Whole genome identification of potential G-quadruplexes and analysis of the G-quadruplex binding domain for SARS-CoV-2

Rongxin Zhang[1], Xiao Ke[1], Yu Gu[1], Hongde Liu[1], Xiao Sun[1,*]

1 State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, China

[*] To whom correspondence should be addressed: xsun@seu.edu.cn.

Abstract

The Coronavirus Disease 2019 (COVID-19) pandemic caused by SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) quickly become a global public health emergency. G-quadruplex, one of the non-canonical secondary structures, has shown potential antiviral values. However, little is known about G-quadruplexes on the emerging SARS-CoV-2. Herein, we characterized the potential G-quadruplexes both in the positive and negative-sense viral stands. The identified potential G-quadruplexes exhibits similar features to the G-quadruplexes detected in the human transcriptome. Within some bat and pangolin related beta coronaviruses, the G-quartets rather than the loops are under heightened selective constraints. We also found that the SUD-like sequence is retained in the SARS-CoV-2 genome, while some other coronaviruses that can infect humans are depleted. Further analysis revealed that the SARS-CoV-2 SUD-like sequence is almost conserved among 16,466 SARS-CoV-2 samples. And the SARS-CoV-2 SUD_{core}-like dimer displayed similar electrostatic potential pattern to the SUD dimer. Considering the potential value of G-quadruplexes to serve as targets in antiviral strategy, we hope our fundamental research could provide new insights for the SARS-CoV-2 drug discovery.

Introduction

The COVID-19 pandemic, which first broke out in China, has rapidly become a global public health emergency within a few months[1]. According to the statistics from the Johns Hopkins Coronavirus Resource Center (https://coronavirus.jhu.edu/map.html), 6.6 million cases have been confirmed, with a death toll rising to 389,000. Since 2000, humans have suffered at least three coronavirus outbreaks, and they were Severe Acute Respiratory Syndrome (SARS) in 2003[2-5], Middle East Respiratory Syndrome (MERS) in 2012[2, 5], and COVID-19. Scientists identified and sequenced the virus early in this outbreak, and named it SARS-CoV-2[6]. The symptoms of the patients infected with the novel coronavirus vary from person to person, and fever, cough, and fatigue are the most common ones[7-10]. The clinical chest CT (Computed tomography) and nucleic acid testing are the most typical methods of diagnosing COVID-19[9, 10]. It is worth noting that the recent achievements in AI (Artificial Intelligence) aid diagnosis technology[11] and CRISPR-Cas12-based detection methods[12] are expected to expand the diagnosis of COVID-19. Despite the great efforts of the researchers, so far, no specific clinical drugs or vaccines have been developed to cope with COVID-19.

SARS-CoV-2 is a Betacoronavirus within the Coronaviridae family that is the culprit responsible for the COVID-19 pandemic[13, 14] (Fig. 1A). Studies have confirmed that SARS-CoV-2 is a positive-sense single-stranded RNA ((+)ssRNA) virus with a total length of approximately 30k. The positive-sense RNA strand of SARS-CoV-2 can serve as a template to produce viral proteins related

to replication, structure composition, and other functions or events[15, 16]. One of the hotspots is how the SARS-CoV-2 entry the host cells. SARS-CoV-2 has shown a great affinity to the angiotensin-converting enzyme 2 (ACE2), which has been proved to be the binding receptor for SARS-CoV-2[17, 18]. After entering the host cells, the viral genomic RNA will be released to the cytoplasm, and the ORF1a/ORF1ab are subsequently translated into replicase polyproteins of pp1a/pp1ab, which will be cleaved into some non-structural proteins (nsps). These non-structure proteins ultimately form the replicase-transcriptase complex for replication and transcription. Along with the full-length positive and negative-sense RNAs, a nested set of sub-genomic RNAs (sgRNAs) are also synthesized, and mainly translated into some structural proteins and accessory proteins. When assembly finished, the mature SARS-CoV-2 particles are released from infected host cells via exocytosis[19]. Mounting evidence suggests that bats and pangolins are the suspected natural host and intermediate host of SARS-CoV-2[20-23]. Intriguingly, a report from Yongyi Shen et al. showed that SARS-CoV-2 might be the recombination product of Bat-CoV-RaTG13-like virus and Pangolin-CoV-like virus[24].

G-quadruplexes are the non-canonical nucleic acid structures usually formed in G-rich regions both in DNA and RNA strands[25-27]. The G-quadruplex is formed by stacking G-quartets (Fig. 1B) on top of each other, in which the four guanines making up a G-quartets are connected via Hoogsteen pairs (Fig. 1C)[25-28]. Extensive research indicated that G-quadruplexes were involved in many critical biological processes, including DNA replication[29-32], telomere regulation[33-37], and RNA translation[38-41]. It has been proved that G-quadruplexes existed in the viral genome and can regulate the viral biological processes, which made it possible to function as potential drug targets for antiviral strategy[42-44]. A study made by Jinzhi Tan et al. demonstrated that the SARS-Unique Domain (SUD) within the nsp3 (non-structural protein 3) of SARS coronavirus (SARS-CoV) exhibits the binding preference to the G-quadruplex structure in the human transcript, and potentially interfere with host cell antiviral response[45]. They also identified several amino acid residues that were tightly associated with its binding capacity. Yet, whether the SARS-CoV-2 contains a SUD-like structure and whether the structure sequence is conserved among SARS-CoV-2 samples needs further interpretation. Besides, the G-quadruplexes in some well-known virus, such as HIV-1[46-49] (Human Immunodeficiency Viruses type 1), ZIKV[50] (ZIKA Virus), HPV[51, 52] (Human Papillomavirus) and EBOV[53] (Ebola virus) have been studied. However, our understanding of the G-quadruplexes, their potential roles, and the SUD-like structures in the emerging SARS-CoV-2 are lacking.

In this study, we depicted the potential G-quadruplexes (PG4s) in the SARS-CoV-2 by combining several G-quadruplex prediction tools. The PG4s in SARS-CoV-2 presented similar features to the two-quartet G-quadruplexes in the human transcriptome, which potentially supported the formation and existence of the G-quadruplexes in SARS-CoV-2. Additionally, we investigated the difference in selective constraints between the G-quartets and other nucleotides in the SARS-CoV-2 genome. To further elucidated the possible pathogenic mechanism of SARS-CoV-2, we examined the SUD-like sequence and structure in SARS-CoV-2 that are critical to binding the G-quadruplexes in host transcripts.

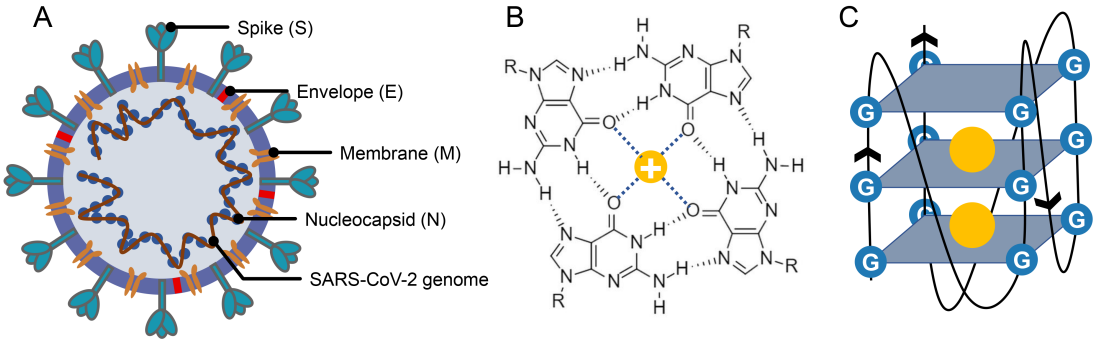

Fig. 1 Structure of SARS-CoV-2, G-quartet, and G-quadruplex. (A) The SARS-CoV-2 particle structure is composed of four structural proteins, which are the spike protein, the envelope protein, the membrane protein, and the nucleocapsid protein. The nucleocapsid proteins are bound to the SARS-CoV-2 genome. (B) Structure of G-quartet, the neighboring guanines are connected via Hoogsteen hydrogen. The cation is indicated by a yellow circle. (C) The G-quadruplex is formed by stacking multiple G-quartets, and the stabilization of the structure is partially determined by the central cations.

Results

Whole genome identification and annotation of potential G-quadruplexes

To get the potential G-quadruplexes in the SARS-CoV-2 genome, we took the strategy described as follows (Fig. 2A): (i) Predicting the PG4s with three software independently. (ii) Merging the prediction results of the PG4s and evaluating the G-quadruplex folding capabilities by the cG/cC scores. (iii) The PG4s with cG/cC scores higher than the threshold were selected as candidates for further analysis. Here, the threshold for determining whether PG4s can be folded was set to 2.05, as described in the study of Jean-Denis Beaudoin et al.[54] In total, we obtained 24 PG4s (Table. 1) in the positive or negative-sense strands for further analysis.

To annotate the PG4s, the reference annotation data (in gff3 format) of SARD-CoV-2 were downloaded from the NCBI database with the accession number of NC_045512. Firstly, we focused on the PG4s on the positive-sense strand. Fifteen of the 24 PG4s (67.5%) were located on the positively-sense strand, the vast majority of them were harbored in non-structural proteins including nsp1, nsp3, nsp4, nsp5, nsp10 and nsp14, with the remaining ones located in the spike protein, orf3a, and the membrane protein. Secondly, we examined the PG4s on the negative-sense strand, which is an intermediate product of replication. Nine PG4s were scattered on the negative-sense strand.

