How the replication and transcription complex of SARS-CoV-2 functions in leader-to-body fusion

Xin Li[1$], Qiang Zhao[1$], Jia Chang[1], Guangyou Duan[2], Jinlong Bei[3]

Tung On Yau[4], Jianyi Yang[5], Jishou Ruan[5], Bingjun He[1]*, Shan Gao[1]*

[1] College of Life Sciences, Nankai University, Tianjin, Tianjin 300071, P.R.China;

[2] School of Life Sciences, Qilu Normal University, Jinan, Shandong 250200, P.R.China;

[3] Guangdong Provincial Key Laboratory for Crop Germplasm Resources Preservation and Utilization, Agro-Biological Gene Research Center, Guangdong Academy of Agricultural Sciences, Guangzhou, Guangdong 510640, P. R. China;

[4] John Van Geest Cancer Research Centre, School of Science and Technology, Nottingham Trent University, Nottingham, NG11 8NS, United Kingdom;

[5] School of Mathematical Sciences, Nankai University, Tianjin, Tianjin 300071, P.R.China.

[$] These authors contributed equally to this paper.

* Corresponding authors.

SG：gao_shan@mail.nankai.edu.cn

BH：hebj@nankai.edu.cn

Abstract

Background: Coronavirus disease 2019 (COVID-19) is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Although unprecedented efforts are underway to develop therapeutic strategies against this disease, scientists have acquired only a little knowledge regarding the structures and functions of the CoV replication and transcription complex (RTC) and 16 non-structural proteins, named NSP1-16.

Results: In the present study, we determined the theoretical arrangement of NSP12-16 in the global RTC structure. This arrangement answered how the CoV RTC functions in the "leader-to-body fusion" process. More importantly, our results revealed the associations between multiple functions of the RTC, including RNA synthesis, NSP15 cleavage, RNA methylation, and CoV replication and transcription at the molecular level. As the most important finding, transcription regulatory sequence (TRS) hairpins were reported for the first time to help understand the multiple functions of CoV RTCs and the strong recombination abilities of CoVs.

Conclusions: TRS hairpins can be used to identify recombination regions in CoV genomes. We provide a systematic understanding of the structures and functions of the RTC, leading to the eventual determination of the global CoV RTC structure. Our findings enrich fundamental knowledge in the field of gene expression and its regulation, providing a basis for future studies. Future drug design targeting SARS-CoV-2 needs to consider protein-protein and protein-RNA interactions in the RTC, particularly the complex structure of NSP15 and NSP16 with the TRS hairpin.

Keyword: coronavirus; RNA methylation; NSP12; NSP15; NSP16

Introduction

Coronavirus disease 2019 (COVID-19) is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1] [2]. SARS-CoV-2 has a genome of ~30 kb [3], including the genes *spike (S)*, *envelope (E)*, *membrane (M)*, *nucleocapsid (N)*, and *ORF1a*, *1b*, *3a*, *6*, *7a*, *7b*, *8* and *10* [4]. The *ORF1a* and *1b* genes encode 16 non-structural proteins (NSPs), named NSP1 through NSP16 [5]. The other 10 genes encode 4 structural proteins (S, E, M and N) and 6 accessory proteins (ORF3a, 6, 7a, 7b, 8 and 10). NSP4-16 are significantly conserved in all known CoVs and have been experimentally demonstrated or predicted to be critical enzymes in CoV RNA synthesis and modification [6], including: NSP12, RNA-dependent RNA polymerase (RdRp) [7]; NSP13, RNA helicase-ATPase (Hel); NSP14, exoribonuclease (ExoN), and methyltransferase; NSP15 endoribonuclease (EndoU) [8]; and NSP16, RNA 2-O-methyltransferase (MT).

NSP1-16 assemble into a replication and transcription complex (RTC) in CoV [7]. The basic function of the RTC is RNA synthesis: it synthesizes genomic RNAs (gRNAs) for replication and subgenomic RNAs (sgRNAs) for transcription [9]. CoV replication and transcription can be explained by the prevailing “leader-to-body fusion” model [9]. For a complete understanding of CoV replication and transcription, much research has been conducted to determine the global structure of the SARS-CoV-2 RTC, since the outbreak of SARS-CoV-2. Although some single protein structures (e.g. NSP15 [8]) and local structures of the RTC (i.e. NSP7-NSP8-NSP12-NSP13 [7] and NSP7-NSP8-NSP12 [10]) have been determined, these structures have not yet answered how the RTC functions in the “leader-to-body fusion” process. As the global structure of the CoV RTC cannot be determined by simple use of any current methods (i.e., NMR, X-ray and Cryo-EM), it is necessary to ascertain the arrangement of all the RTC components, leading to the eventual determination of its global structure.

