Evidence for a long-range RNA-RNA interaction between *ORF8* and the downstream region of the *Spike* polybasic insertion of SARS-CoV-2

Filipe Pereira[1,2], Amirhossein Manzourolajdad[3]

1. Centre for Functional Ecology, Department of Life Sciences, University of Coimbra, Calçada Martim de Freitas, 3000-456 Coimbra, Portugal, e-mail: fpereirapt@gmail.com
2. IDENTIFICA genetic testing, Rua Simão Bolívar 259 3º Dir Tras. 4470-214 Maia, Portugal.
3. Department of Biohealth Informatics, School of Informatics and Computing, Indiana University Purdue University, 535 W Michigan St, Indianapolis, IN 46202, USA, e-mail: amanzour@iu.edu

Abstract

SARS-CoV-2 has affected people all over the world as the causative agent of COVID-19. The virus is related to the highly lethal SARS-CoV responsible for the 2002-2003 SARS outbreak in Asia. Intense research is ongoing to understand why both viruses have different spreading capacities and mortality rates. Similar to other betacoronaviruses, long-range RNA-RNA interactions occur between different parts of the viral genomic RNA, resulting in discontinuous transcription and production of various sub-genomic RNAs. These sub-genomic RNAs are then translated into different viral proteins. An important difference between both viruses is a polybasic insertion in the *Spike* region of SARS-CoV-2, absent in SARS-CoV. Here we show that a 26-base-pair long-range RNA-RNA interaction occurs between the genomic region downstream of the Spike insertion and *ORF8* in SARS-CoV-2. Predictions suggest that the corresponding *ORF8* region forms the most energetically favorable interaction with that of *Spike* region from amongst all possible candidate regions within SARS-CoV-2 genomic RNA. We also found signs of sequence covariation in the predicted interaction using a large dataset with 27,592 full-length SARS-CoV-2 genomes. In particular, a synonymous mutation in ORF8 accommodated for base pairing with Spike [G23675 C28045U], and a non-synonymous mutation in Spike accommodated for base pairing with ORF8 [C23679U G28042] both of which were in close proximity of one another. The predicted interactions can potentially be related to regulation of sub-genomic RNA production rates.

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a highly transmissible and pathogenic coronavirus that emerged in late 2019 and has caused a pandemic of acute respiratory disease, named 'coronavirus disease 2019' (COVID-19) (Hu, Guo et al. 2020). SARS-CoV-2 is related to SARS-CoV, a highly lethal virus responsible for an outbreak in 2002-2003 that was contained after intense public health mitigation measures (Standl, Jockel et al. 2020). SARS-CoV-2 is an enveloped, positive sense, single stranded RNA virus from the Betacoronavirus genus, whose genome is around 30,000 nucleotides (nt) long (Lu, Zhao et al. 2020).

Replication and transcription of SARS-CoV-2 occur in the cytoplasm of infected cells. The virus replicates through synthesis of the negative sense strand, which in turn produces a plus RNA strand virus genome. The complex replication/transcription machinery enables discontinuous transcription, in which a series of sub-genomic RNAs are also produced through the process of template switching during negative sense RNAs synthesis (Sola, Almazán et al. 2015, Snijder, Limpens et al. 2020). These sub-genomic RNAs encode for spike (S), envelope (E), membrane (M) and nucleocapsid (N) virus proteins, as well as and several accessory proteins known as Open Reading Frames (ORF) 3a, 6, 7a, 7b, 8, and 10 (Kim, Lee et al. 2020).

The genomes SARS-CoV-2, SARS-CoV share around 89% sequence identity (Naqvi, Fatima et al. 2020). The genomic differences explain the disparities in the dispersal and immune evasion of both viruses (Ortiz-Fernandez and Sawalha 2020). A major difference between the genome of both viruses is a polybasic insertion at the S1/S2 cleavage site, resulting from a 12-nt insert in the Spike region of SARS-CoV-2 that does not exist in SARS-CoV. The 12-nt insert may influence the Spike protein infectivity. A previous study identified a conserved RNA stable structure at the 12-nt insert and its vicinity (Rangan, Zheludev et al. 2020). However, it remains to be determined which interactions may occur at the RNA level for the polybasic insertion at the S1/S2 cleavage site.

The 12-nt insert has a high GC content (CCUCGGCGGGCA; positions 23,603-23,614 of the reference), which suggest a possible RNA structural role. Generally, RNA structures can play critical roles in the life cycle of betacoronaviruses. Such structures can be formed either in specific regions (local) or be the result of long-range interactions between distant sections of the genome. It has been shown that betacoronaviruses can form long-range high-order RNA-RNA interactions that contribute to template switch and consequently production of sub-genomic mRNAs (Sola, Almazán et al. 2015). Recent studies have also found locally stable RNA structures within the SARS-CoV-2 genome (Andrews, Peterson et al. 2020, Bartas, Brázda et al. 2020, Lan, Allan et al. 2020, Simmonds 2020). Moreover, *in-vivo* RNA structure prediction methods such as dimethyl sulfate mutational profiling with sequencing (DMS-MaP-seq) suggest that SARS-CoV-2 forms RNA structures within the majority of its genome (Lan, Allan et al. 2020), some of possible relevance to the virus life cycle. These structural elements may be the target of RNA-based therapeutic applications (Manfredonia, Nithin et al. 2020, Rouskin, Lan et al. 2021), or may lead to methods for inhibiting viral growth (Huston, Wan et al. 2021). Therefore, further investigation of RNA structure, especially those unique to SARS-CoV-2 can be of great importance.

Here we show that the 12-nt insert is associated with long-range RNA-RNA interactions, particularly with ORF8. At first glance, the insert can take part in increasing stability of local RNA structure, due to its high GC composition. In our long-range RNA-RNA interaction analysis, however, we found that a sequence segment immediately downstream of the 12-nt insert that has RNA binding affinity with a segment on *ORF8*, raising the possibility that the 12-nt insert may be

allosterically impacting (i.e. facilitating or impeding) the predicted long-range RNA-RNA interaction.

Material and Methods

Genomic sequences

We used the SARS-CoV-2 (NC_045512) and SARS-CoV **(**NC_004718) reference sequences. Additional 27,592 full-length SARS-CoV-2 genome sequences were downloaded from the Global Initiative on Sharing All Influenza Data (GISAID) platform (https://www.gisaid.org/) on May 20, 2020. The alignment of full-length SARS-CoV-2 sequences is here referred to as **msa-0520**.