To further characterize the potential canonical secondary structures competitive with G-quadruplexes, the landscape of thermodynamic stability of the SARS-CoV-2 genome was depicted by using ΔG° z-score[55]. In general, a positive ΔG° z-score implies that the secondary structure of this region tends to be less stable than the randomly shuffled sequence with the identical nucleotide composition, while a negative ΔG° z-score signifies higher stability than the randomly shuffled sequence. For each nucleotide in the SARS-CoV-2 genome, the ΔG° z-score was calculated for all the 120 nt windows covering the nucleotide, and an average ΔG° z-score was deduced then. Several PG4s are located in positions with a locally higher average ΔG° z-scores (Fig. 2B) which implied the relative instability of a canonical secondary structure and the lower possibility to adopt such a competitive structure against the G-quadruplex, which may ultimately favor the formation of G-quadruplex.

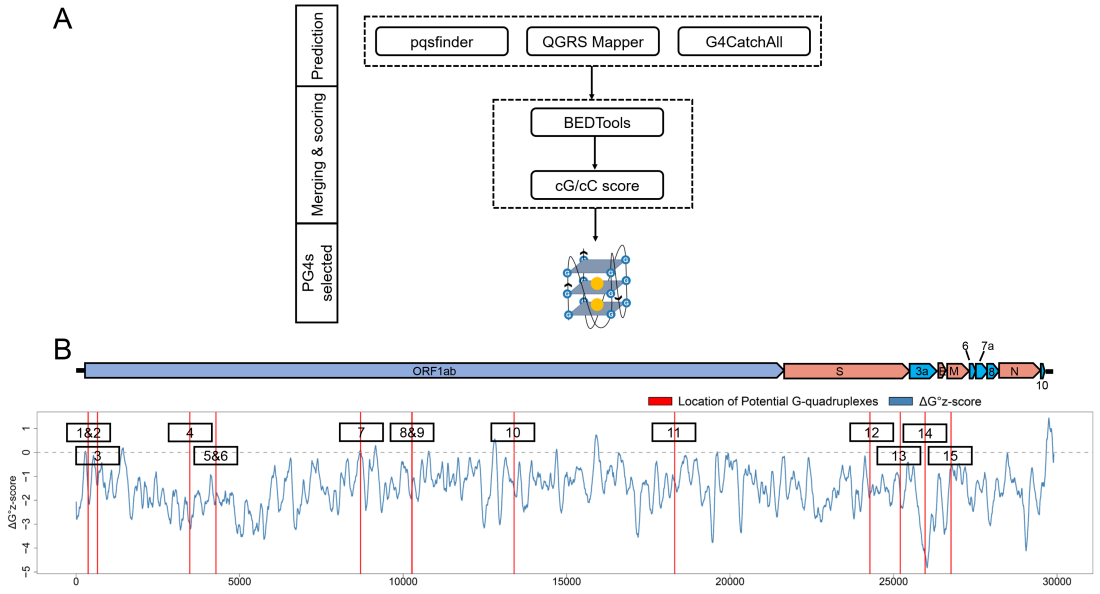

Fig. 2 Detection and annotation of the PG4s. (A) The schematic flow of PG4s detection. The G-quadruplex prediction tools, pqsfinder, QGRS Mapper, and G4CatchAll were utilized for the prediction of PG4s. BEDTools and the cG/cC scoring system were applied to merging and scoring the PG4s. After screening, the PG4s that used in this study were generated. (B) Visualization of the PG4s in the SARS-CoV-2 genome. Top panel, the genome organization of SARS-CoV-2. Bottom panel, the average Δ G° z-score for each nucleotide (blue curve) in the SARS-CoV-2 genome, the location of PG4s are plotted with red vertical lines. The order of the PG4s is marked with a black box. Please note that only the PG4s in the positive-sense strand are visualized.

Table. 1
The PG4s found in the SARS-CoV-2.

No.	start	end	strand	sequence (5'→3')	annotation
1	15	37	-	GGTTGGTTTGTTACCTGGGAAGG	-
2	353	377	+	GGCTTTGGAGACTCCGTGGAGGAG G	nsp1
3	359	377	+	GGAGACTCCGTGGAGGAGG	nsp1
4	644	663	+	GGTAATAAAGGAGCTGGTGG	nsp1
5	2449	2472	-	GGGGCTTTTAGAGGCATGAGTAGG	-
6	3467	3483	+	GGAGGAGGTGTTGCAGG	nsp3
7	4261	4289	+	GGGTTTAAATGGTTACACTGTAGAG GAGG	nsp3
8	4262	4289	+	GGTTTAAATGGTTACACTGTAGAGG AGG	nsp3
9	4886	4901	-	GGTGGAATGTGGTAGG	-
10	6011	6027	-	GGATATGGTTGGTTTGG	-
11	8687	8709	+	GGATACAAGGCTATTGATGGTGG	nsp4
12	10015	10030	-	GGTTTGTGGTGGTTGG	-
13	10015	10039	-	GGTGATAGAGGTTTGTGGTGGTTGG	-

14	10019	10039	-	GGTGATAGAGGTTTGTGGTGG	-
15	10255	10282	+	GGTACAGGCTGGTAATGTTCAACTCAGG	nsp5
16	10261	10290	+	GGCTGGTAATGTTCAACTCAGGGTTATTGG	nsp5
17	13385	13404	+	GGTATGTGGAAAGGTTATGG	nsp10
18	15924	15941	-	GGATCTGGGTAAGGAAGG	-
19	18296	18318	+	GGATTGGCTTCGATGTCGAGGGG	nsp14
20	24268	24291	+	GGCTTATAGGTTTAATGGTATTGG	S-S2
21	25197	25218	+	GGCCATGGTACATTTGGCTAGG	S-S2
22	25951	25979	+	GGTGGTTATACTGAAAAATGGGAATCTGG	ORF3a
23	26746	26775	+	GGATCACCGGTGGAATTGCTATCGCAATGG	M
24	26889	26917	-	GGTCTGGTCAGAATAGTGCCATGGAGTGG	-

Potential G-quadruplexes in SARS-CoV-2 show analogical features with the rG4s in the human transcriptome

In 2016, Chun Kit Kwok and co-workers profiled the RNA G-quadruplexes in the HeLa transcriptome by using the RNA G-quadruplex sequencing (rG4-seq) technology, and quantified the diversity of these RNA G-quadruplexes[56]. We set out to address the question of whether the potential G-quadruplexes in SARS-CoV-2 showed analogical features with the G-quadruplexes found in the human transcriptome and if these PG4s have the ability to form G-quadruplex structures. We noticed that the PG4s in SARS-CoV-2 are all in the two-quartet style. Therefore we retrieved the two-quartet RNA G-quadruplex sequence data generated in the rG4-seq experiment under the condition of K^+ and pyrdiostatin (PDS). However, for some RTS (Reverse Transcriptase Stalling) sites labeled as two-quartet, there may exist overlapping G-quadruplexes with different loops (e.g., GGCACAGCAGGCATCGGAGGTGAGGCGGGG), and it is difficult to determine which one was formed in the experiment. In order to eliminate the ambiguity, only the RTS sites containing non-overlapping two-quartet G-quadruplex (e.g., GTCATTTTTGTGTTTGGTTTGGTGGTGGC) were considered.

Firstly, we investigated the loop length distribution pattern of the two-quartet PG4s in both SARS-CoV-2 and the human transcriptome (Fig. 3A). As a whole, the two-quartet PG4s in SARS-CoV-2 and the human transcriptome displayed similar loop length distribution patterns, and the loop length of the PG4s in SARS-CoV-2 falls into the scope of the ones from the human transcriptome. The distributions of loop length between the SARS-CoV-2 PG4s and the human two-quartet G-quadruplexes did not show discrepancies (Fig. S1, Wilcoxon test, *p*-value = 0.4552).

Considering the fact that the presence of multiple-cytosine tracks may hinder the formation of G-quadruplexes[54, 57], we examined the cytosine ratio in G-quadruplex loops (Fig. 3B). No significant difference in loop cytosine ratios was observed between the SARS-CoV-2 PG4s and the human two-quartet G-quadruplexes (Wilcoxon test, *p*-value = 0.9911), which suggested that the loop cytosine ratios between the two types of G-quadruplex were similar.

Taken together, our results suggested that the PG4s in SARS-CoV-2 displayed similar features to the rG4s in the human transcriptome.

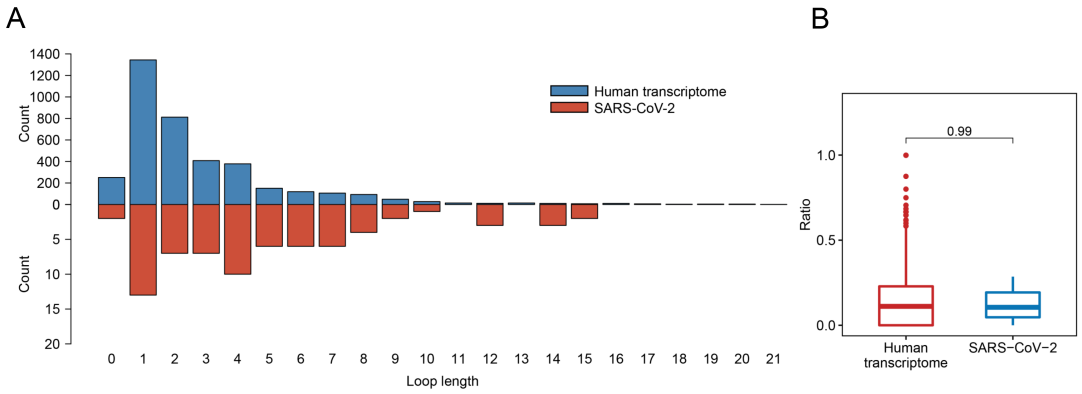

Fig. 3 Feature comparison of potential G-quadruplexes found in rG4-seq and SARS-CoV-2. (A) The Histogram represents the number of two-quartet G-quadruplex loops with different lengths in the human transcriptome and SARS-CoV-2, respectively. (B) The red boxplot shows the ratio of cytosine in human transcriptome two-quartet G-quadruplex loops. And the blue boxplot displays the ratio of cytosine in SARS-CoV-2 PG4 loops.