In our previous study, we provided a molecular basis to explain the "leader-to-body fusion" model by identifying the cleavage sites of NSP15 and proposed a negative feedback model to explain the regulation of CoV replication and transcription. In the present study, we aimed to determine the theoretical arrangement of NSP12-16 in the global structure of the CoV RTC by comprehensive analysis of data from different sources, and to elucidate the functions of CoV RTC in the "leader-to-body fusion" process.

Results

Molecular basis of "leader-to-body fusion" model

Here, we provide a brief introduction to the "leader-to-body fusion" model proposed in an early study [9] and its molecular basis proposed in our recent study [11]. CoV replication and transcription require gRNAs(+) as templates for the synthesis of antisense genomic RNAs [gRNAs(-)] and antisense subgenomic RNAs [sgRNAs(-)] by RdRP. When RdRP pauses, as it crosses a body transcription regulatory sequence (TRS-B) and switches the template to the leader TRS (TRS-L), sgRNAs(-) are formed through discontinuous transcription (also referred to as polymerase jumping or template switching). Otherwise, RdRP reads gRNAs(+) continuously, without interruption, resulting in gRNAs(-). Thereafter, gRNAs(-) and sgRNAs(-) are used as templates to synthesize gRNAs(+) and sgRNAs(+), respectively; gRNAs(+) and sgRNAs(+) are used as templates for the translation of NSP1-16 and 10 other proteins (S, E, M, N, and ORF3a, 6, 7a, 7b, 8 and 10), respectively. The molecular basis of the "leader-to-body fusion" model as proposed in our previous study is that NSP15 cleaves gRNAs(-) and sgRNAs(-) at TRS-Bs(-). Then, the free 3' ends (~6 nt) of TRS-Bs(-) hybridize the junction regions of TRS-Ls for template switching. NSP15 may also cleave gRNAs(-) and sgRNAs(-) at TRS-Ls(-), which is not necessary for template switching. This molecular basis

preliminarily answered how the RTC functions in the "leader-to-body fusion" process. However, the associations between NSP12, NSP15 and other 14 NSPs are still unknown.

NSP15 cleavage, RNA methylation and 3' polyadenylation

The cleavage sites of NSP15 contain the canonical TRS motif "GTTCGT" [11], read on the antisense strands of CoV genomes. One study reported that RNA methylation sites contain the "AAGAA-like" motif (including AAGAA and other A/G-rich sequences) throughout the viral genome, particularly enriched in genomic positions 28,500-29,500 [4]. Nanopore RNA-seq, a direct RNA sequencing method [12], was used in the study that shares data for our reanalysis in the previous [11] and present studies. By reanalysing the Nanopore RNA-seq data [4], we found that the "AAGAA-like" motif co-occurred with the canonical TRS motif "GTTCGT" (**Figure 1A**) in TRS-Bs of eight genes (*S, E, M, N,* and *ORF3a, 6, 7a* and *8*). In addition, each of these TRS-B contains many possible hairpins. These hairpins are encoded by complemented palindrome sequences, which explained a finding reported in our previous study: complemented palindromic small RNAs (cpsRNAs) with lengths ranging from 14 to 31 nt are present throughout the SARS-CoV genome, however, most of them are semipalindromic or heteropalindromic [13].

In the present study, we defined the hairpins containing the canonical and non-canonical TRS motif as canonical and non-canonical TRS hairpins, respectively. In addition, we defined the hairpins opposite to the TRS hairpins as opposite TRS hairpins (**Figure 1A**). As the global structure of CoV RTC is asymmetric (**Figure 1B**), opposite TRS hairpins may not be present. All the complemented palindrome sequences in TRS hairpins are semipalindromic or heteropalindromic. By analysing the junction regions between TRS-Bs and the TRS-L of SARS-CoV-2, we found that NSP15 cleaves the canonical TRS hairpin of *ORF3a* at an unexpected breakpoint "GTTCGTTTAT|N" (the vertical line indicates the breakpoint and N indicates any

nucleotide base), rather than the end of the canonical TRS motif “GTTCGT|TTATN”. Here, we defined the breakpoints “GTTCGT|TTATN” and “GTTCGTTTAT|N” as canonical and non-canonical TRS breakpoints, respectively. The discovery of this non-canonical TRS breakpoint indicated that the recognition of NSP15 cleavage sites is structure-based rather than sequence-based. In addition, we found non-canonical TRS hairpins in many non-canonical junction regions [11]. Thus, we proposed a hypothesis that CoV recombinant events occurred due to the cleavage of non-canonical TRS hairpins. Then, we validated that non-canonical TRS hairpins are present in 7 recombination regions that in the *ORF1a*, *S* and *ORF8* genes, using 292 genomes of betacoronavirus subgroup B, in our previous study [14]. Non-canonical TRS hairpins are also present in 5 typical recombination regions (**Figure 2**) analyzed in our previous study [11]. Therefore, TRS hairpins can be used to identify recombination regions in CoV genomes.