RNA structural analyses

Different components of the RNAstructure software package (Reuter and Mathews 2010) along with other tools were used for secondary structure predictions. Minimum Free Energy structure (MFE) prediction was performed using both fold (Mathews 2004) and ViennaRNA software package (Lorenz, Bernhart et al. 2011) using default parameters. Individual base-pair probabilities are according to McCaskill's partition function (McCaskill 1990, Mathews 2004). The Maximum Expected Accuracy (MEA) (Lu, Gloor et al. 2009, Reuter and Mathews 2010) and pseudoknot predictions were done using ProbKnot (Bellaousov and Mathews 2010, Zhang, Zhang et al. 2020) using default parameters. Genome-wide RNA-RNA interaction targets candidates were identified using IntaRNA (Busch, Richter et al. 2008, Wright, Georg et al. 2014, Mann, Wright et al. 2017, Raden, Ali et al. 2018), with different search mode parameters (i.e. `--mode M, S, X, B,` and `P`). IntaRNA results using different parameters were mostly consistent, if not identical. We used IntaRNA to find the most thermodynamically stable base pairing from any SARS-CoV-2 region with the *Spike* region 23,600 - 24,107, using as query two sequence segments: 23,600 - 24,107 and 23,917 - 24,118. A similar procedure was used for SARS-CoV.

Bimolecular RNA structural predictions were performed using bifold (Mathews, Burkard et al. 1999) default parameters. The bifold program takes two sequences as arguments and predicts the most energetically favorable bimolecular conformation. Three-way RNA structural predictions (i.e., simultaneous RNA-RNA interaction amongst three RNA molecules) were predicted by combining information from bimolecular folding predictions. From amongst the three RNA sequences, the longest of the three was identified as the 'main sequence'. Two bimolecular folding predictions were then performed between the main sequence and either of the two remaining RNA sequences. The resulting RNA-RNA pairing regions were then deleted from the main sequence. The remaining sections of the sequence were then concatenated to construct a shortened main sequence. The structure of the shortened main sequence was then predicted using ProbKnot. The final trimolecular RNA structure predictions would then consist of restoring the main sequence by adding RNA-RNA binding regions in their original locations. RNA secondary structure schematics were done using VARNA (Darty, Denise et al. 2009).

Compensatory Mutations Analysis of Long-range RNA-RNA interactions

Compensatory mutations within the multiple sequence alignments were investigated using the R-scape software package (Rivas, Clements et al. 2017, Rivas 2020, Rivas, Clements et al. 2020, Rivas and Eddy 2020), which analyzes covariation in nucleotide pairs in the population to infer possible compensatory mutations in an RNA base pair. If the consensus RNA secondary structure is not provided by the user, the software is also capable of predicting the consensus structure from the population of sequences using an implementation of the CaCoFold algorithm.

Compensatory (covarying) mutations for long-range RNA-RNA interactions were analyzed by retrieving the two sequence segments that constitute the desired RNA-RNA interaction from all downloaded SARS-CoV-2 sequences. Sequences with runs of ambiguous nucleotides in selected regions were filtered out. Then, the long-range RNA-RNA interacting structure was predicted by finding the consensus secondary structure within the population of sequences in the dataset using R-scape implementation of CaCoFold. The consensus structure was compared to bifold predictions for verification. Nucleotide pairs belonging to the consensus structure were then examined within the dataset for evidence of covariation using the built-in survival function that plots the distribution of base pairs with respect to their corresponding covariation scores (t). For a typical survival function plot, the black depicts the survival function for the null alignment. Blue depicts the survival function for the alignment. The black line fits to the tail of the null distribution. E-value corresponding to a pair is obtained by drawing a vertical line from the point to cross the black line. If a dot representing a base pair crosses the 0.05-threshold vertical blue line, it is considering as significantly covarying (See R-scape manual for more details).

Results

Downstream region of the Spike 12-nt insert contains locally stable RNA structural features

The flanking regions of the 12-nt insert in *Spike* (23,600-24,107) showed a consistent local RNA structure. The SARS-CoV and SARS-CoV-2 predicted structures differ in this region due to their low sequence similarity (76% similarity; p-distance ~ 0.24). The SARS-CoV-2 predicted structure has a higher stability (-141.5 kcal/mol) that the structure predicted for the homologous region in SARS-CoV (-125.00 kcal/mol; Figures 1A and 1B). The 12-nt insert tends to form a stem with complementary nucleotides approximately 500-nt downstream. The stem named S1 located immediately after the 12-nt insert is common to both SARS-CoV-2 and SARS-CoV (Figures 1A and 1B), and has been found conserved across betacoronaviruses (Rangan, Zheludev et al. 2020). Both MEA and Probknot predicted structures downstream of the insert are similar (See structure in boxes; Figure 1A). Together, these results suggest the existence of stable RNA structural features downstream of the 12-nt insert.

