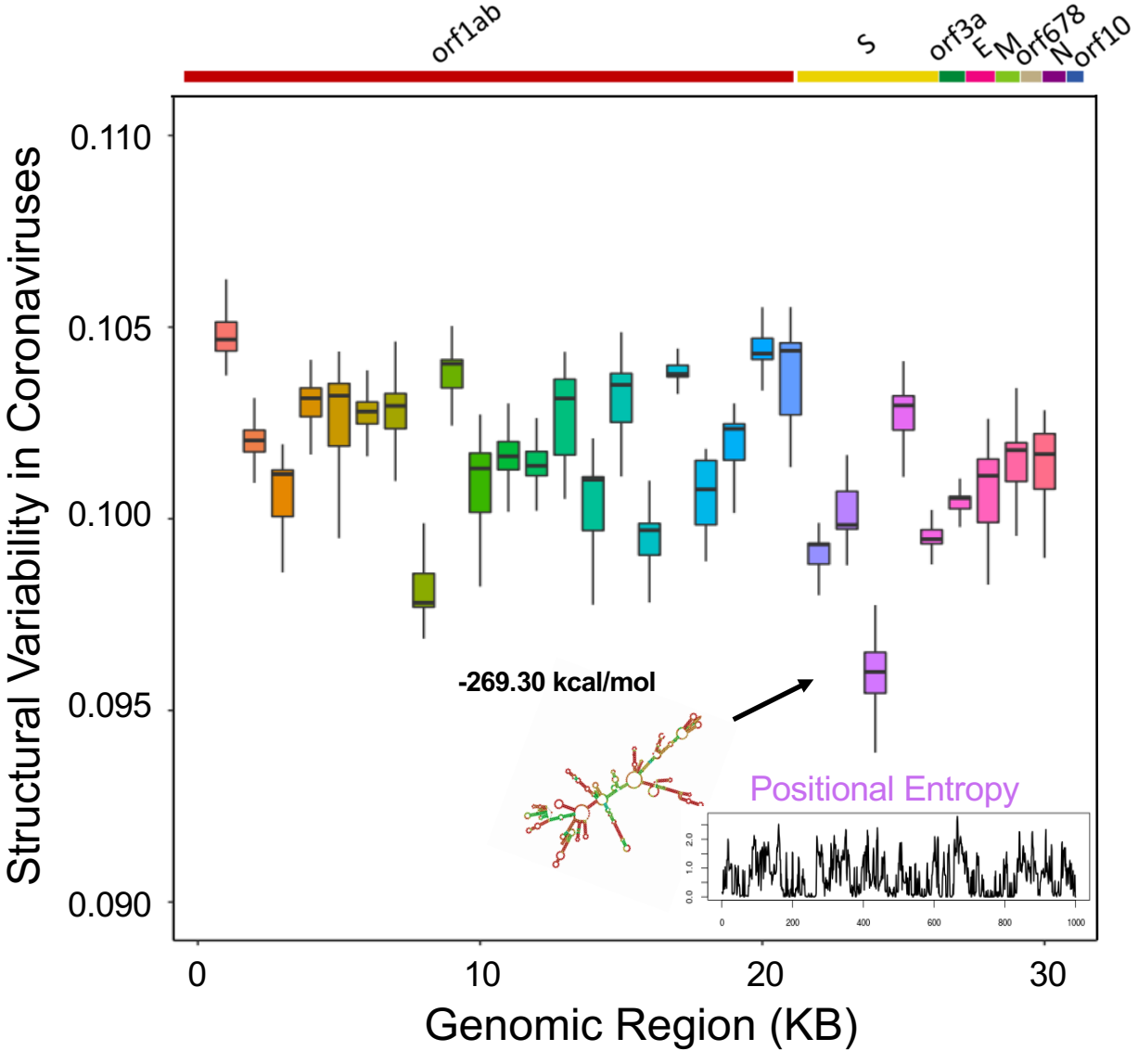

orf1ab
S
orf3a
E
M
orf678
N
orf10
Structural Variability in Coronaviruses
0.110
0.105
0.100
0.095
0.090
-269.30 kcal/mol
Positional Entropy
0
200
400
600
800
1000
0
10
20
30
Genomic Region (KB)