Another important phenomenon reported in the previous study [4] that merits further analysis was that the “AAGAA-like” motif associates with the 3' poly(A) lengths of gRNAs and sgRNAs. However, the study did not investigate the association between the “AAGAA-like” motif and the nascent RNAs cleaved by NSP15. The methylation at the “AAGAA-like” motif (read on the antisense strands of CoV genomes) in TRS hairpins may affect the downstream 3' polyadenylation that prevents the quick degradation of nascent RNAs (**Figure 1B**). Although the type of methylation is unknown, a preliminary study was conducted, which revealed that modified RNAs of SARS-CoV-2 have shorter 3' poly(A) tails than unmodified ones [4]. However, there are two shortcomings in the interpretation of CoV RNA methylation in this previous study: (1) it was not explained that many methylation sites are far from 3' ends, which are unlikely to contributes to the3' poly(A) tails; and (2) the “AAGAA-like” motif on the antisense strand was not analyzed (**See below**), as only a few antisense reads were obtained using Nanopore RNA-seq. Although the associations between NSP15 cleavage, RNA methylation and 3' polyadenylation are

still unclear, they suggest that the RTC has a local structure composed of NSP15 and 16 to contain a TRS hairpin. This special local structure is able to facilitate the NSP15 cleavage and RNA methylation of the TRS hairpin at the opposite sides (**Figure 1B**).

How RTC functions in "leader-to-body fusion"

Since several A-rich and T-rich regions are alternatively present in each TRS-B, it contains many possible hairpins. Thus, to determining which one is the TRS hairpin needs decisive information. After comparing all possible hairpins in the TRS-Bs of betacoronavirus subgroup B, we found that they can be classified into three classes. Using the TRS-B of the S gene of SARS-CoV-2 as an example, the first class (**Figure 3A**) and the third class (**Figure 3C**) require the "AAGAA-like" motif involved in the Watson-Crick pairing. However, the methylation at the "AAGAA-like" motif is not in favour of Watson-Crick pairing. Further analysis of the "AAGAA-like" motif on the antisense strand (**See above**) inspired us to propose a novel explanation of CoV RNA methylation. RNA methylation of CoVs participates in the determination of the RNA secondary structures by affecting the formation of hairpins. The methylation of flanking sequences containing the "AAGAA-like" motif ensures that the NSP15 cleavage site resides in the loops of the second class of hairpins (**Figure 3B**). Therefore, the second class of hairpin structures is the best choice for both the NSP15 cleavage and the "AAGAA-like" motif. The NSP15 cleavage site exposed in a small loop, which facilitates the contacts of NSP15. This structure verified the results of mutation experiments in a previous study [15] that the recognition of NSP15 cleavage sites is independent on the TRS motif, but dependent on its context. These findings confirmed that the recognition of NSP15 cleavage sites is structure-based rather than sequence-based (**See above**).

NSP12-16 form the main structure of the RTC and work as a pipeline (**Figure 1B**). The RTC pipeline starts with NSP13 that unwind template RNAs [7]. Using single-strand templates, NSP12 synthesizes RNAs with error correction by NSP14.

Then, the nascent RNAs are methylated, respectively. At last, TRS hairpins are cleaved by NSP15 under specific conditions. Based on the available protein structure data, NSP7 and NSP8, acting as the cofactors of nsp12, assemble the central RTC [7]. The results of biological experiments suggest that NSP8 is able to interact with NSP15 [16]. Therefore, the hexameric NSP15 [8] connects to NSP8 in the global structure of CoV RTC. However, what conditions or local structures decide whether NSP15 cleave the nascent RNAs is still unknown. Another unknown topic is which enzyme is responsible for the methylation at the "AAGAA-like" motif. A recent study reported that NSP16-NSP10 (PDB: 7BQ7), as 2'-O-RNA methyltransferase (MTase), is crucial for RNA cap formation [17]. Although the previous study excluded METTL3-mediated m6A (for lack of canonical motif RRACH), there is still a possibility that METTL3 or its family members function for the methylation at the "AAGAA-like" motif. Another possibility is that NSP10 methylate guanosines in both the caps and the "AAGAA-like" motif. More molecular experiments need be conducted to verify these findings and inferences. The key step leading to the proposal of the arrangement of NSP12-16 in the global RTC structure was that NSP15 cleavage sites are associated to RNA methylation sites. The arrangement of NSP12-16 was proposed mainly due to the integration of information from many aspects, particularly considering: (1) the identification of NSP15 cleavage sites in our previous study [13]; (2) TRS hairpins eight genes (*S*, *E*, *M*, *N*, and *ORF3a*, *6*, *7a* and *8*) are conserved in 292 genomes of betacoronavirus subgroup B; (3) the associations between NSP15 cleavage, RNA methylation and 3' polyadenylation; (4) *ORF1b*, without recombination regions, is much more conservative than *ORF1a*, with two recombination regions, in 292 genomes of betacoronavirus subgroup B [14]; and (5) the extremely high ratio between sense and antisense reads.