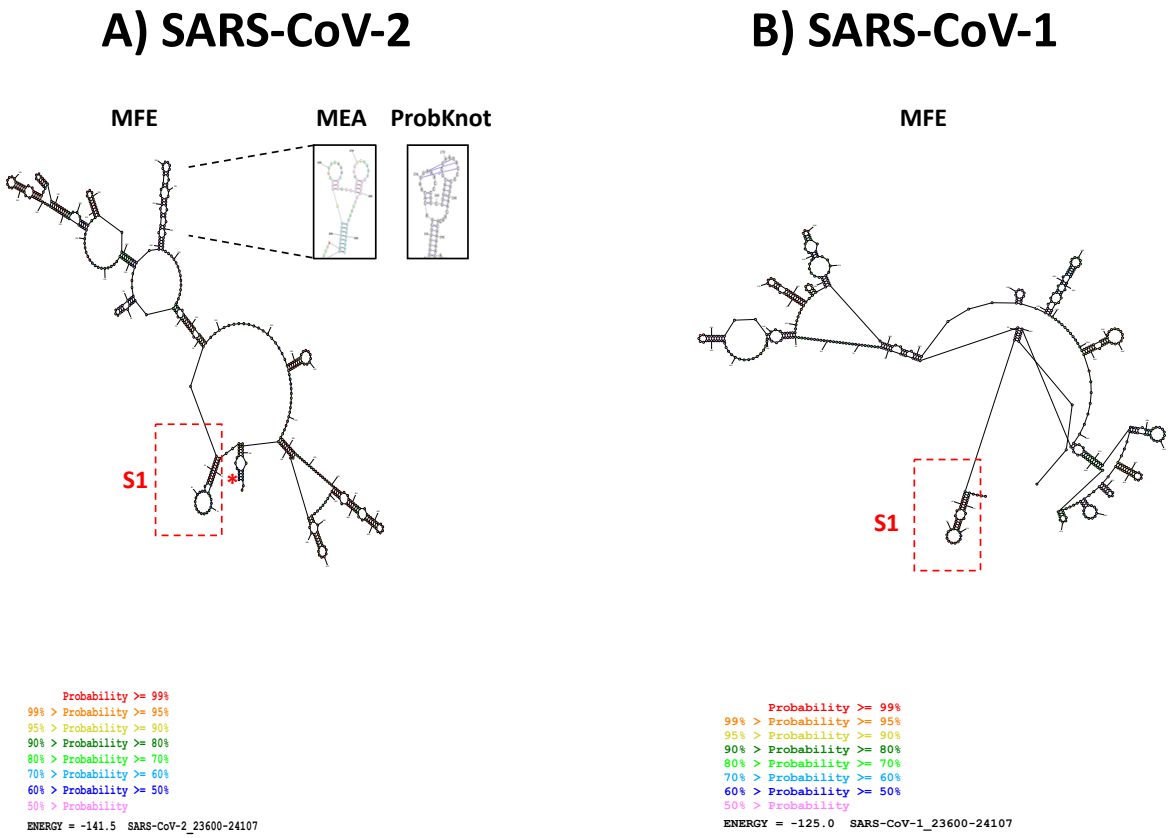


Figure 1. RNA secondary structure predictions, MFE, MEA, and ProbKnot for the 23,600-24,107 *Spike* region of SARS-CoV-2 (A) and SARS-CoV (B). The corresponding sequence of SARS-CoV was found by aligning SARS-CoV-2 and SARS-CoV with 121/508 nucleotides being different between the two sequence segments (sequence similarity 76%; p-distance ~ 0.24). The red asterisk represents the location of the 12-nt insert on SARS-CoV-2. Stem S1 is shown in red.

The downstream of Spike 12-nt Insert is predicted to interact with *ORF8*

The possibility of long-range RNA-RNA interaction between the 12-nt insert region and other genomic regions was investigated using IntaRNA. Figure 2A shows IntaRNA results on the complete SARS-CoV genomes. In almost all cases, results were consistent across different choices of mode parameters. Predictions using query 23,917 - 24,118 suggest a binding to Orf1ab. In specific, region 24,084-24,114 (*Spike*) formed an RNA interaction with region 17,012-17,046 (*Orf1ab*), containing ~30 base pairs between the two regions. The predicted structure was also found in SARS-CoV, with similar interacting energies (Figures 2A and 2B). On the contrary, other regions yield different results for both viruses. The region 23,663-23,804 (*Spike*) of SARS-CoV was predicted to bind 29,548-29,696 (3'UTR), while the same region (23,660-23,703) of SARS-CoV-2 was predicted to bind region 28,025-28,060 (*ORF8*).

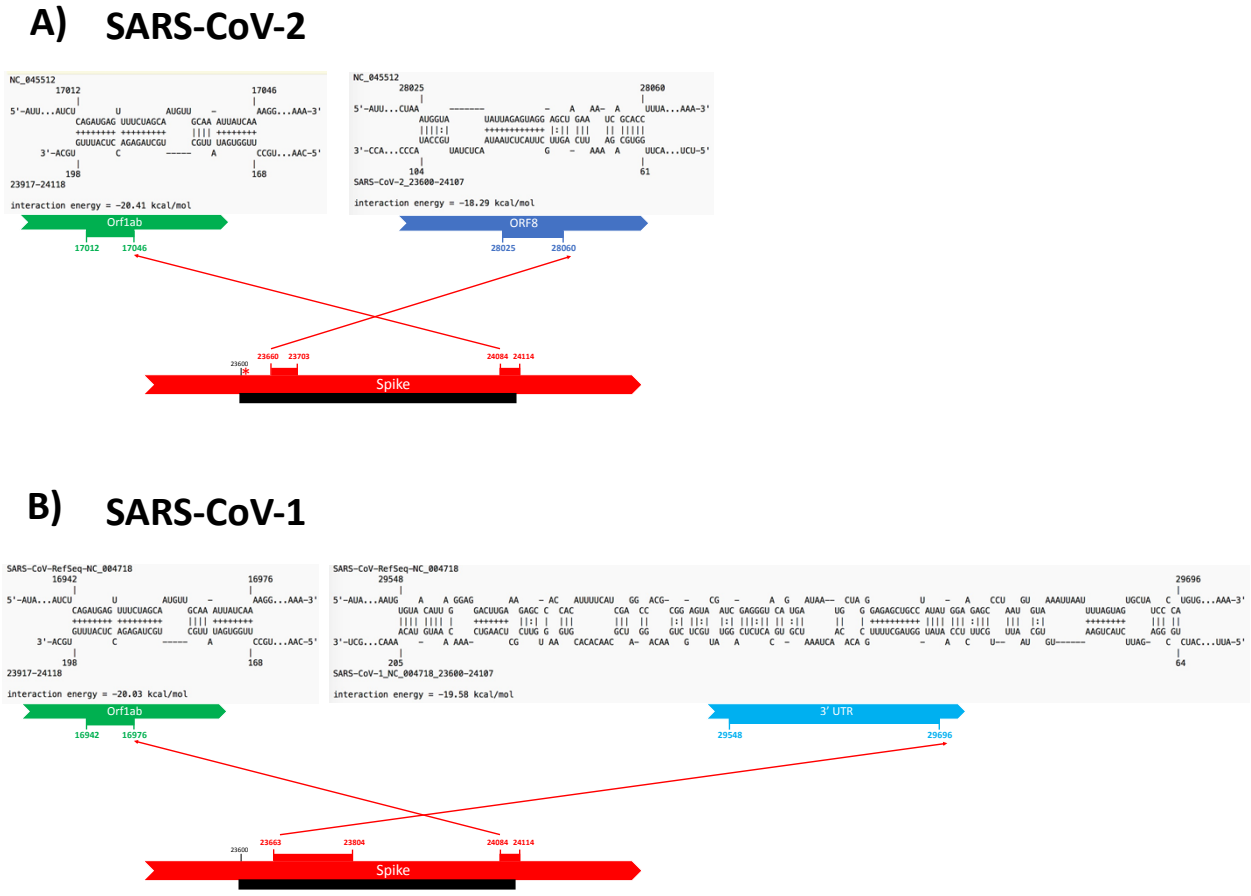


Figure 2. Long-range RNA-RNA interactions in SARS-CoV-2 (A) and SARS-CoV (B). Query parameters were set to regions (23600-24107 *Spike*) and (23917-24118 *Spike*), while the target parameter was the complete genome of the SARS-CoV-2 RNA. The IntaRNA software was used for predictions (See Materials and Methods). Region highlighted by the black rectangle shows region of query sequence. Relative location of the 12-nt insert is shown by a red asterisk.