Potential G-quadruplexes are under selective constraints in bat and pangolin related beta coronaviruses

Recent research revealed that the G-quadruplexes in human UTRs (Untranslated Regions) are under selective pressures[58], and some coronaviruses on bats and pangolins are closely related to SARS-CoV-2. The conservation of the potential G-quadruplexes in the SARS-CoV-2 genome under selective constraints were analyzed. We collected some beta coronavirus genomic sequences of bats and pangolins from several public databases and used the NJ (Neighbor-Joining) method to construct the phylogenetic tree with 1,000 bootstrap replications (Fig. S2). The RS (Rejected Substitutions) score for each site in the SARS-CoV-2 reference genome was evaluated by using the GERP++ software.

We checked the RS score difference between the G-tract (continuous runs of G) nucleotides and other nucleotides. A significant discrepancy was observed, which means that the G-tracts nucleotides exhibit heightened selective constraints than other nucleotides in the SARS-CoV-2 genome (Fig. 4A, Wilcoxon test, p-value = 9.254×10^{-8}). Considering that the G-tracts are composed of guanines, the conservation of guanines in and outside the G-tracts in the SARS-CoV-2 genome were also compared. We found that the guanines in G-tracts are under heightened selective constraints (Fig. 4B, Wilcoxon test, p-value = 3.363×10^{-3}). The nucleotides within G-tracts are more relevant to the G-quadruplexes structural maintenance than loops. Then we compared the G-tract and loop RS scores. As expected, the G-tract RS scores were significantly higher than loops (Fig. 4C, Wilcoxon test, p-value = 3.962×10^{-7}), which suggests that the G-tracts experienced stronger selective constraints.

We also checked that if the PG4s that are under heightened selective constraints is relevant to its inherent properties or potential functions rather than the sequence contexts. A random test was performed to check whether the fragments containing PG4s manifested different average RS scores compared with random fragments in the SARS-CoV-2 genome. The fragments containing PG4 were

designated as the sequence 100 nt upstream and downstream of the PG4 centers. We conducted 1,000 rounds of tests. In each test, we randomly selected 50 fragments from the SARS-CoV-2 genome with a length of 200 nt and carried out the Wilcoxon test to assess the average RS score difference among the randomly selected fragments and the fragments containing PG4s. The *p*-value for each round was retained. As a result, no evident difference was observed as few *p*-values (13/1000) were less than 0.05 (Fig. 4D), suggesting that PG4s that are under heightened selective constraints is more likely to be related to its inherent properties or potential functions rather than sequence contexts.

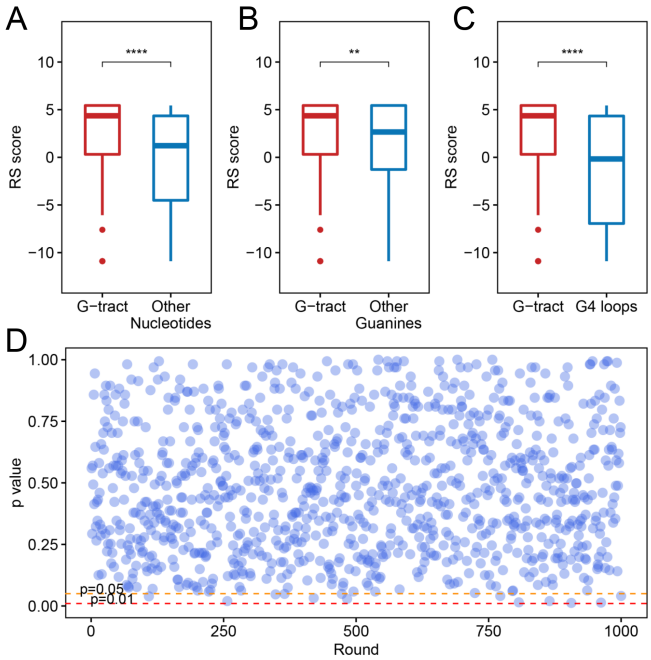

Fig. 4 Potential G-quadruplexes exhibit heightened selective constraints in bat and pangolin related coronavirus. (A-C) Boxplot showing the difference of nucleotide RS scores in G-tract, other nucleotides, other guanines, and PG4 loops (**$p \leq 0.01$, ****$p \leq 0.0001$). (D) Tests of the RS score difference between the fragments containing PG4s and randomly selected fragments. The abscissa indicates the round of the test, while the ordinate represents the *p*-value for each round.

SARS-CoV-2 contains similar SUD to SARS-CoV

Both SARS-CoV and SARS-CoV-2 could cause acute disease symptoms, and the above coronavirus shares similar nucleic acid sequence compositions. There is a SUD in the SARS-CoV genome that can binding to the G-quadruplex structures and it is unclear if the SARS-CoV-2 genome possess the resemble structure. Thus, we started to explore whether the SARS-CoV-2 genome contains the protein-coding sequence similar to SUD and whether SARS-CoV-2 retains the ability to bind RNA G-quadruplexes. We collected the ORF1ab amino acid sequences of some coronaviruses, including seven known coronaviruses, which can infect humans and other coronaviruses belonging to different genera. Surprisingly, the SUD protein sequence is absent in some coronaviruses, especially in alpha, gamma, and delta coronaviruses (Fig. S3). In contrast, the SUD protein sequence is retained in several beta-coronavirus, particularly in bat and pangolin associated beta coronavirus. Moreover, among the seven coronaviruses that can infect humans, only SRAS-CoV and SARS-CoV-2 keep the SUD sequence, while the SUD sequence in MERS-CoV, HCoV-229E, HCoV-NL63, HCoV-OC43

and HCoV-HKU1 is depleted. Next, we examined eight key amino acid residues in SUD that previously reported associated with G-quadruplex binding affinity (Fig. 5A). Almost all the key amino acid residues are reserved in SARS-CoV-2, except one conservative replacement of K (Lysine) > R (Arginine). We hypothesized that if the G-quadruplex binding ability is essential for the SARS-CoV-2, the above amino acid residues should be conservative. We then investigated the conservation of the eight amino acid residues within SARS-CoV-2 samples. We retrieved the sequence alignment file of 16,466 SARS-CoV-2 samples from the GISAID database and calculated the mutation frequency for each nucleotide. We observed the frequency of nucleotide mutations in the above eight codons. As a result, a limited mutation frequency was found as compared to the whole genome average mutation frequency (Fig. 5B, frequency = 3.96). Although eight mutations were detected in glutamate (2432 E), seven of them were synonymous mutations. Next, we checked the electrostatic potential pattern in the SARS-CoV-2 SUD_{core}-like dimer structure. The SARS-CoV-2 SUD_{core}-like dimer structure is defined as the dimer structure formed by the amino acid residues in SARS-CoV-2 corresponding to the SUD of SARS. We found that the SUD_{core}-like dimer of SARS-CoV-2 and the SUD_{core} of SARS present analogical electrostatic potential patterns. The positively charged patches were observed in the core of the SUD_{core}-like dimer, which was surrounded by negatively charged patches (Fig. 5C). In contrast, when the dimer is rotated 180°, a slightly inclined narrow cleft with negative potential accompanied by the positively charged patches was discovered (Fig. 5D). And the above patterns also appeared in the SUD dimer. In the previous reports, several positively charged patches located in the center and back of the dimer were presumed to bind the G-quadruplex structures. By comparison with the electrostatic potential of the SARS SUD_{core} dimer, we identified the positively charged patches located in the center and back of the SARS-CoV-2 SUD_{core}-like dimer, which can potentially bind the G-quadruplexes (Fig. 5C-D).