Conclusion and Discussion

In the present study, we determined the theoretical arrangement of NSP12-16 in the global RTC structure. This arrangement answered how the CoV RTC functions in the "leader-to-body fusion" process. Our model did not rule out the involvement of other proteins (e.g., NSP7) in the global RTC structure or the "leader-to-body fusion" process. More importantly, our results revealed the associations between multiple functions of the RTC, including RNA synthesis, NSP15 cleavage, RNA methylation, and CoV replication and transcription, at the molecular level. Future research needs to be conducted to determine the structures of NSP12&14, NSP12&15 and NSP15&16&TRS haripin by Cryo-EM. These local RTC structures can be used to assemble a global RTC structure by protein-protein docking calculation, particularly using deep learning methods. Future drug design targeting SARS-CoV-2 needs to consider protein-protein and protein-RNA interactions, particularly the contacts between NSP15, NSP16 and the stem in the TRS hairpin.

Materials and Methods

1,265 genome sequences of betacoronaviruses (in subgroups A, B, C and D) were downloaded from the NCBI Virus database (https://www.ncbi.nlm.nih.gov/labs/virus) in our previous study [12]. Among these genomes, 292 belongs to betacoronavirus subgroup B (including SARS-CoV and SARS-CoV-2). Nanopore RNA-seq data was downloaded from the website (https://osf.io/8f6n9/files/) for reanalysis. Protein structure data (PDB: 6X1B, 7BQ7, 7CXN) were used to analyzed NSP15, NSP10-NSP16 and NSP7-NSP8-NSP12-NSP13, respectively. SARS-CoV were detected from 4 runs of small RNA-seq data (NCBI SRA: SRR452404, SRR452406, SRR452408 and SRR452410). Data cleaning and quality control were performed using Fastq_clean [18]. Statistics and plotting were conducted using the software R v2.15.3 with the Bioconductor packages [19]. The structures of NSP12-16 were predicted using trRosetta [20].

Supplementary information

Declarations

Ethics approval and consent to participate

Not applicable.

Consent to publish

Not applicable.

Availability of data and materials

All data used in the present study was download from the public data sources.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Natural Science Foundation of China (31871992) to Bingjun He, Tianjin Key Research and Development Program of China (19YFZCSY00500) to Shan Gao and National Natural Science Foundation of China (31700787) to Guangyou Duan. The funding bodies played no role in the study design, data collection, analysis, interpretation or manuscript writing.

Authors' contributions

Shan Gao conceived the project. Shan Gao and Bingjun He supervised this study. Guangyou Duan and Jia Chang performed programming. Xin Li and Qiang Zhao,

Jinlong Bei and Zhenguang Chai downloaded, managed and processed the data. Jianyi Yang predicted the protein structures. Shan Gao drafted the main manuscript text. Shan Gao and Jishou Ruan revised the manuscript.

Acknowledgments

We are grateful for the help from the following faculty members of College of Life Sciences at Nankai University: Xuetao Cao, Deling Kong, Quan Chen, Wenjun Bu, Tao Zhang, Dawei Huang, Mingqiang Qiao, Yanqiang Liu and Zhen Ye. We would like to thank Editage (www.editage.cn) for polishing part of this manuscript in English language. This manuscript was online as a preprint on Feb 5th, 2021 at https://www.researchgate.net/publication/349054954_How_the_replication_and_transcription_complex_of_SARS-CoV-2_functions_in_leader-to-body_fusion.

REFERENCES

[1] X. Li, G. Duan, W. Zhang, J. Shi, J. Chen, S. Chen, S. Gao, and J. Ruan. A Furin Cleavage Site Was Discovered in the S Protein of the 2019 Novel Coronavirus. Chinese Journal of Bioinformatics (In Chinese) 2020, 18(2): 103-108.

[2] G. Duan, J. Shi, Y. Xuan, J. Chen, C. Liu, J. Ruan, S. Gao, and X. Li. 5' UTR Barcode of the 2019 Novel Coronavirus Leads to Insights into Its Virulence. Chinese Journal of Virology (In Chinese) 2020, 36(3): 365-369.

[3] C. Jiayuan, S.Jinsong , O. Yau Tung, L.Chang, L. Xin, Z.Qiang, R. Jishou, and G. Shan. Bioinformatics Analysis of the 2019 Novel Coronavirus Genome. Chinese Journal of Bioinformatics (In Chinese) 2020, 18(2): 96-102.