Covarying mutations suggestive of Spike-ORF8 long-range RNA-RNA interaction

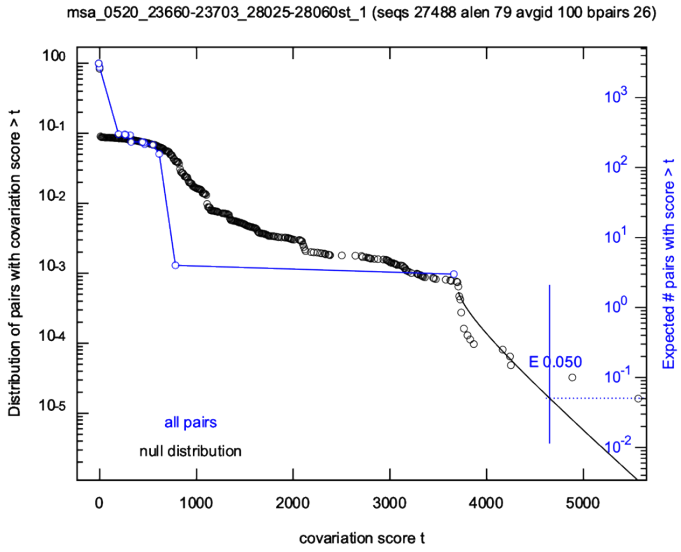
The predicted interactions between the *Spike* (23,660-23,703) and *ORF8* (28,025-28,060) regions was analyzed for compensatory mutations. Sequence identity on the interacting regions for the 27,488 genomes was 99.98% (certain sequences with gaps were eliminated from the dataset with original size of 27,592). Three significantly covarying nucleotide pairs were observed across the population, none of which belonged to the predicted structure and all of which belonged to *ORF8* region. The three covarying pairs refer to *ORF8* non-synonymous mutations and are summarized in Supplementary Materials Text S1, Figure S1, and Tables S1 and S2.

Table 1 shows the global position and frequency of these covarying pairs on *Spike* and *ORF8* reference sequences. The top four pairs with the highest numbers of substitutions within 27,488 sequences are: 23,675-28,045 with *n*=2, 23,673-28,048 with *n*=41, 23,660-28,057 with *n*=29 and 23,679-28,042 with *n*=29. The high frequency mutation (C28045U, synonymous) within the predicted Spike-ORF8 interaction was located in *ORF8*, being observed in 41 times within 27,488 sequences. The second most frequent mutation (C23679U) was non-synonymous and occurred on the second codon in Spike. The two base pairs with the highest mutation frequency are only a few positions apart (Figure 3B). The substitution rate within the predicted RNA base pairs, however, was not high enough to pass the 0.05 statistical significance threshold. Figure 3A shows the survival function of scores of all pairs in the Spike-ORF8 interaction region. The black depicts the survival function for the null alignment (i.e., all hypothetical base pairs within the region). Blue depicts the survival function for the alignment (i.e., all predicted base pairs within the structure). As we can see, none of the predicted base pairs (shown in blue) cross the line representing 0.05 significance threshold, although base pair [23675 28045] seems to be an outlier within the distribution (See the blue population in Figure 3A).

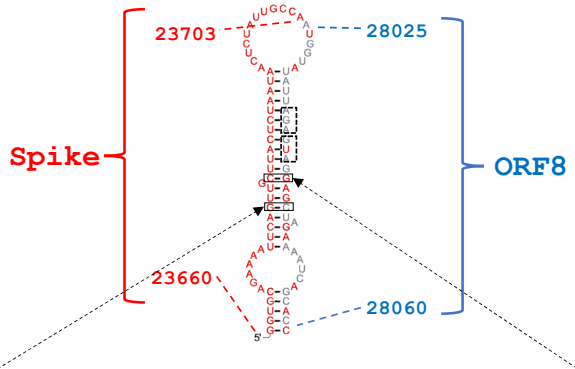
Table 1. Number of variations for each base of predicted base pairing between (23660-23703 *Spike*) and (28025-28060 *ORF8*). 27488 SARS-CoV-2 sequences worldwide retrieved from GISAID. R-scape program was used on the two regions. The top four base pairs with highest substitutions are shown in bold.

Right (*Spike*)	**Left (*ORF8*)**	**Substitutions**	**Power**
23660	28060	8	0.05
23661	28059	8	0.05
23662	28058	4	0
23663	**28057**	**29**	**0.3**
23664	28056	5	0.01
23666	28054	17	0.17
23667	28053	1	0
23671	28050	1	0
23672	28049	5	0.01
23673	**28048**	**41**	**0.41**
23674	28046	1	0
23675	**28045**	**92**	**0.74**
23676	28044	5	0.01
23677	28043	15	0.14
23679	**28042**	**29**	**0.3**
23680	28041	13	0.12
23681	28040	1	0
23682	28039	0	0
23683	28038	14	0.13
23684	28037	10	0.08
23685	28036	4	0
23686	28035	10	0.08
23687	28034	1	0
23688	28033	4	0
23689	28032	4	0
23690	28031	1	0

A)



B)



23675 Spike	28045 ORF8	Count
G	C	27443
G	U	41
A	C	1
G	A	1
G	G	1
G	N	1

23679 Spike	28042 ORF8	Count
C	G	27471
U	G	14
C	N	1
N	G	1
Y	G	1

Figure 3. Covarying substitutions in *Spike-ORF8* interaction. RNA base pairing between (23,660-23,703 *Spike*) and (28,025-28,060 *ORF8*) and major covarying nucleotide substitutions are shown. Size of the resulting dataset considered for R-scape program (msa_0520_23660-23703_28025-28060st) was 27488. Average pairwise similarity was 99.98. Number of predicted base pairs between the two regions was 26. R-scape program was used concatenation of the two regions. Base pairing configurations were according to IntaRNA. R-scape consensus structure prediction was consistent with IntaRNA and other thermodynamic-based predictions on the reference sequence segments. (A) shows the survival function of scores all pairs in the alignment of (23,660-23,703 *Spike*) and (28,025-28,060 *ORF8*). See Methods and R-scape manual for more details. (B) shows the predicted RNA base pairing between the two regions. The top two base pairs with highest substitutions [23,675-28,045] and [23,679-28,042] are shown in tables along with number of observed counts for each variation. Highest variation in each location shown in red. Covariation corresponding to the base pair with highest substitutions, shown in the red box, was [23,675-28,045].