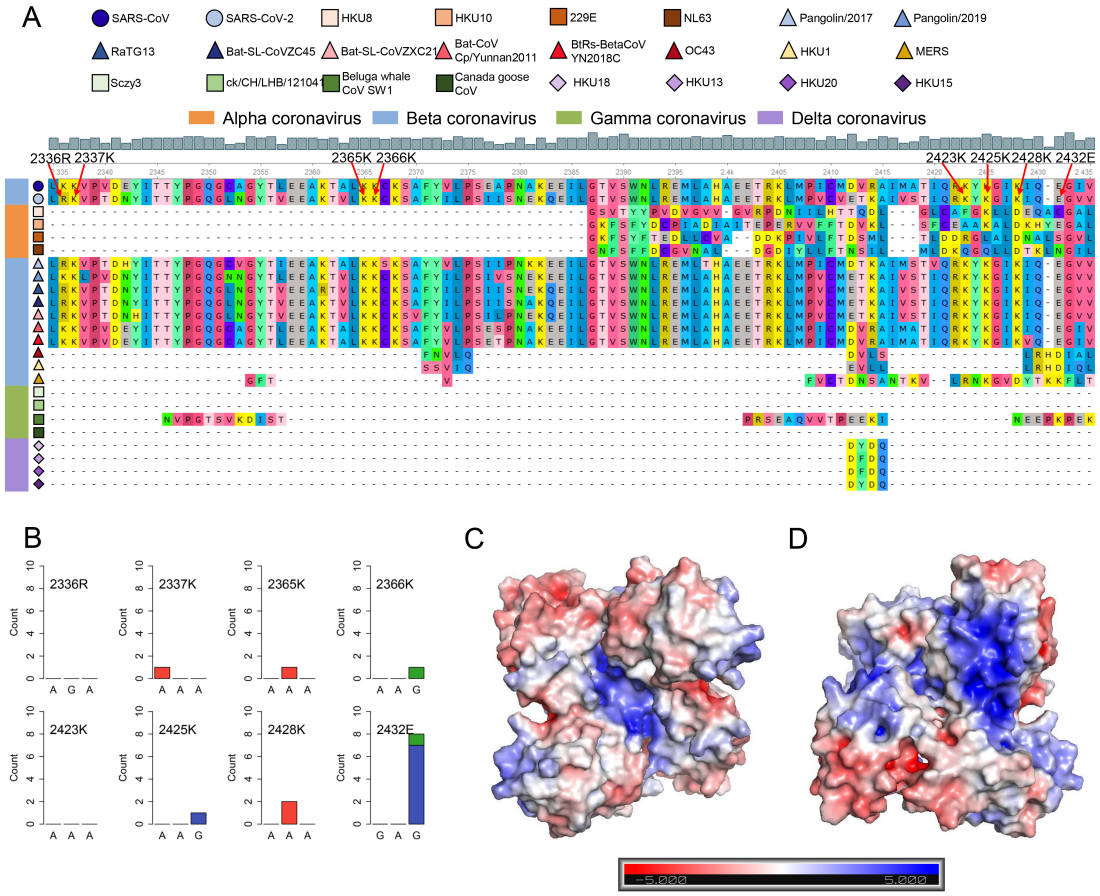

Fig. 5 SARS-CoV-2 contains the SUD_{core}-like sequence and dimer structure. (A) Sequence alignment of four different genera coronaviruses. The shapes in various colors mark different kinds of coronaviruses. The color bar represents different genera of coronaviruses (orange, alpha coronavirus; blue, beta coronavirus; green, gamma coronavirus; purple, delta coronavirus). The grey histogram shows the consensus of the alignment sites. The eight amino acid residues related to the G-quadruplex binding affinity are labeled by red arrows. (B) Nucleotide mutation count of eight amino acid residues among 16,466 SARS-CoV-2 samples. (C-D) The electrostatic potential surface of the SARS-CoV-2 SUDcore-like dimer with different orientations (the front (C) and rotate 180° (D)). The blue and red showed positive and negative potential, respectively.

Discussion

The COVID-19 pandemic has caused huge losses to humans and made people pay more attention to public health. A large number of scientists all over the world have been engaged in the fight against the outbreak. The SARS-CoV-2 coronavirus is the key culprit responsible for the outbreak, and no specific inhibitor drugs have been developed yet. G-quadruplexes have shown tremendous potential for the development of anticancer[59-62] and antiviral drugs[44, 63, 64], as G-quadruplexes can interfere with many biological processes that are critical to cancer cells and viruses. Therefore, it is necessary to quantify and characterize the PG4s in the SARS-CoV-2 genome to provide a possible novel method for the treatment of COVID-19.

In this study, besides three popular G-quadruplexes prediction tools, the cG/cC scoring system, which is specially designed for the identification of RNA G-quadruplexes, was adopted to determine the PG4s. Indeed, we did not find the G-quadruplexes with three or more G-quartets, which are generally considered to be more stable than the two-quartet G-quadruplexes. One of the controversial issues lies on the stability of the two-quartet G-quadruplexes, especially the folding capability of those G-quadruplexes *in vivo*. However, it is well-acknowledged that the RNA G-quadruplexes is more stable than their DNA counterparts[65, 66] and SARS-CoV-2 is a single-strand RNA virus, which may be conducive to its structure formation. Several emerging studies have demonstrated the formation of two-quartet G-quadruplexes in viral sequences, which could serve as antiviral elements under the presence of G-quadruplex ligands[53, 67, 68]. Moreover, the K^+ (potassium ion), one of the primary positive ions inside human cells, can strongly support the formation of G-quadruplexes. Nevertheless, whether the SARS-CoV-2 G-quadruplexes could form *in vivo* requires overwhelming proofs.

Most of the PG4s we detected were located in the positive-sense strand. The G-quadruplex forming sequences in the SARS-CoV genome were presumed to function as the chaperones of SUD, and their interaction was essential for the SARS-CoV genome replication[69]. ORF1ab that encodes the replicase proteins is required for the viral replication and transcription. Some PG4s were found to harbored in ORF1ab, and whether these PG4s were related to the replication of the viral genome and interact with SUD-like structures like in SARS-CoV, is worthy of further investigation. In addition to ORF1ab, there exists several PG4s in the structural and accessory protein-coding sequences as well as the sgRNAs that containing the above protein sequences. Some studies have characterized the impact of G-quadruplex structures on the translation of human transcripts, and an apparent inhibitory effect was observed [38, 57, 70]. The translation of some SARS-CoV-2 proteins requires the involvement of human ribosomes; thus, it is possible to repress the translation of SARS-CoV-2 proteins via stabilizing the G-quadruplex structures. In fact, this inhibition effect has been reported in some other viral studies [67, 71]. The negative-sense strand serves as templates for the synthesis of the positive-sense strand and the sub-genomic RNAs. The identified potential G-quadruplexes were broadly distributed in the negative-sense strand. Notably, we observed one PG4 located at the 3' end of the negative-sense strand. A previous study confirmed that the stable G-quadruplex structures located at the 3' end of the negative-sense strand could inhibit the RNA synthesis by reducing the activity of the RdRp (RNA-dependent RNA polymerase) [72]. Therefore, it is necessary to further investigate whether the PG4 at the 3' end of the negative-sense strand of SARS-CoV-2 could inhibit RNA synthesis. In addition, recent research revealed that the high-frequency trinucleotide mutations (G28881A, G2882A and G28883C) were detected in the SARS-CoV-2 genome [73, 74]. G28881A and G28882A always co-occur within the same codon, which means a positive selection of amino acid [75]. We noticed that the trinucleotide mutations were in the G-rich sequence from 28881 nt to 28917 nt (5' GGGGAACTTCTCCTGCTAGAAT**GGCTGGCAATGGCGG** 3'). The potential G-quadruplex downstream of the trinucleotide mutations was filtered by the cG/cC score system as the presence of cytosine tracks within and flanking of the potential G-quadruplex reduce the cG/cC score; however, in fact, this potential G-quadruplex showed a relative lower MFE (Minimum Free Energy) among all the potential G-quadruplexes we detected. The consequence of the trinucleotide mutations was still elusive. Whether the mutations have an internal causality with the G-rich sequence still needs to be elucidated.

The SUD in SARS, which is thought to be related to its terrible pathogenicity, has displayed binding preference to the G-quadruplexes in human transcripts[45]. Our analysis revealed that the novel coronavirus SARS-CoV-2 contained a similar domain to SUD as well. Furthermore, several amino acid residues previously reported to be an indispensable part of the G-quadruplexes binding capability are retained in SARS-CoV-2. Further exploration indicated that the eight key amino acid residues were conserved in numerous SARS-CoV-2 samples across countries all over the world, suggesting the essentiality of the above residues. It is supposed that the binding of SUD to G-quadruplexes could affect transcripts stability and translation, hence impairing the immune response of host cells. The expression of host genes in SARS-CoV-2 infected cells is extremely inhibited[15]; therefore, we speculate that the SARS-CoV-2 may possess the similar mechanism with SARS-CoV that can inhibit the expression of some important immune-related genes to escape immune defense. Herein, we briefly depict the possible role of G-quadruplexes in the antiviral mechanism and pathogenicity, and the development of certain G-quadruplex specific ligands might be a promising antiviral strategy (Fig. 6). We call for more researchers to shed light on the relationship between G-quadruplexes and coronaviruses. Only if we have a deeper understanding of coronaviruses can we better cope with the possible novel coronavirus pandemics in the future.