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

[5] S.J.R. da Silva, C.T. Alves da Silva, R.P.G. Mendes, and L. Pena. Role of nonstructural proteins in the pathogenesis of SARS-CoV-2. J Med Virol 2020, 92: 1427-1429.

[6] Denison MR, Graham RL, Donaldson EF, Eckerle LD, Baric RS.Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity. RNA Biol. 2011, 8(2):270-279.

[7] L. Yan, Y. Zhang, J. Ge, L. Zheng, Y. Gao, T. Wang, Z. Jia, H. Wang, Y. Huang, M. Li, Q. Wang, Z. Rao, and Z. Lou, Architecture of a SARS-CoV-2 mini replication and transcription complex. Nature Communications 2020, 2020(2020): 1-6.

[8] Y. Kim, R. Jedrzejczak, N.I. Maltseva, M. Wilamowski, M. Endres, A. Godzik, K. Michalska, and A. Joachimiak. Crystal structure of NSP15 endoribonuclease NendoU from SARS-CoV-2. Protein Science 2020, 29(7): 1596-1605.

[9] S.G. Sawicki, and D.L. Sawicki. A New Model for Coronavirus Transcription. in: L. Enjuanes, S.G. Siddell, and W. Spaan, (Eds.). Coronaviruses and Arteriviruses, Springer US, Boston, MA, 1998, pp. 215-219.

[10] H.S. Hillen, G. Kokic, L. Farnung, C. Dienemann, and P. Cramer. Structure of replicating SARS-CoV-2 polymerase. Nature 2020, 584(7819): 1-6.

[11] Xin Li, Zhi Cheng, Fang Wang, Jia Chang, Qiang Zhao, Hao Zhou, Chang Liu, Jishou Ruan, Guangyou Duan, Shan Gao. A negative feedback model to explain regulation of SARS-CoV-2 replication and transcription. Frontiers in Genetics 2021, 10: 1-11.

[12] X. Xu, H. Ji, X. Jin, Z. Cheng, X. Yao, Y. Liu, Q. Zhao, T. Zhang, J. Ruan, W. Bu, Z. Chen, and S. Gao. Using pan RNA-seq analysis to reveal the ubiquitous existence of 5' and 3' end small RNAs. Frontiers in Genetics 2019, 10: 1-11.

[13] Liu C, Chen Z, Hu Y, Ji H, Yu D, Shen W, Li S, Ruan J, Bu W, Gao S. Complemented Palindromic Small RNAs First Discovered from SARS Coronavirus. Genes 2018, 9(9): 1-11.

[14] Xin Li, Jia Chang, Shunmei Chen, Liangge Wang, Tung On Yau, Qiang Zhao, Zhangyong Hong, Jishou Ruan, Guangyou Duan and Shan Gao. Genomic feature analysis of betacoronavirus provides insights into SARS and COVID-19 pandemics. bioRxiv. 2020(2020): 1-8.

[15] Yount B , Roberts R , Lindesmith L, et al. Rewiring the severe acute respiratory syndrome coronavirus (SARS-CoV) transcription circuit: Engineering a recombination-resistant genome[J]. Proceedings of the National Academy of Sciences of the United States of America, 2006, 103(33) : 12546–12551.

[16] Lianqi Z , Lei L , Liming Y, et al. Structural and Biochemical Characterization of Endoribonuclease Nsp15 Encoded by Middle East Respiratory Syndrome Coronavirus. Journal of Virology, 2018, 92.

[17] Krafcikova P , Silhan J , Nencka R , et al. Structural analysis of the SARS-CoV-2 methyltransferase complex involved in RNA cap creation bound to sinefungin. Nature Communications, 2020, 11(1):3717.

[18] M. Zhang, F. Zhan, H. Sun, X. Gong, Z. Fei, and S. Gao. Fastq_clean: An optimized pipeline to clean the Illumina sequencing data with quality control, Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on, IEEE, 2014, pp. 44-48.

[19] S. Gao, J. Ou, and K. Xiao. R language and Bioconductor in bioinformatics applications(Chinese Edition), Tianjin Science and Technology Translation Publishing Ltd, Tianjin, 2014.

[20] Yang J, Anishchenko I, Park H, Peng Z, Baker D. Improved protein structure prediction using predicted interresidue orientations. Proceedings of the National Academy of Sciences 2020, 117(3), 1496-1503.