Figure 4A shows the structure prediction between the *ORF8* region and an extension of the *Spike* region, using bifold program. The 12-nt insert appears to be unpaired in the RNA-RNA duplex structure. Structural prediction of (23,600-24,107 in *Spike*) and (17,012-17,046 in *Orf1ab*) is partially shown in Figure 4B, where the corresponding *Orf1ab* region is shown in green. As we can see, the 12-nt insert appears to partially bind to a region on (23,600-24,107 *Spike*) different from originally predicted local structure (Figure 1A). Figure 4C illustrates the approximated three-way RNA structure prediction amongst all three regions (23,600-24,107 in *Spike*), (28,025-28,060 in *ORF8*), and (17,012-17,046 in *Orf1ab*). In this scenario, the 12-nt insert remains unpaired but a new 3-bp pseudo-knot is stabilized.

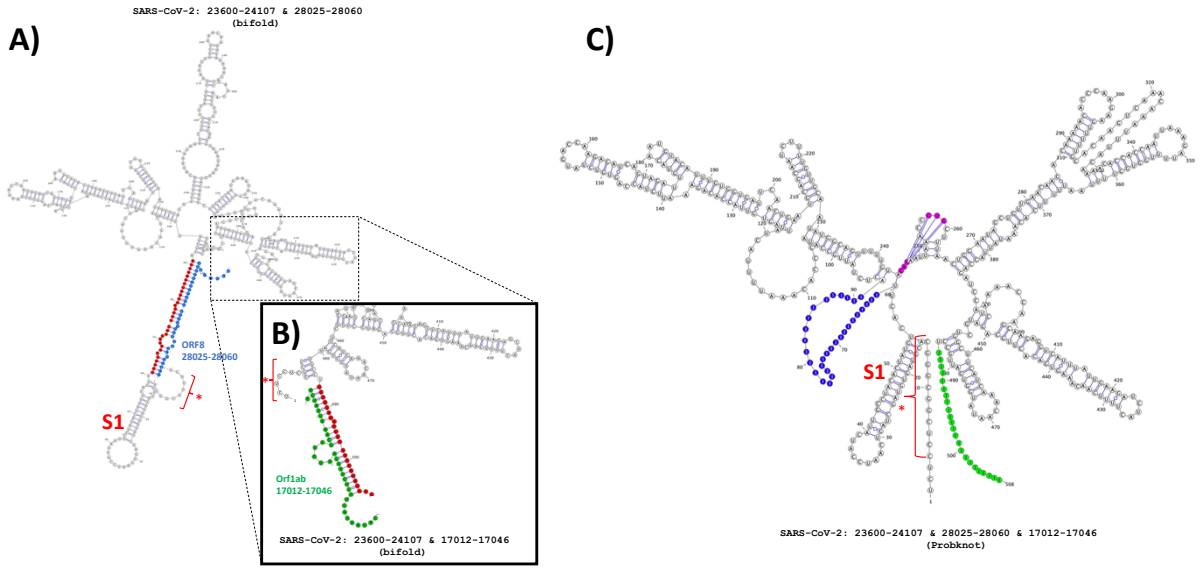


Figure 4. Predicted structures from interaction amongst SARS-CoV-2 (23,600-24,107 in *Spike*), (28,025-28,060 in *ORF8*), and (17,012-17,046 in *Orf1ab*). (A) RNA duplex prediction between region (23,600-24,107 in *Spike*) in black, and region (28,025-28,060 in *ORF8*) in blue. S1 shows Stem 1. (B) RNA duplex prediction between region (23,600-24,107 in *Spike*) in black, and region (17,012-17,046 in *Orf1ab*) in green. (A) and (B): Red nucleotides show regions in Spike interacting with ORF8 and Orf1ab. Predictions performed using bifold program (See Methods). (C) Three-way RNA structure prediction amongst (23600-

24107 *Spike*), (28025-28060 *ORF8*), and (17012-17046 *Orf1ab*). (A), (B), and (C): Asterisk shows location of the 12-bp insert.

Discussion

Betacoronaviruses use long-range genomic RNA-RNA interactions to control sub-genomic transcript production (Sola, Almazán et al. 2015). Mediated by stabilizing proteins, such interactions impact the tertiary structure of the genomic RNA, facilitating binding of the 5' UTR Transcript Regulatory Sequences (TRS) to the TSR upstream of a particular gene, leading to the switching of minus strand template to that of the gene's sub-genomic transcript. Regulation of the *N*-gene sub-genomic transcript is a fair example of such high-order RNA-RNA interactions (Sola, Almazán et al. 2015). Although some efforts have been made to investigate RNA-RNA interactions of SARS-CoV-2 (Ziv, Price et al. 2020), It is generally very difficult to identify all the genomic RNA regions that are involved in such intricate interactions, presenting computational challenges to finding novel interacting regions within the virus (Rouskin, Lan et al. 2021).

The focus of the RNA structure analyses described here was the region corresponding to the 12-nt insert in the *Spike* region, present in SARS-CoV-2 but absent in SARS-CoV. The insert has a high GC content and is immediately followed by a conserved RNA stem-loop immediately downstream of it (Rangan, Zheludev et al. 2020). We found RNA structural differences between corresponding regions in SARS-CoV-2 and SARS-CoV. The locally stable RNA structure containing the 12-insert region was more stable in SARS-CoV-2 than in SARS-CoV with a pseudo-knotted structure not observed in SARS-CoV (Figure 1). In SARS-CoV-2, the downstream region of the 12-nt insert was predicted to base pair with *ORF8*, while in SARS-CoV, the same region preferred the 3'UTR (Figure 2).

The aligned SARS-CoV-2 sequences were investigated for compensatory mutations that might occur within and between *Spike-ORF8* binding location. Although not any significantly covarying mutations were observed, a polymorphism in *ORF8*, C28045U, was mildly in support of the *In-silico* prediction about *Spike-ORF8* RNA-RNA base pairing (Figure 3). Being a synonymous mutation, C28045U has been previously identified as one of the polymorphic positions of *ORF8* (Pereira 2020). In the local RNA secondary structure prediction of *ORF8*, C28045U is unpaired, while in the predicted long-range RNA-RNA interaction with *Spike*, it pairs with G23675. Another observed mutation [C23679U G28042], though non-synonymous, accommodates for the base pairing and occurs on *Spike*. The above discussed predicted pairs are only few base pairs apart. Neither of the two base pairs, however, individually passed the significance threshold (Figure 3). Given the relatively low mutation rate of the virus and given that not any covarying mutations in 5'UTR passed the significance threshold (data not shown) either, it may be possible that the downloaded dataset is not diverse enough to statistically support predicted [G23675 C28045U] and [C23679U G28042] base pairings.