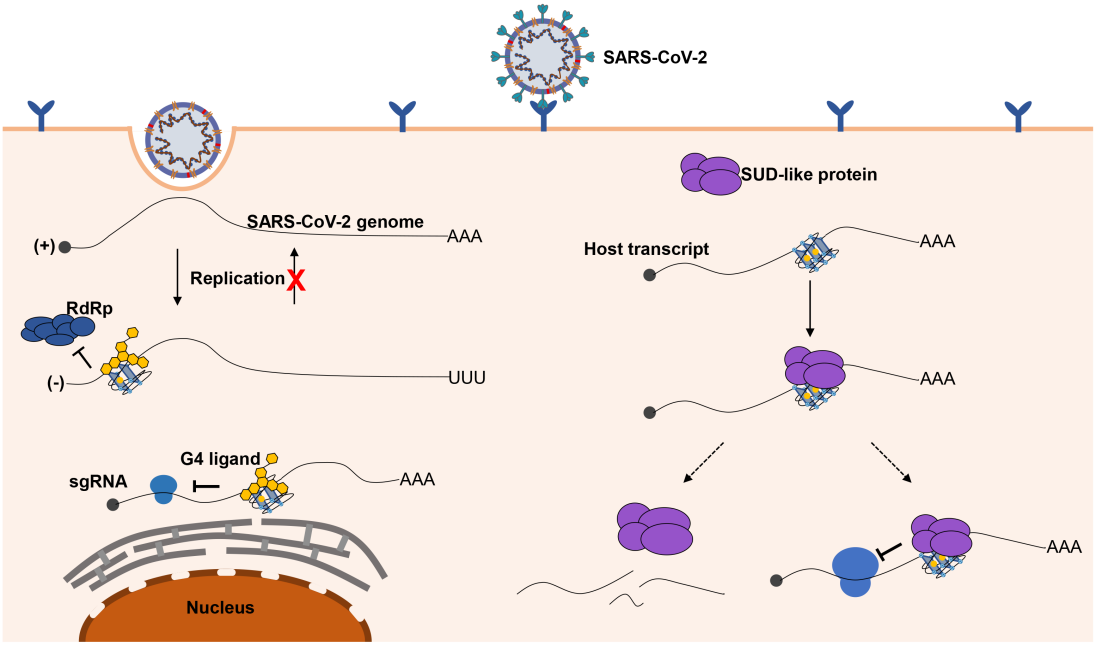

Fig. 6 Possible role of G-quadruplexes in the antiviral mechanism and pathogenicity. Left part, G-quadruplexes can function as inhibition elements in the SARS-CoV-2 life cycle. Both the replication and translation could be affected by the G-quadruplexes structures. The stable G-quadruplexes in the 3' end of the negative-sense strand may interfere with the activity of RdRp; hence, the replication of the negative-sense strands to the positive-sense strands is repressed, so that the SARS-CoV-2 genomes cannot be produced in large quantities. The G-quadruplex structures can suppress the translation process by impairing the elongating of ribosomes, which can hinder the production of proteins required for the virus. The G-quadruplex structures could be stabilized by the specific ligands to enhance the inhibitory effects, which is a promising antiviral strategy. Right part, a possible mechanism for SARS-CoV-2 to impede the expression of human genes. G-quadruplex structures, particularly with longer G-stretches, are the potential binding targets for SUD-like

proteins. And the interaction of the SUD-like proteins with G-quadruplex structures possibly lead to the instability of host transcripts or obstructing the translation efficiency.

Methods

Data collection

We obtained a total of 77 full-length bat-associated beta coronaviruses from the DBatVir (http://www.mgc.ac.cn/DBatVir/) database[76]. We also downloaded the bat coronavirus RaTG13 genome from the NCBI virus database (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/), which has shown a high sequence similarity to the SARS-CoV-2 reference genome in previous reports. We acquired the SARS-CoV-2 reference genome from the NCBI virus database under the accession number of NC_045512. In addition to those sequences, nine pangolin coronaviruses were derived from GISAID (https://www.gisaid.org/) database[77].

Pairwise and multiple sequence alignment, phylogenetic and conservation analysis

The EMBOSS Needle software, which is based on the Needleman-Wunsch algorithm and, is a part of the EMBL-EBI web tools[78], was employed for the pairwise sequence alignment. Clustal Omega[79, 80] is a reliable and accurate multiple sequence alignment (MSA) tool that can be performed on large data sets. We choose this MSA tool for the alignment of viral genomes and the alignment of protein sequences under the default paraments. UGENE[81] is a powerful and user-friendly bioinformatics software, and we choose UGENE to visualize the pairwise and multiple sequence alignment results. We used the MEGA X software[82] to construct the Neighbor-Joining phylogenetic tree with 1,000 bootstrap replications. To depict the conservation state for each nucleotide site, the GERP++ software[83] was applied to calculate the "Rejected Substitutions" score column by column, which can reflect the constraints strength for each nucleotide sites.

Potential G-quadruplex detection

Several open-source G-quadruplex detection software was used to search the PG4s both in the SARS-CoV-2 positive-sense and negative-sense strands. G4CatchAll[84], pqsfinder[85], and QGRS Mapper[86] were employed to predict the putative G-quadruplexes, respectively; Please see ref[87] for more information about the comparison of those tools mentioned above. The minimum G-tract length was set to two in the three software, while the max length of the predicted G-quadruplexes was limited to 30. Specifically, the minimum score of the predicted G-quadruplex was set to 10 when using pqsfinder. We utilized BEDTools[88] to sort the PG4s according to their coordinates. Apart from this, we adopted the cG/cC scoring system[54] proposed by Jean-Pierre Perreault et al. to delineate the sequence context influence on PG4s. The PG4s along with 15 nt upstream and downstream sequence contexts were used to calculate cG/cC score, and 2.05 was taken as the threshold for the preliminary inference of the G-quadruplex folding capability[54]. Using a customized python script, we implemented the cG/cC scoring system.

Homo-dimer homology modeling and electrostatic potential calculation

The SARS-CoV-2 SUD_{core}-like homo-dimer structure was modeled based on the template of the SARS-CoV SUD structure (PDB ID: 2W2G) through homology modeling. All the modeling process were performed in the Swiss Model[89] website (https://swissmodel.expasy.org/). The electrostatic

potential was calculated and visualized in the PyMOL software by using the APBS (Adaptive Poisson-Boltzmann Solver) plugin.

ΔG° z-score analysis

The ΔG° z-score for the SARS-CoV-2 genome was retrieved from RNAStructuromeDB (https://structurome.bb.iastate.edu/sars-cov-2). The ΔG° z-score is described as follows.

$$\Delta G^\circ\ z-score = \frac{(MFE_{native} - \overline{MFE}_{random})}{\sigma} \quad (1)$$

Where the MFE_{native} means the MFE (minimum free energy) ΔG° value predicted by the RNAfold software with a window of 120 nt and step of one nt. And the $\overline{MFE}_{random}$ represents the MFE ΔG° value generated by the randomly shuffled sequence with the identical nucleotide composition. The σ is the standard deviation across all the MFE values.

To depict the ΔG° z-score for each nucleotide in the SARS-CoV-2 genome, we utilized the following formula.

$$z_i = \frac{\sum_{m=1}^{w} \Delta G^\circ\ z-score_m}{w} \quad (2)$$

Where z_i is the average ΔG° z-score for nucleotide i, w denotes the total number of the sliding windows that covering the nucleotide i. $\Delta G^\circ\ z-score_m$ indicates the ΔG° z-score for the m-th window. For example, when considering the nucleotide 1000 under the setting of 120 nt window length and one nt step, there are 120 sliding windows covering the nucleotide 1000. So, the z_{200}, which means the average ΔG° z-score for nucleotide 200, is calculated as the sum of the ΔG° z-score of 120 sliding windows divided by the total number of the sliding windows.

Funding

This work was supported by the National Natural Science Foundation of China (61972084).

Supplementary

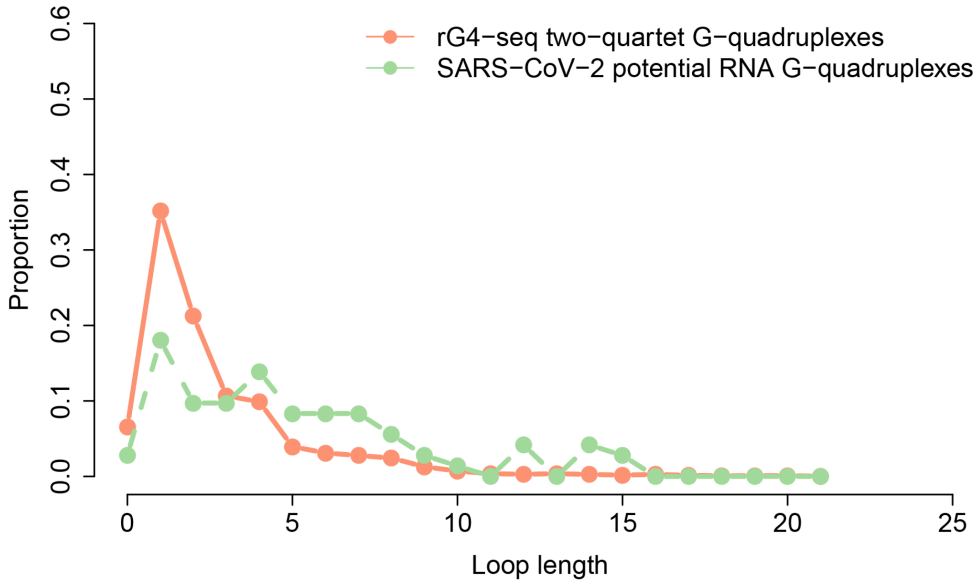

Fig. S1 Proportion of loops with different lengths. The x-axis and y-axis show the loop length and loop proportion, respectively. The blue curve represents the loop in potential G-quadruplexes, while the red curve indicates the loop in two-quartet G-quadruplexes derived from rG4-seq.

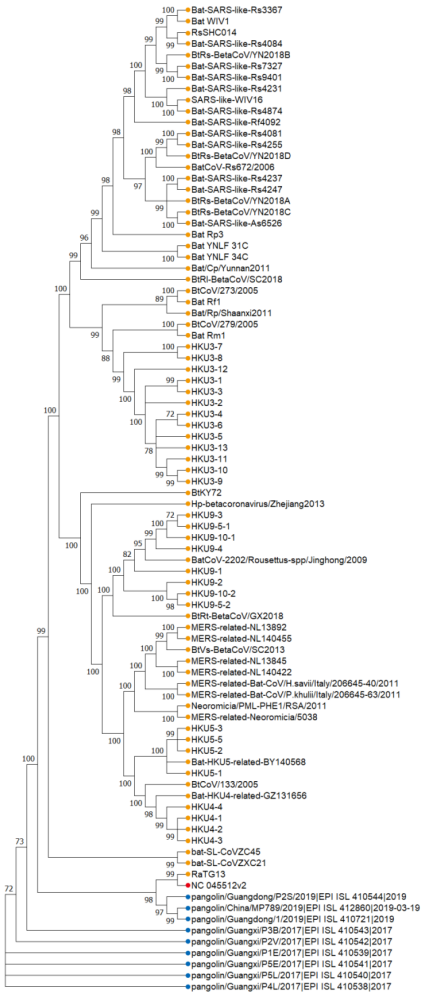

Fig. S2 Phylogenetic tree of bat and pangolin related beta-coronavirus. The phylogenetic tree was constructed using the Neighbor-Joining method with 1,000 bootstrap replications. The bootstrap values lower than 70 were removed from the phylogenetic tree nodes. The SARS-CoV-2 reference sample is marked in the red dot, and the orange dots indicate the bat-related beta coronavirus samples, while the blue dots represent pangolin related beta-coronavirus samples.