Figure legends

Figure 1 How RTC functions in "leader-to-body fusion"

Read on the antisense strands of the SARS-CoV-2 genome (GenBank: MN908947.3), "AAGAA" (in red color) and "GUUCGU" (in blue color) represent RNA methylation sites and NSP15 cleavage sites, respectively. **A**. The positions are the start and end positions of hairpins in the SARS-CoV-2 genome. NSP15 cleave the single RNA after U (indicated by arrows). These hairpins including the transcription regulatory sequence (TRS) motifs are defined as TRS hairpins (below polyN), which can be used to identify recombination regions in CoV genomes. We defined the hairpins opposite to the TRS hairpins as opposite TRS hairpins (above polyN). **B**. 5'-3' represents the strand of the SARS-CoV-2 genome. NSP15 cleave the single RNA after the last U (indicated by *).

Figure 2 TRS hairpins in 5 typical recombination regions

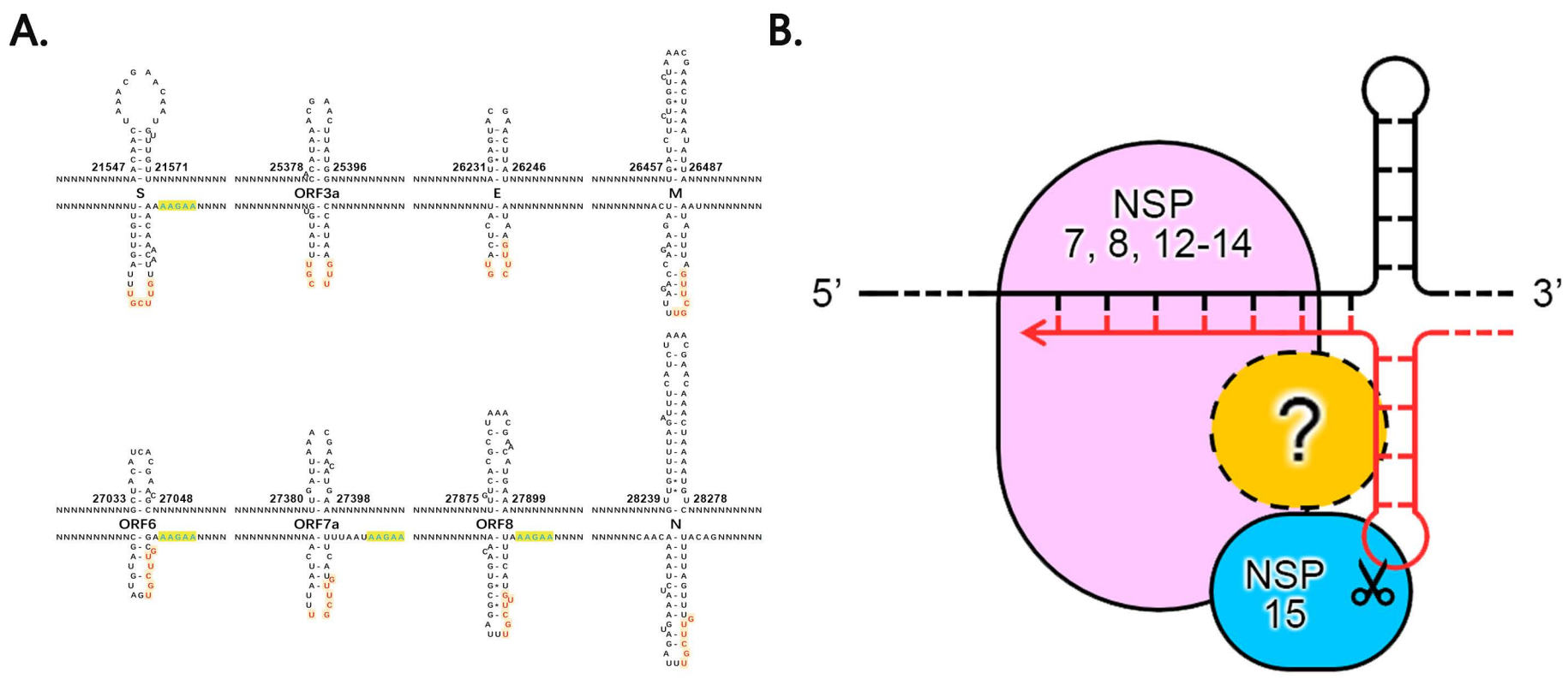
A-E have already been published in our previous study [11]. **A**. The genome sequences are from the SARS-like CoV strain WIV1 from bats (GenBank: KF367457) and SARS-CoV-2 (GenBank: MN908947); **B**. The genome sequences are from SARS-CoV-2 (GenBank: MN908947) and the SARS-CoV-2 strain Hongkong (GISAID: EPI_ISL_417443); **C**. The genome sequences are from SARS-CoV-2 (GenBank: MN908947) and the SARS-CoV-2 strain Singapore (GISAID: EPI_ISL_414378, EPI_ISL_414379 and EPI_ISL_414380); **D**. The genome sequences are from SARS-CoV-2 (GenBank: MN908947) and the mink SARS2-like CoV strain (GenBank: MT457390); **E**. The genome sequences are from the SARS-CoV strain GD01 (GenBank: AY278489) and the SARS-CoV strain Tor2 (GenBank: AY274119). **F**. CoV recombinant events occurred due to the cleavage of these non-canonical TRS hairpins.