The contribution of the 12-nt insert in stabilizing the *Spike-ORF8* interaction is still unclear. Although the 12-nt insert consists of high GC content, it appears as unpaired in most of the long-range RNA-RNA interacting predictions (Figure 4). This occurs while the 12-nt insert secures base pairing in 500-nt downstream in local structure prediction (Figure 1A). Ironically, Stem S1 remains stabilized in both local and long-range structure predictions (Figures 1A and 4). It is possible that the insert stabilizes a tertiary interaction or binds to a protein or other regions of the virus not

predicted here. In any case, it is unclear if the 12-nt insert contributes to long-range interactions or stabilizes against them.

The predicted *Spike-ORF8* long-range RNA-RNA interaction can potentially impact template switch during negative strand synthesis. Template switch in betacoronaviruses might occur if the TSR element downstream of the 5'UTR is in proximity of the TSR element immediately upstream of a viral gene. Such complex genomic conformation may involve other RNA-RNA as mediators. The dE-pE (Figure 2 of (Sola, Almazán et al. 2015)) act as such mediator RNA binding locations to facilitate a discontinuous negative strand synthesis of the viral genome, leading to N-gene sub-genomic RNA. The coronavirus nucleocapsid (N) is known to be a structural protein that forms complexes with genomic RNA, interacts with the viral membrane protein during virion assembly and plays a critical role in enhancing the efficiency of virus transcription and assembly (Sola, Almazán et al. 2015). The predicted *Spike*-ORF8 binding location here is 200nt upstream of the N-gene TSR (Pereira 2020). Although high-order RNA-RNA interactions needed for template switch can be more complex and may involve the 5'UTR as well, it might be that the predicted *Spike-ORF8* RNA binding is acting as an additional mediator step to bring the TRS elements of 5'UTR and the coronavirus *N*-gene closer to each other. It could be speculated that the sequence immediately downstream of the 12-nt insert are evolved to bind to *ORF8* for the purpose of regulating sub-genomic RNA production. Since the first gene downstream of *Spike-ORF8* binding location happens to be the *N*-gene, the binding location might be affecting the *N*-gene sub-genomic RNA production.

Amongst coronaviruses, *ORF8* is a rapidly evolving hypervariable gene that undergoes deletions to possibly adapt to human host (Gong, Tsao et al. 2020, Pereira 2020, Su, Anderson et al. 2020, Zinzula 2021). It has also been previously observed that patients infected with SARS-CoV-2 variants with a 382-nucleotide deletion (Δ382) in *ORF8* had milder symptoms (Young, Fong et al. 2020). In addition, *ORF8* contains RNA structural features (Pereira 2020). While this observation may very well be due to impact of absence of the translated protein, ORF8 RNA structural characteristics of the genome may also play a role in the viral life cycle, making long-range RNA-RNA prediction with Spike a less remote possibility. A comprehensive exploration of predicted *Spike-ORF8* binding in other SARS-CoV-2 variants and evaluating corresponding sub-genomic RNA production rates of these variants may lead to further clues about the predicted long-range RNA-RNA interaction, which can be rewarding for therapeutic purposes.

Acknowledgements

We thank researches who have kindly shared genomes on public databases and the Global Initiative on Sharing All Influenza Data (GISAID) platform.

Supplementary Information

Text S1. Non-synonymous mutations observed within selected *ORF8* region. The R-scape was applied to regions (23,660-23,703 *Spike*) and (28,025-28,060 *ORF8*). For each SARS-CoV-2 sequence in the dataset, the two regions were cut and appended to each other as a single sequence segment. The secondary structure was chosen by folding the sequence segment corresponding the reference genome (See Supplementary Figure S1). All other approaches including results from IntaRNA, bimolecular folding, and population-based consensus structure prediction via CaCoFold had almost identical structures. The predicted structure was passed to R-scape for assessing

covariation of its base pairs across the population of sequences referred to here as `msa_0520_23660-23703_28025-28060st`. Sequence identity was 99.98 across the population of 27,488 (certain sequences with gaps were eliminated from the dataset with original size of 27,592). Three significantly covarying nucleotides were observed across the population, none of which belonged to the predicted structure and all of which belonged to *ORF8* region. All the three significantly covarying mutations had E-value of 0.02975. These significantly covarying mutations referred to non-synonymous mutations for the *ORF8* coding region. From amongst the 27,488 viral sequences analyzed here, two sequences from patients from Ecuador and one from Canada had different amino acids on two consecutive locations on ORF8: ORF8 AA (48-49), corresponding to genomic coordinates (28,035-28,040 *ORF8*). Supplementary Figure S2 shows the location of the two codons Supplementary Table S1 describe the statistical figures associated with the above significantly covarying nucleotides, given by R-scape. Supplementary Table S2 denotes information about individuals carrying the viruses with the above codon variations.

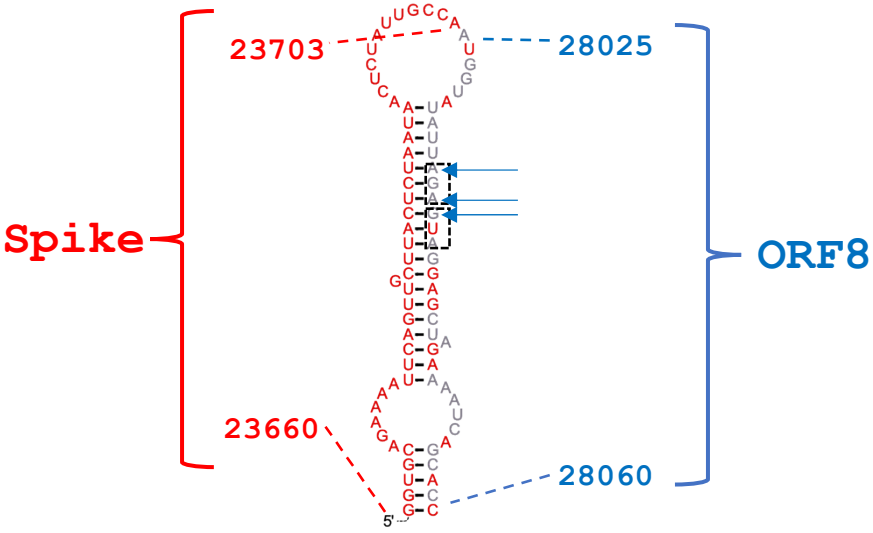


Figure S1. Covarying mutations analysis of SARS-CoV-2 (23,660-23,703 *Spike*) and (28,025-28,060 *ORF8*). R-scape was used to investigate covarying mutations. (A) shows the consensus pairing between the two regions. Blue arrows point to position of significantly covarying mutations A28035, A28037, and G28038. Dotted black squares represent codon positioning. (B) shows the covariation significance between the three pairs. (C) shows covariation score of nucleotide pairs (blue) is shown against a random background (black). Total number of sequences was of 27488. Average pairwise similarity was 99.98.