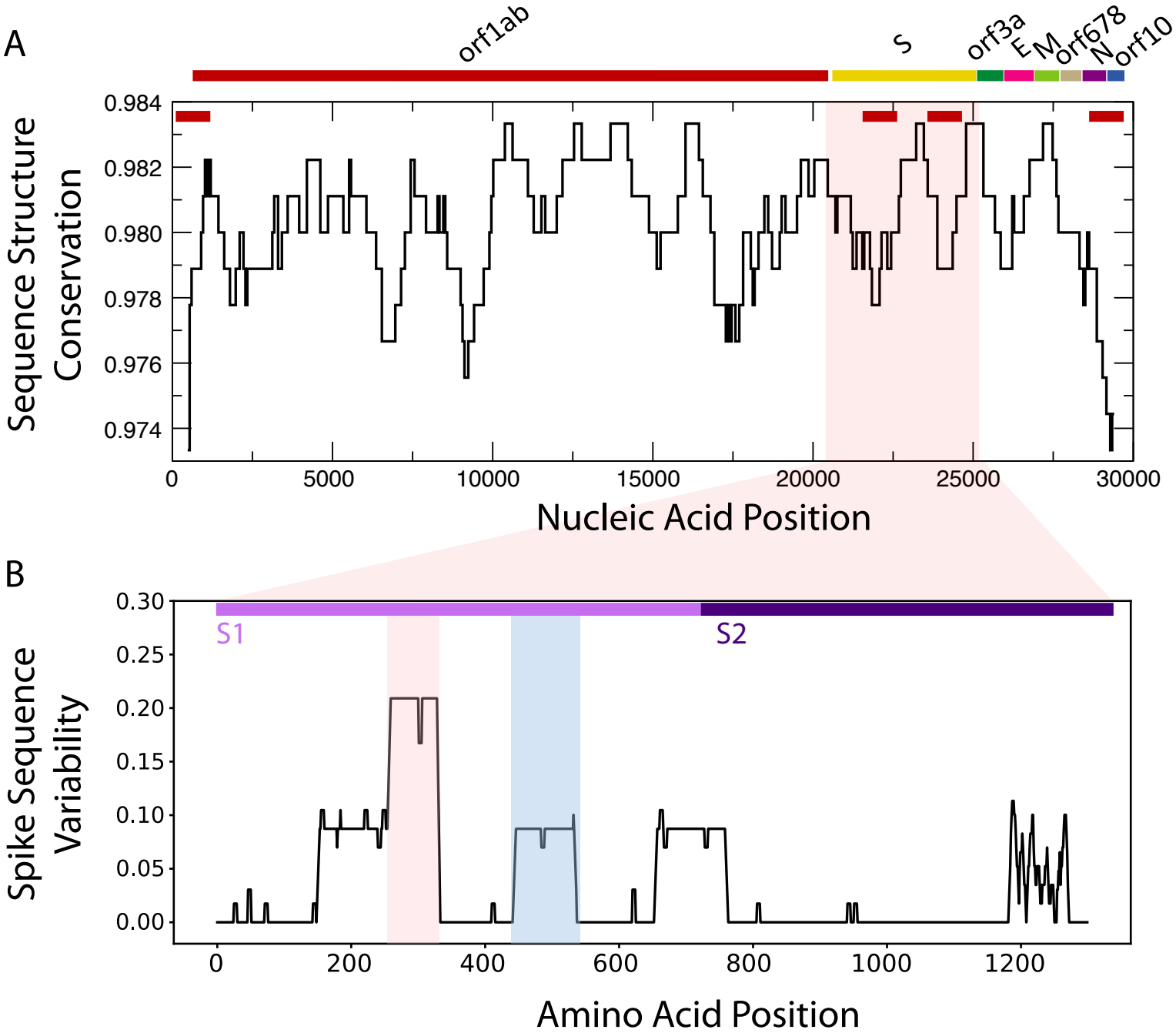

A
orf1ab
S
orf3a
E
M
orf678
N
orf10
Sequence Structure Conservation
0.984
0.982
0.980
0.978
0.976
0.974
0
5000
10000
15000
20000
25000
30000
Nucleic Acid Position
B
S1
S2
Spike Sequence Variability
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
200
400
600
800
1000
1200
Amino Acid Position