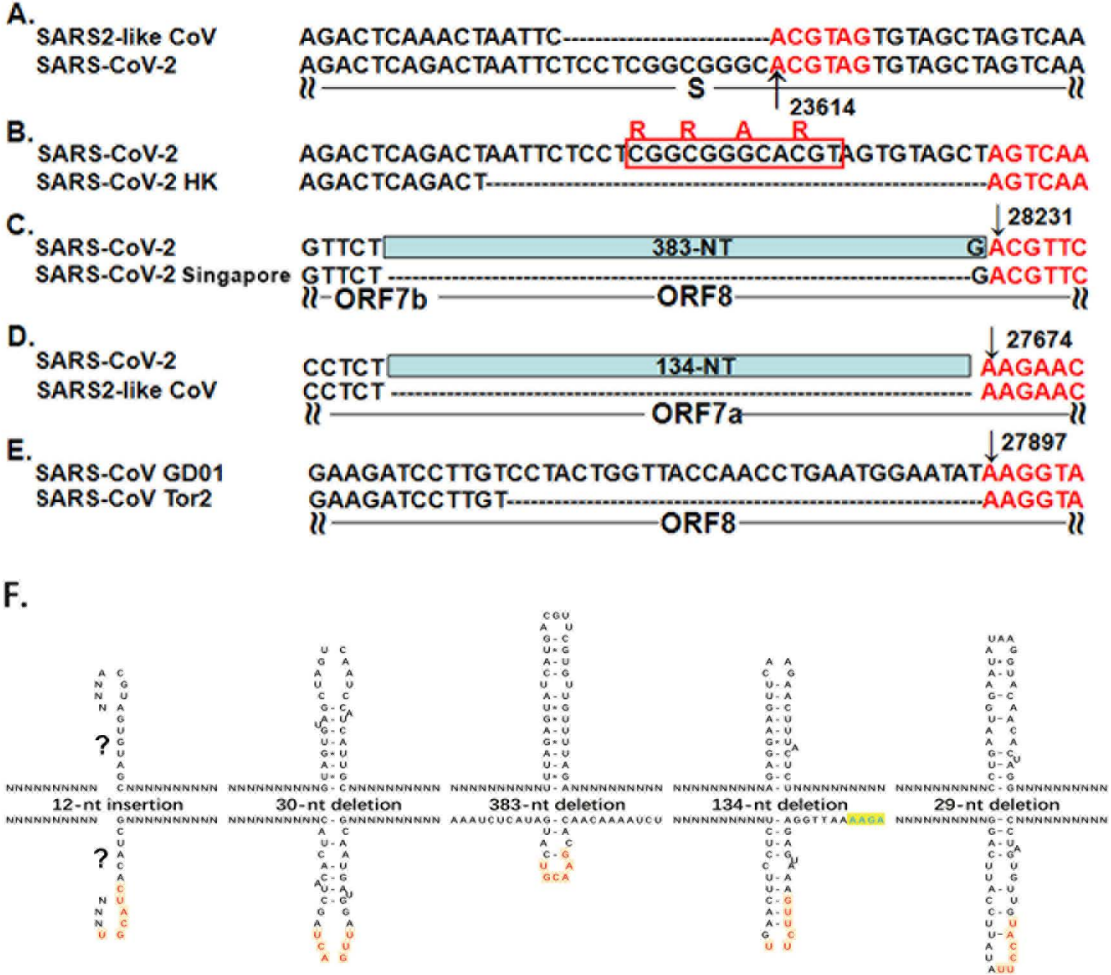
Figure 3 Three classes of possible hairpins