Table S1. Significantly nucleotide pairs across 27488 SARS-CoV-2 sequences worldwide retrieved from GISAID in the (23660-23703 *Spike*) and (28025-28060 *ORF8*). R-scape program was used. The nucleotides within two codons (28035-28040 *ORF8*) were found to be significantly varying.

Left position	Right Position	Score	E-value	Substitutions	Power

61	64	3662.87682	0.0297468	20	0.2
63	64	3662.87682	0.0297468	20	0.2
61	63	3662.87682	0.0297468	20	0.2

Table S2. Codons with significantly covarying mutations in 27488 SARS-CoV-2 sequences worldwide retrieved from GISAID in the (23660-23703 *Spike*) and (28025-28060 *ORF8*). R-scape program was used on regions (23660-23703 *Spike*) and (28025-28060 *ORF8*). The nucleotides within two codons (28035-28040 *ORF8*) were found to be significantly varying.

GISAID ID	**Count**	**(28035-28040 *ORF8*)**	**ORF8 AA (48-49)**
*hCoV-19/worldwide	27485	AGA GUA	R V
*hCoV-19/Ecuador/HGSQ-USFQ-010/2020|EPI_ISL_422565|2020-03-30|SouthAmerica *hCoV-19/Ecuador/HGSQ-USFQ-007/2020|EPI_ISL_422564|2020-03-30|SouthAmerica	2	GGG UUA	G L
*hCoV-19/Canada/ON_PHL1898/2020|EPI_ISL_418377|2020-03-13|NorthAmerica	1	UAU AUU	Y I

References

Andrews, R. J., J. M. Peterson, H. S. Haniff, J. Chen, C. Williams, M. Grefe, M. D. Disney and W. N. Moss (2020). "An in silico map of the SARS-CoV-2 RNA Structurome." bioRxiv: 2020.2004.2017.045161.

Bartas, M., V. Brázda, N. Bohálová, A. Cantara, A. Volná, T. Stachurová, K. Malachová, E. B. Jagelská, O. Porubiaková, J. Červeň and P. Pečinka (2020). "In-Depth Bioinformatic Analyses of Nidovirales Including Human SARS-CoV-2, SARS-CoV, MERS-CoV Viruses Suggest Important Roles of Non-canonical Nucleic Acid Structures in Their Lifecycles." Frontiers in microbiology **11**: 1583-1583.

Bellaousov, S. and D. H. Mathews (2010). "ProbKnot: fast prediction of RNA secondary structure including pseudoknots." RNA (New York, N.Y.) **16**(10): 1870-1880.

Busch, A., A. S. Richter and R. Backofen (2008). "IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions." Bioinformatics **24**(24): 2849-2856.

Darty, K., A. Denise and Y. Ponty (2009). "VARNA: Interactive drawing and editing of the RNA secondary structure." Bioinformatics **25**(15): 1974-1975.

Gong, Y. N., K. C. Tsao, M. J. Hsiao, C. G. Huang, P. N. Huang, P. W. Huang, K. M. Lee, Y. C. Liu, S. L. Yang, R. L. Kuo, K. F. Chen, Y. C. Liu, S. Y. Huang, H. I. Huang, M. T. Liu, J. R. Yang, C. H. Chiu, C. T. Yang, G. W. Chen and S. R. Shih (2020). "SARS-CoV-2 genomic surveillance in Taiwan revealed

novel ORF8-deletion mutant and clade possibly associated with infections in Middle East." Emerg Microbes Infect **9**(1): 1457-1466.
Hu, B., H. Guo, P. Zhou and Z. L. Shi (2020). "Characteristics of SARS-CoV-2 and COVID-19." Nat Rev Microbiol.
Huston, N. C., H. Wan, M. S. Strine, R. de Cesaris Araujo Tavares, C. B. Wilen and A. M. Pyle (2021). "Comprehensive in vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms." Molecular Cell **81**(3): 584-598.e585.
Kim, D., J.-Y. Lee, J.-S. Yang, J. W. Kim, V. N. Kim and H. Chang (2020). "The Architecture of SARS-CoV-2 Transcriptome." Cell **181**(4): 914-921.e910.
Lan, T. C. T., M. F. Allan, L. E. Malsick, S. Khandwala, S. S. Y. Nyeo, M. Bathe, A. Griffiths and S. Rouskin (2020). "Structure of the full SARS-CoV-2 RNA genome in infected cells." bioRxiv: 2020.2006.2029.178343.
Lorenz, R., S. H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler and I. L. Hofacker (2011). "ViennaRNA Package 2.0." Algorithms for Molecular Biology **6**(1): 26.
Lu, R., X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu, Y. Bi, X. Ma, F. Zhan, L. Wang, T. Hu, H. Zhou, Z. Hu, W. Zhou, L. Zhao, J. Chen, Y. Meng, J. Wang, Y. Lin, J. Yuan, Z. Xie, J. Ma, W. J. Liu, D. Wang, W. Xu, E. C. Holmes, G. F. Gao, G. Wu, W. Chen, W. Shi and W. Tan (2020). "Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding." The Lancet **395**(10224): 565-574.
Lu, Z. J., J. W. Gloor and D. H. Mathews (2009). "Improved RNA secondary structure prediction by maximizing expected pair accuracy." RNA (New York, N.Y.) **15**(10): 1805-1813.
Manfredonia, I., C. Nithin, A. Ponce-Salvatierra, P. Ghosh, T. K. Wirecki, T. Marinus, N. S. Ogando, Eric J. Snijder, M. J. van Hemert, J. M. Bujnicki and D. Incarnato (2020). "Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements." Nucleic Acids Research **48**(22): 12436-12452.
Mann, M., P. R. Wright and R. Backofen (2017). "IntaRNA 2.0: enhanced and customizable prediction of RNA–RNA interactions." Nucleic Acids Research **45**(W1): W435-W439.
Mathews, D. H. (2004). "Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization." RNA (New York, N.Y.) **10**(8): 1178-1190.
Mathews, D. H., M. E. Burkard, S. M. Freier, J. R. Wyatt and D. H. Turner (1999). "Predicting oligonucleotide affinity to nucleic acid targets." RNA **5**(11): 1458-1469.
McCaskill, J. S. (1990). "The equilibrium partition function and base pair binding probabilities for RNA secondary structure." Biopolymers **29**(6-7): 1105-1119.
Naqvi, A. A. T., K. Fatima, T. Mohammad, U. Fatima, I. K. Singh, A. Singh, S. M. Atif, G. Hariprasad, G. M. Hasan and M. I. Hassan (2020). "Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach." Biochimica et biophysica acta. Molecular basis of disease **1866**(10): 165878-165878.
Pereira, F. (2020). "Evolutionary dynamics of the SARS-CoV-2 ORF8 accessory gene." Infect Genet Evol **85**: 104525.
Pereira, F. (2020). "Evolutionary dynamics of the SARS-CoV-2 ORF8 accessory gene." Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases **85**: 104525-104525.