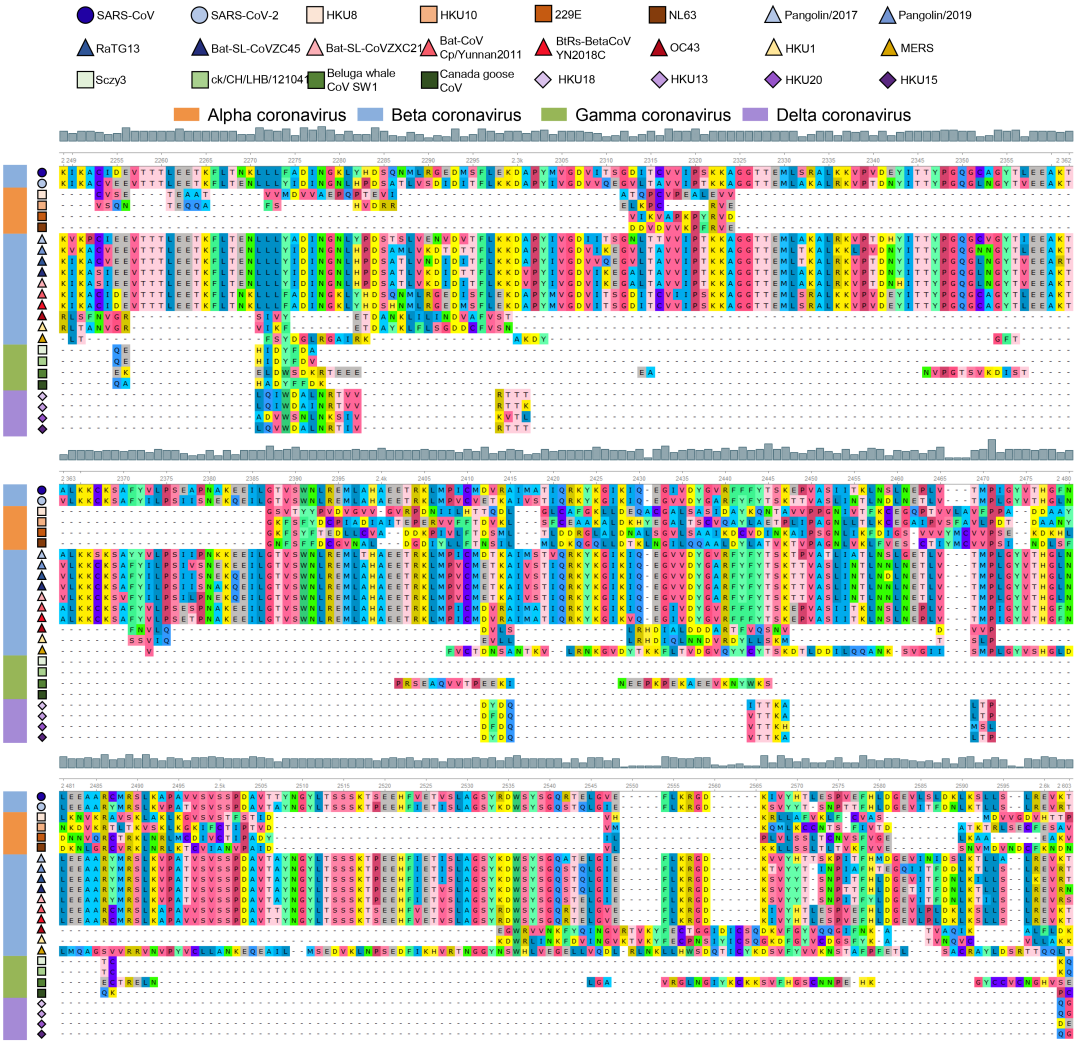

Fig. S3 Alignment of several coronavirus amino acid sequences. The figure shows the sequence alignment corresponding to SARS SUD. The shapes in various colors mark different kinds of coronaviruses. The color bar represents different genera of coronaviruses (orange, alpha coronavirus; blue, beta coronavirus; green, gamma coronavirus; purple, delta coronavirus). The grey histogram shows the consensus of the alignment sites.

Reference

[1] C.-C. Lai, T.-P. Shih, W.-C. Ko, H.-J. Tang, P.-R. Hsueh, Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges, International Journal of Antimicrobial Agents 55(3) (2020) 105924.

[2] J. Cui, F. Li, Z.-L. Shi, Origin and evolution of pathogenic coronaviruses, Nature Reviews Microbiology 17(3) (2019) 181-192.

[3] N.S. Zhong, B.J. Zheng, Y.M. Li, L.L.M. Poon, Z.H. Xie, K.H. Chan, P.H. Li, S.Y. Tan, Q. Chang, J.P. Xie, X.Q. Liu, J. Xu, D.X. Li, K.Y. Yuen, J.S.M. Peiris, Y. Guan, Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003, The Lancet 362(9393) (2003) 1353-1358.

[4] J.S.M. Peiris, Y. Guan, K.Y. Yuen, Severe acute respiratory syndrome, Nature Medicine 10(12) (2004) S88-S97.

[5] A. Zumla, J.F.W. Chan, E.I. Azhar, D.S.C. Hui, K.-Y. Yuen, Coronaviruses — drug discovery and therapeutic options, Nature Reviews Drug Discovery 15(5) (2016) 327-347.

[6] A.E. Gorbalenya, S.C. Baker, R.S. Baric, R.J. de Groot, C. Drosten, A.A. Gulyaeva, B.L. Haagmans, C. Lauber, A.M. Leontovich, B.W. Neuman, D. Penzar, S. Perlman, L.L.M. Poon, D.V. Samborskiy, I.A. Sidorov, I. Sola, J. Ziebuhr, V. Coronaviridae Study Group of the International Committee on Taxonomy of, The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2, Nature Microbiology 5(4) (2020) 536-544.

[7] W.-j. Guan, Z.-y. Ni, Y. Hu, W.-h. Liang, C.-q. Ou, J.-x. He, L. Liu, H. Shan, C.-l. Lei, D.S.C. Hui, B. Du, L.-j. Li, G. Zeng, K.-Y. Yuen, R.-c. Chen, C.-l. Tang, T. Wang, P.-y. Chen, J. Xiang, S.-y. Li, J.-l. Wang, Z.-j. Liang, Y.-x. Peng, L. Wei, Y. Liu, Y.-h. Hu, P. Peng, J.-m. Wang, J.-y. Liu, Z. Chen, G. Li, Z.-j. Zheng, S.-q. Qiu, J. Luo, C.-j. Ye, S.-y. Zhu, N.-s. Zhong, Clinical Characteristics of Coronavirus Disease 2019 in China, New England Journal of Medicine 382(18) (2020) 1708-1720.

[8] H.A. Rothan, S.N. Byrareddy, The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak, Journal of Autoimmunity 109 (2020) 102433.

[9] Z.Y. Zu, M.D. Jiang, P.P. Xu, W. Chen, Q.Q. Ni, G.M. Lu, L.J. Zhang, Coronavirus Disease 2019 (COVID-19): A Perspective from China, Radiology (2020) 200490.

[10] Y. Jin, H. Yang, W. Ji, W. Wu, S. Chen, W. Zhang, G. Duan, Virology, Epidemiology, Pathogenesis, and Control of COVID-19, Viruses 12(4) (2020).

[11] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, K. Cao, D. Liu, G. Wang, Q. Xu, X. Fang, S. Zhang, J. Xia, J. Xia, Artificial Intelligence Distinguishes COVID-19 from Community Acquired Pneumonia on Chest CT, Radiology (2020) 200905.

[12] J.P. Broughton, X. Deng, G. Yu, C.L. Fasching, V. Servellita, J. Singh, X. Miao, J.A. Streithorst, A. Granados, A. Sotomayor-Gonzalez, K. Zorn, A. Gopez, E. Hsu, W. Gu, S. Miller, C.-Y. Pan, H. Guevara, D.A. Wadford, J.S. Chen, C.Y. Chiu, CRISPR–Cas12-based detection of SARS-CoV-2, Nature Biotechnology (2020).

[13] J. Zheng, SARS-CoV-2: an Emerging Coronavirus that Causes a Global Threat, Int J Biol Sci 16(10) (2020) 1678-1685.

[14] Y.-R. Guo, Q.-D. Cao, Z.-S. Hong, Y.-Y. Tan, S.-D. Chen, H.-J. Jin, K.-S. Tan, D.-Y. Wang, Y. Yan, The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak – an update on the status, Military Medical Research 7(1) (2020) 11.

[15] D. Kim, J.-Y. Lee, J.-S. Yang, J.W. Kim, V.N. Kim, H. Chang, The Architecture of SARS-CoV-2 Transcriptome, Cell 181(4) (2020) 914-921.e10.