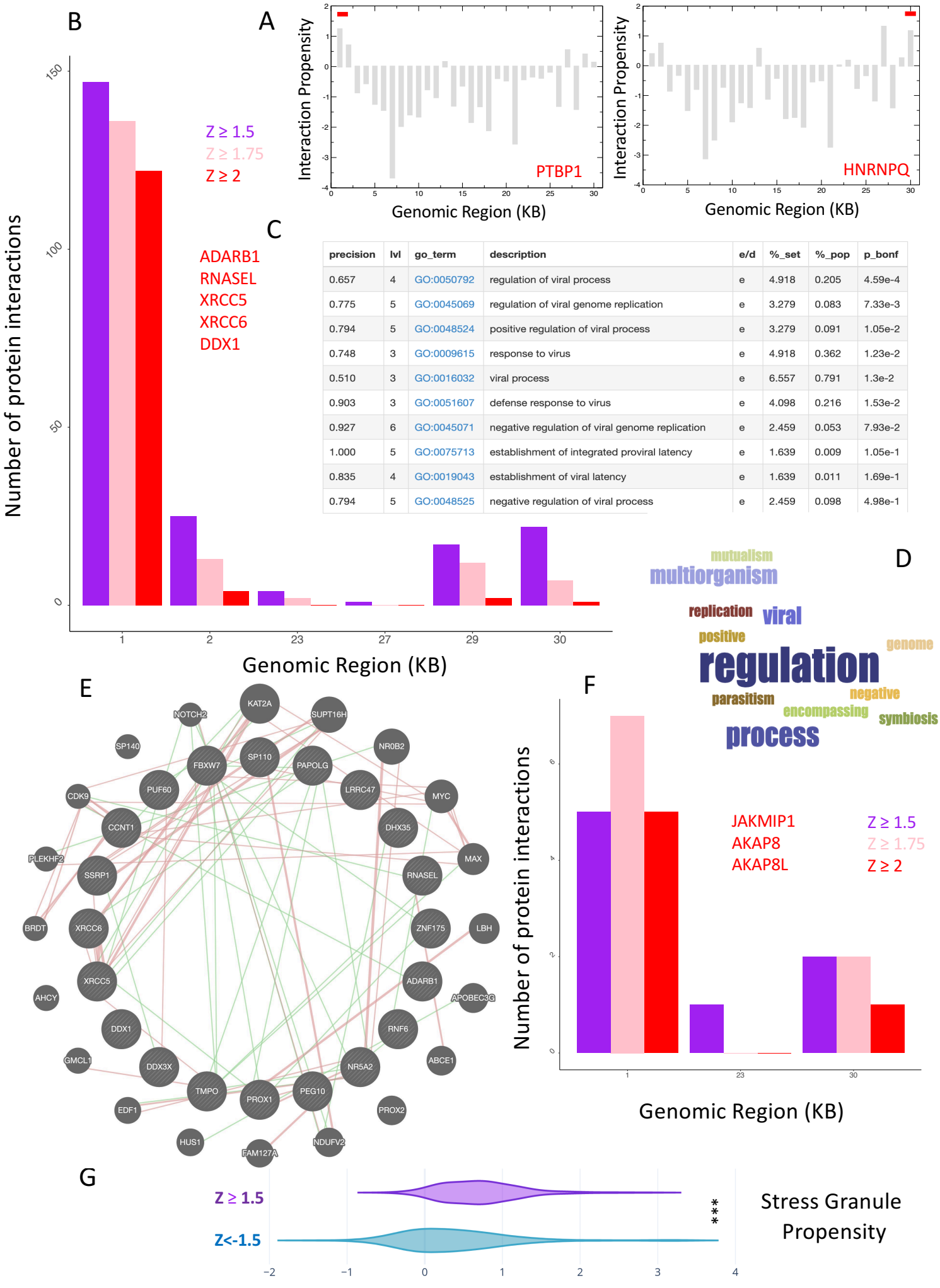


precision	lvl	go_term	description	e/d	%_set	%_pop	p_bonf
0.657	4	GO:0050792	regulation of viral process	e	4.918	0.205	4.59e-4
0.775	5	GO:0045069	regulation of viral genome replication	e	3.279	0.083	7.33e-3
0.794	5	GO:0048524	positive regulation of viral process	e	3.279	0.091	1.05e-2
0.748	3	GO:0009615	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


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
Structural Variability in Coronaviruses
0.110
0.105
0.100
0.095
0.090
-269.30 kcal/mol
Positional Entropy
0
200
400
600
800
1000
0
10
20
30
Genomic Region (KB)


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
S1
S2
Spike Sequence Variability
0.30
0.25
0.20
0.15
0.10
0.05
0.00
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