Raden, M., S. M. Ali, O. S. Alkhnbashi, A. Busch, F. Costa, J. A. Davis, F. Eggenhofer, R. Gelhausen, J. Georg, S. Heyne, M. Hiller, K. Kundu, R. Kleinkauf, S. C. Lott, M. M. Mohamed, A. Mattheis, M. Miladi, A. S. Richter, S. Will, J. Wolff, P. R. Wright and R. Backofen (2018). "Freiburg RNA tools: a central online resource for RNA-focused research and teaching." Nucleic Acids Research **46**(W1): W25-W29.

Rangan, R., I. N. Zheludev and R. Das (2020). "RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses." bioRxiv : the preprint server for biology: 2020.2003.2027.012906.

Reuter, J. S. and D. H. Mathews (2010). "RNAstructure: software for RNA secondary structure prediction and analysis." BMC Bioinformatics **11**(1): 129.

Rivas, E. (2020). "RNA structure prediction using positive and negative evolutionary information." bioRxiv: 2020.2002.2004.933952.

Rivas, E., J. Clements and S. R. Eddy (2017). "A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs." Nature Methods **14**(1): 45-48.

Rivas, E., J. Clements and S. R. Eddy (2020). "Estimating the power of sequence covariation for detecting conserved RNA structure." Bioinformatics **36**(10): 3072-3076.

Rivas, E. and S. R. Eddy (2020). "Response to Tavares et al., “Covariation analysis with improved parameters reveals conservation in lncRNA structures”." bioRxiv: 2020.2002.2018.955047.

Rouskin, S., T. Lan, M. Allan, L. Malsick, S. Khandwala, S. Nyeo, Y. Sun, J. Guo, M. Bathe and A. Griffiths (2021). Insights into the secondary structural ensembles of the full SARS-CoV-2 RNA genome in infected cells, Research Square.

Simmonds, P. (2020). "Pervasive RNA Secondary Structure in the Genomes of SARS-CoV-2 and Other Coronaviruses." mBio **11**(6): e01661-01620.

Snijder, E. J., R. W. A. L. Limpens, A. H. de Wilde, A. W. M. de Jong, J. C. Zevenhoven-Dobbe, H. J. Maier, F. F. G. A. Faas, A. J. Koster and M. Bárcena (2020). "A unifying structural and functional model of the coronavirus replication organelle: Tracking down RNA synthesis." PLoS biology **18**(6): e3000715-e3000715.

Sola, I., F. Almazán, S. Zúñiga and L. Enjuanes (2015). "Continuous and Discontinuous RNA Synthesis in Coronaviruses." Annual review of virology **2**(1): 265-288.

Sola, I., F. Almazán, S. Zúñiga and L. Enjuanes (2015). "Continuous and Discontinuous RNA Synthesis in Coronaviruses." Annu Rev Virol **2**(1): 265-288.

Su, Y. C. F., D. E. Anderson, B. E. Young, M. Linster, F. Zhu, J. Jayakumar, Y. Zhuang, S. Kalimuddin, J. G. H. Low, C. W. Tan, W. N. Chia, T. M. Mak, S. Octavia, J.-M. Chavatte, R. T. C. Lee, S. Pada, S. Y. Tan, L. Sun, G. Z. Yan, S. Maurer-Stroh, I. H. Mendenhall, Y.-S. Leo, D. C. Lye, L.-F. Wang, G. J. D. Smith and S. Schultz-Cherry (2020). "Discovery and Genomic Characterization of a 382-Nucleotide Deletion in ORF7b and ORF8 during the Early Evolution of SARS-CoV-2." mBio **11**(4): e01610-01620.

Wright, P. R., J. Georg, M. Mann, D. A. Sorescu, A. S. Richter, S. Lott, R. Kleinkauf, W. R. Hess and R. Backofen (2014). "CopraRNA and IntaRNA: predicting small RNA targets, networks and interaction domains." Nucleic Acids Research **42**(W1): W119-W123.

Young, B. E., S.-W. Fong, Y.-H. Chan, T.-M. Mak, L. W. Ang, D. E. Anderson, C. Y.-P. Lee, S. N. Amrun, B. Lee, Y. S. Goh, Y. C. F. Su, W. E. Wei, S. Kalimuddin, L. Y. A. Chai, S. Pada, S. Y. Tan, L. Sun, P. Parthasarathy, Y. Y. C. Chen, T. Barkham, R. T. P. Lin, S. Maurer-Stroh, Y.-S. Leo, L.-F. Wang, L. Renia, V. J. Lee, G. J. D. Smith, D. C. Lye and L. F. P. Ng (2020). "Effects of a major

deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: an observational cohort study." The Lancet **396**(10251): 603-611.
Zhang, L., H. Zhang, D. H. Mathews and L. Huang (2020). "ThreshKnot: Thresholded ProbKnot for Improved RNA Secondary Structure Prediction." arXiv:1912.12796 [physics, q-bio].
Zinzula, L. (2021). "Lost in deletion: The enigmatic ORF8 protein of SARS-CoV-2." Biochemical and biophysical research communications **538**: 116-124.
Ziv, O., J. Price, L. Shalamova, T. Kamenova, I. Goodfellow, F. Weber and E. A. Miska (2020). "The Short- and Long-Range RNA-RNA Interactome of SARS-CoV-2." Molecular Cell **80**(6): 1067-1077.e1065.