[16] Y. Chen, Q. Liu, D. Guo, Emerging coronaviruses: Genome structure, replication, and pathogenesis, Journal of Medical Virology 92(4) (2020) 418-423.

[17] M. Hoffmann, H. Kleine-Weber, S. Schroeder, N. Krüger, T. Herrler, S. Erichsen, T.S. Schiergens, G. Herrler, N.-H. Wu, A. Nitsche, M.A. Müller, C. Drosten, S. Pöhlmann, SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor, Cell 181(2) (2020) 271-280.e8.

[18] A.C. Walls, Y.-J. Park, M.A. Tortorici, A. Wall, A.T. McGuire, D. Veesler, Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein, Cell 181(2) (2020) 281-292.e6.

[19] M.A. Shereen, S. Khan, A. Kazmi, N. Bashir, R. Siddique, COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses, Journal of Advanced Research 24 (2020) 91-98.

[20] T. Zhang, Q. Wu, Z. Zhang, Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak, Current Biology 30(7) (2020) 1346-1351.e2.

[21] T.T.-Y. Lam, M.H.-H. Shum, H.-C. Zhu, Y.-G. Tong, X.-B. Ni, Y.-S. Liao, W. Wei, W.Y.-M. Cheung, W.-J. Li, L.-F. Li, G.M. Leung, E.C. Holmes, Y.-L. Hu, Y. Guan, Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins, Nature (2020).

[22] K.G. Andersen, A. Rambaut, W.I. Lipkin, E.C. Holmes, R.F. Garry, The proximal origin of SARS-CoV-2, Nature Medicine 26(4) (2020) 450-452.

[23] P. Zhou, X.-L. Yang, X.-G. Wang, B. Hu, L. Zhang, W. Zhang, H.-R. Si, Y. Zhu, B. Li, C.-L. Huang, H.-D. Chen, J. Chen, Y. Luo, H. Guo, R.-D. Jiang, M.-Q. Liu, Y. Chen, X.-R. Shen, X. Wang, X.-S. Zheng, K. Zhao, Q.-J. Chen, F. Deng, L.-L. Liu, B. Yan, F.-X. Zhan, Y.-Y. Wang, G.-F. Xiao, Z.-L. Shi, A pneumonia outbreak associated with a new coronavirus of probable bat origin, Nature 579(7798) (2020) 270-273.

[24] K. Xiao, J. Zhai, Y. Feng, N. Zhou, X. Zhang, J.-J. Zou, N. Li, Y. Guo, X. Li, X. Shen, Z. Zhang, F. Shu, W. Huang, Y. Li, Z. Zhang, R.-A. Chen, Y.-J. Wu, S.-M. Peng, M. Huang, W.-J. Xie, Q.-H. Cai, F.-H. Hou, W. Chen, L. Xiao, Y. Shen, Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins, Nature (2020).

[25] M.L. Bochman, K. Paeschke, V.A. Zakian, DNA secondary structures: stability and function of G-quadruplex structures, Nature Reviews Genetics 13(11) (2012) 770-780.

[26] C.K. Kwok, C.J. Merrick, G-Quadruplexes: Prediction, Characterization, and Biological Application, Trends in Biotechnology 35(10) (2017) 997-1013.

[27] D. Varshney, J. Spiegel, K. Zyner, D. Tannahill, S. Balasubramanian, The regulation and functions of DNA and RNA G-quadruplexes, Nature Reviews Molecular Cell Biology (2020).

[28] J. Spiegel, S. Adhikari, S. Balasubramanian, The Structure and Function of DNA G-Quadruplexes, Trends in Chemistry 2(2) (2020) 123-136.

[29] P. Prorok, M. Artufel, A. Aze, P. Coulombe, I. Peiffer, L. Lacroix, A. Guédin, J.-L. Mergny, J. Damaschke, A. Schepers, B. Ballester, M. Méchali, Involvement of G-quadruplex regions in mammalian replication origin activity, Nature Communications 10(1) (2019) 3274.

[30] A.-L. Valton, M.-N. Prioleau, G-Quadruplexes in DNA Replication: A Problem or a Necessity?, Trends in Genetics 32(11) (2016) 697-706.

[31] S. Hoshina, K. Yura, H. Teranishi, N. Kiyasu, A. Tominaga, H. Kadoma, A. Nakatsuka, T. Kunichika, C. Obuse, S. Waga, Human Origin Recognition Complex Binds Preferentially to G-quadruplex-preferable RNA and Single-stranded DNA, 288(42) (2013) 30161-30171.

[32] A.-L. Valton, V. Hassan-Zadeh, I. Lema, N. Boggetto, P. Alberti, C. Saintomé, J.-F. Riou, M.-N. Prioleau, G4 motifs affect origin positioning and efficiency in two vertebrate replicators, EMBO J 33(7) (2014) 732-746.

[33] L.I. Jansson, J. Hentschel, J.W. Parks, T.R. Chang, C. Lu, R. Baral, C.R. Bagshaw, M.D. Stone, Telomere DNA G-quadruplex folding within actively extending human telomerase, Proceedings of the National Academy of Sciences 116(19) (2019) 9350.

[34] Q. Wang, J.-q. Liu, Z. Chen, K.-w. Zheng, C.-y. Chen, Y.-H. Hao, Z. Tan, G-quadruplex formation at the 3' end of telomere DNA inhibits its extension by telomerase, polymerase and unwinding by helicase, Nucleic Acids Res 39(14) (2011) 6229-6237.

[35] A.L. Moye, K.C. Porter, S.B. Cohen, T. Phan, K.G. Zyner, N. Sasaki, G.O. Lovrecz, J.L. Beck, T.M. Bryan, Telomeric G-quadruplexes are a substrate and site of localization for human telomerase, Nature Communications 6(1) (2015) 7643.

[36] K. Takahama, A. Takada, S. Tada, M. Shimizu, K. Sayama, R. Kurokawa, T. Oyoshi, Regulation of Telomere Length by G-Quadruplex Telomere DNA- and TERRA-Binding Protein TLS/FUS, Chemistry & Biology 20(3) (2013) 341-350.

[37] J. Tang, Z.-y. Kan, Y. Yao, Q. Wang, Y.-h. Hao, Z. Tan, G-quadruplex preferentially forms at the very 3′ end of vertebrate telomeric DNA, Nucleic Acids Res 36(4) (2007) 1200-1208.

[38] S. Kumari, A. Bugaut, J.L. Huppert, S. Balasubramanian, An RNA G-quadruplex in the 5′ UTR of the NRAS proto-oncogene modulates translation, Nature Chemical Biology 3(4) (2007) 218-221.

[39] R. Jodoin, J.C. Carrier, N. Rivard, M. Bisaillon, J.-P. Perreault, G-quadruplex located in the 5′ UTR of the BAG-1 mRNA affects both its cap-dependent and cap-independent translation through global secondary structure maintenance, Nucleic Acids Res 47(19) (2019) 10247-10266.

[40] D. Gomez, A. Guédin, J.-L. Mergny, B. Salles, J.-F. Riou, M.-P. Teulade-Fichou, P. Calsou, A G-quadruplex structure within the 5′ -UTR of TRF2 mRNA represses translation in human cells, Nucleic Acids Res 38(20) (2010) 7187-7198.

[41] P. Murat, G. Marsico, B. Herdy, A. Ghanbarian, G. Portella, S. Balasubramanian, RNA G-quadruplexes at upstream open reading frames cause DHX36- and DHX9-dependent translation of human mRNAs, Genome Biol 19(1) (2018) 229.

[42] N. Saranathan, P. Vivekanandan, G-Quadruplexes: More Than Just a Kink in Microbial Genomes, Trends in Microbiology 27(2) (2019) 148-163.

[43] M. Métifiot, S. Amrane, S. Litvak, M.-L. Andreola, G-quadruplexes in viruses: function and potential therapeutic applications, Nucleic Acids Res 42(20) (2014) 12352-12366.

[44] E. Ruggiero, S.N. Richter, G-quadruplexes and G-quadruplex ligands: targets and tools in antiviral therapy, Nucleic Acids Res 46(7) (2018) 3270-3283.

[45] J. Tan, C. Vonrhein, O.S. Smart, G. Bricogne, M. Bollati, Y. Kusov, G. Hansen, J.R. Mesters, C.L. Schmidt, R. Hilgenfeld, The SARS-unique domain (SUD) of SARS coronavirus contains two macrodomains that bind G-quadruplexes, PLoS Pathog 5(5) (2009) e1000428-e1000428.

[46] E. Butovskaya, B. Heddi, B. Bakalar, S.N. Richter, A.T. Phan, Major G-Quadruplex Form of HIV-1 LTR Reveals a (3 + 1) Folding Topology Containing a Stem-Loop, Journal of the American Chemical Society 140(42) (2018) 13654-13662.

[47] D. Piekna-Przybylska, M.A. Sullivan, G. Sharma, R.A. Bambara, U3 Region in the HIV-1 Genome Adopts a G-Quadruplex Structure in Its RNA and DNA Sequence, Biochemistry 53(16) (2014) 2581-2593.

[48] E. Butovskaya, P. Soldà, M. Scalabrin, M. Nadai, S.N. Richter, HIV-1 Nucleocapsid Protein Unfolds Stable RNA G-Quadruplexes in the Viral Genome and Is Inhibited by G-Quadruplex Ligands, ACS Infectious Diseases 5(12) (2019) 2127-2135.

[49] R. Perrone, E. Butovskaya, D. Daelemans, G. Palù, C. Pannecouque, S.N. Richter, Anti-HIV-1 activity of the G-quadruplex ligand BRACO-19, Journal of Antimicrobial Chemotherapy 69(12) (2014) 3248-3258.

[50] A.M. Fleming, Y. Ding, A. Alenko, C.J. Burrows, Zika Virus Genomic RNA Possesses Conserved G-Quadruplexes Characteristic of the Flaviviridae Family, ACS Infectious Diseases 2(10) (2016) 674-681.

[51] K. Tlučková, M. Marušič, P. Tóthová, L. Bauer, P. Šket, J. Plavec, V. Viglasky, Human Papillomavirus G-Quadruplexes, Biochemistry 52(41) (2013) 7207-7216.

[52] M. Marušič, L. Hošnjak, P. Krafčikova, M. Poljak, V. Viglasky, J. Plavec, The effect of single nucleotide polymorphisms in G-rich regions of high-risk human papillomaviruses on structural diversity of DNA, Biochimica et Biophysica Acta (BBA) - General Subjects 1861(5, Part B) (2017) 1229-1236.

[53] S.-R. Wang, Q.-Y. Zhang, J.-Q. Wang, X.-Y. Ge, Y.-Y. Song, Y.-F. Wang, X.-D. Li, B.-S. Fu, G.-H. Xu, B. Shu, P. Gong, B. Zhang, T. Tian, X. Zhou, Chemical Targeting of a G-Quadruplex RNA in the Ebola Virus L Gene, Cell Chemical Biology 23(9) (2016) 1113-1122.

[54] J.-D. Beaudoin, R. Jodoin, J.-P. Perreault, New scoring system to identify RNA G-quadruplex folding, Nucleic Acids Res 42(2) (2014) 1209-1223.

[55] R.J. Andrews, J.M. Peterson, H.S. Haniff, J. Chen, C. Williams, M. Grefe, M.D. Disney, W.N. Moss, An in silico map of the SARS-CoV-2 RNA Structurome, bioRxiv (2020) 2020.04.17.045161.

[56] C.K. Kwok, G. Marsico, A.B. Sahakyan, V.S. Chambers, S. Balasubramanian, rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome, Nature Methods 13(10) (2016) 841-844.

[57] J.-D. Beaudoin, J.-P. Perreault, 5'-UTR G-quadruplex structures acting as translational repressors, Nucleic Acids Res 38(20) (2010) 7022-7036.

[58] D.S.M. Lee, L.R. Ghanem, Y. Barash, Integrative analysis reveals RNA G-quadruplexes in UTRs are selectively constrained and enriched for functional associations, Nature Communications 11(1) (2020) 527.

[59] H. Han, L.H. Hurley, G-quadruplex DNA: a potential target for anti-cancer drug design, Trends in Pharmacological Sciences 21(4) (2000) 136-142.

[60] S. Neidle, Quadruplex nucleic acids as targets for anticancer therapeutics, Nature Reviews Chemistry 1(5) (2017) 0041.

[61] K.M. Miller, R. Rodriguez, G-quadruplexes: selective DNA targeting for cancer therapeutics?, Expert Review of Clinical Pharmacology 4(2) (2011) 139-142.

[62] S. Balasubramanian, L.H. Hurley, S. Neidle, Targeting G-quadruplexes in gene promoters: a novel anticancer strategy?, Nat Rev Drug Discov 10(4) (2011) 261-275.

[63] E. Ruggiero, S.N. Richter, Viral G-quadruplexes: New frontiers in virus pathogenesis and antiviral therapy, Annu Rep Med Chem (2020) 10.1016/bs.armc.2020.04.001.

[64] R. Perrone, S. Artusi, E. Butovskaya, M. Nadai, C. Pannecouque, S.N. Richter, G-Quadruplexes in the Human Immunodeficiency Virus-1 and Herpes Simplex Virus-1: New Targets for Antiviral Activity by Small Molecules, in: V.V. Toi, T.H. Lien Phuong (Eds.) 5th International Conference on Biomedical Engineering in Vietnam, Springer International Publishing, Cham, 2015, pp. 207-210.

[65] F. Zaccaria, C. Fonseca Guerra, RNA versus DNA G-Quadruplex: The Origin of Increased Stability, Chemistry 24(61) (2018) 16315-16322.

[66] A. Joachimi, A. Benz, J.S. Hartig, A comparison of DNA and RNA quadruplex structures and stabilities, Bioorganic & Medicinal Chemistry 17(19) (2009) 6811-6815.

[67] P. Majee, S. Kumar Mishra, N. Pandya, U. Shankar, S. Pasadi, K. Muniyappa, D. Nayak, A. Kumar, Identification and characterization of two conserved G-quadruplex forming motifs in the Nipah virus genome and their interaction with G-quadruplex specific ligands, Scientific Reports 10(1) (2020) 1477.

[68] R. Perrone, M. Nadai, J.A. Poe, I. Frasson, M. Palumbo, G. Palù, T.E. Smithgall, S.N. Richter, Formation of a Unique Cluster of G-Quadruplex Structures in the HIV-1 nef Coding Region: Implications for Antiviral Activity, PLOS ONE 8(8) (2013) e73121.

[69] Y. Kusov, J. Tan, E. Alvarez, L. Enjuanes, R. Hilgenfeld, A G-quadruplex-binding macrodomain within the "SARS-unique domain" is essential for the activity of the SARS-coronavirus replication-transcription complex, Virology 484 (2015) 313-322.

[70] R. Shahid, A. Bugaut, S. Balasubramanian, The BCL-2 5′ Untranslated Region Contains an RNA G-Quadruplex-Forming Motif That Modulates Protein Expression, Biochemistry 49(38) (2010) 8300-8306.

[71] S.-R. Wang, Y.-Q. Min, J.-Q. Wang, C.-X. Liu, B.-S. Fu, F. Wu, L.-Y. Wu, Z.-X. Qiao, Y.-Y. Song, G.-H. Xu, Z.-G. Wu, G. Huang, N.-F. Peng, R. Huang, W.-X. Mao, S. Peng, Y.-Q. Chen, Y. Zhu, T. Tian, X.-L. Zhang, X. Zhou, A highly conserved G-rich consensus sequence in hepatitis C virus core gene represents a new anti-hepatitis C target, Sci Adv 2(4) (2016) e1501535-e1501535.

[72] C. Jaubert, A. Bedrat, L. Bartolucci, C. Di Primo, M. Ventura, J.-L. Mergny, S. Amrane, M.-L. Andreola, RNA synthesis is modulated by G-quadruplex formation in Hepatitis C virus negative RNA strand, Scientific Reports 8(1) (2018) 8120.

[73] C. Yin, Genotyping coronavirus SARS-CoV-2: methods and implications, Genomics (2020).

[74] H. Yao, X. Lu, Q. Chen, K. Xu, Y. Chen, L. Cheng, F. Liu, Z. Wu, H. Wu, C. Jin, M. Zheng, N. Wu, C. Jiang, L. Li, Patient-derived mutations impact pathogenicity of SARS-CoV-2, medRxiv (2020) 2020.04.14.20060160.

[75] A. Mishra, A.K. Pandey, P. Gupta, P. Pradhan, S. Dhamija, J. Gomes, B. Kundu, P. Vivekanandan, M.B. Menon, Mutation landscape of SARS-CoV-2 reveals three mutually exclusive clusters of leading and trailing single nucleotide substitutions, bioRxiv (2020) 2020.05.07.082768.

[76] L. Chen, B. Liu, J. Yang, Q. Jin, DBatVir: the database of bat-associated viruses, Database 2014 (2014).

[77] Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data – from vision to reality, 22(13) (2017) 30494.

[78] F. Madeira, Y.M. Park, J. Lee, N. Buso, T. Gur, N. Madhusoodanan, P. Basutkar, A.R.N. Tivey, S.C. Potter, R.D. Finn, R. Lopez, The EMBL-EBI search and sequence analysis tools APIs in 2019, Nucleic Acids Res 47(W1) (2019) W636-W641.

[79] F. Sievers, A. Wilm, D. Dineen, T.J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J.D. Thompson, D.G. Higgins, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, Molecular Systems Biology 7(1) (2011) 539.

[80] F. Sievers, D.G. Higgins, Clustal Omega for making accurate alignments of many protein sequences, Protein Science 27(1) (2018) 135-145.

[81] K. Okonechnikov, O. Golosova, M. Fursov, U.t. the, Unipro UGENE: a unified bioinformatics toolkit, Bioinformatics 28(8) (2012) 1166-1167.

[82] S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms, Mol Biol Evol 35(6) (2018) 1547-1549.

[83] E.V. Davydov, D.L. Goode, M. Sirota, G.M. Cooper, A. Sidow, S. Batzoglou, Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++, PLOS Computational Biology 6(12) (2010) e1001025.

[84] O. Doluca, G4Catchall: A G-quadruplex prediction approach considering atypical features, Journal of Theoretical Biology 463 (2019) 92-98.

[85] J. Hon, T. Martínek, J. Zendulka, M. Lexa, pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R, Bioinformatics 33(21) (2017) 3373-3379.

[86] O. Kikin, L. D'Antonio, P.S. Bagga, QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences, Nucleic Acids Res 34(suppl_2) (2006) W676-W682.

[87] E. Puig Lombardi, A. Londoño-Vallejo, A guide to computational methods for G-quadruplex prediction, Nucleic Acids Res 48(1) (2019) 1-15.

[88] A.R. Quinlan, I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics 26(6) (2010) 841-842.

[89] A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F.T. Heer, T.A P. de Beer, C. Rempfer, L. Bordoli, R. Lepore, T. Schwede, SWISS-MODEL: homology modelling of protein structures and complexes, Nucleic Acids Res 46(W1) (2018) W296-W303.