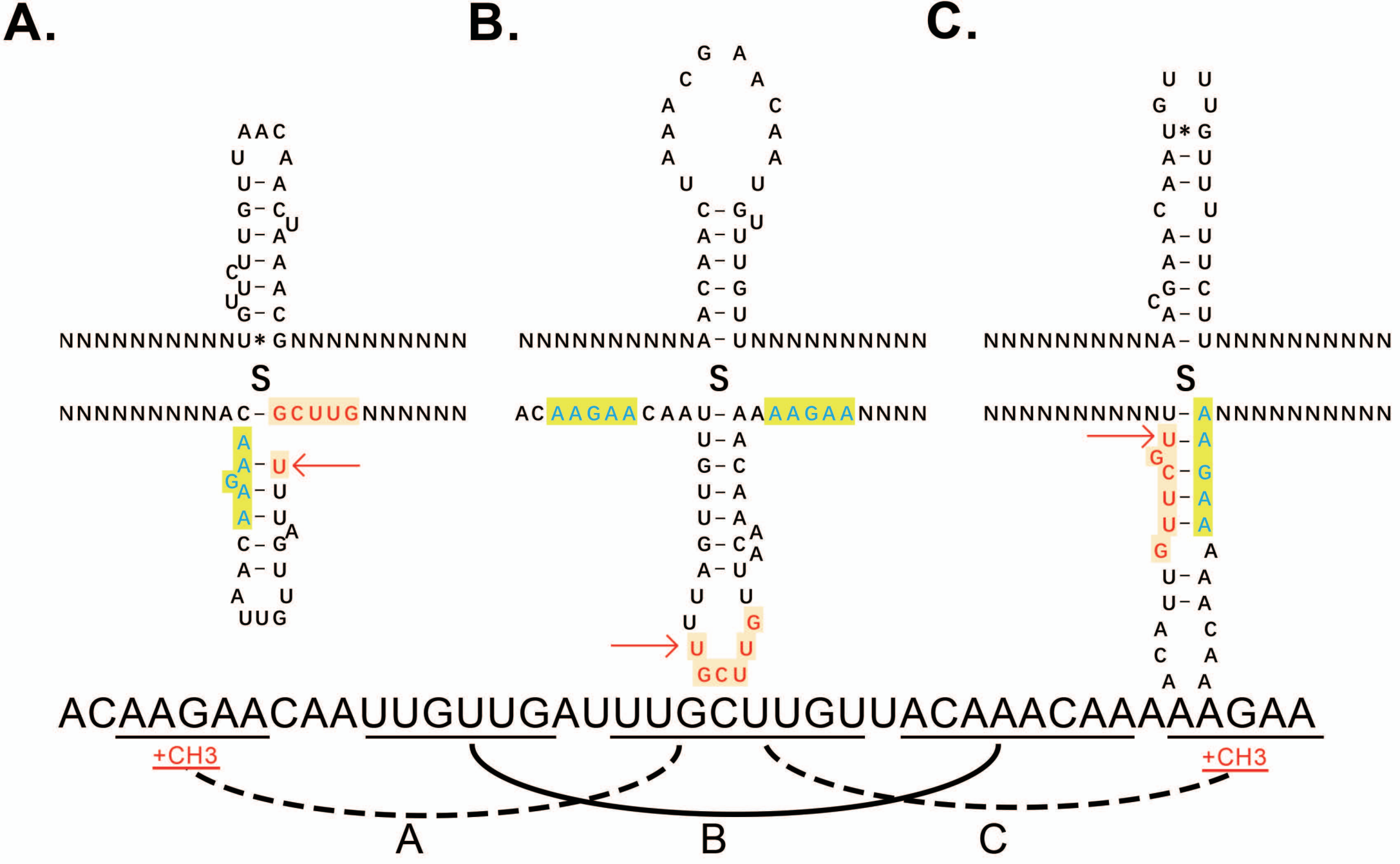
All possible hairpins in the TRS-Bs were classified into three classes. In the class 1 and 2. Using the TRS-B of the S gene of SARS-CoV-2 as an example, the first class (**A**) and the third class (**C**) require the "AAGAA-like" motif involved in the Watson-Crick pairing. However, the methylation at the "AAGAA-like" motif (in blue

377 color) is not in favour of Watson-Crick pairing, which ensures that the NSP15
378 cleavage site (in red color) resides in the loops of the second class of hairpins (**B**).

379


A.
21547
21571
S
25378
25396
ORF3a
26231
26246
E
26457
26487
M
27033
27048
ORF6
27380
27398
ORF7a
27875
27899
ORF8
28239
28278
N
B.
5’
3’
NSP
7, 8, 12-14
?
NSP
15


A.
SARS2-like CoV
SARS-CoV-2
AGACTCAAACTAATTC----------ACGTAGTGTAGCTAGTCAA
AGACTCAGACTAATTCTCCTCGGCGGGCACGTAGTGTAGCTAGTCAA
S
23614
B.
R R A R
SARS-CoV-2
SARS-CoV-2 HK
AGACTCAGACTAATTCTCCTCGGCGGGCACGTAGTGTAGCTAGTCAA
AGACTCAGACT----------AGTCAA
C.
28231
SARS-CoV-2
SARS-CoV-2 Singapore
GTTCT 383-NT GACGTTC
GTTCT ----------GACGTTC
ORF7b
ORF8
D.
27674
SARS-CoV-2
SARS2-like CoV
CCTCT 134-NT AAGAAC
CCTCT ---------- AAGAAC
ORF7a
E.
27897
SARS-CoV GD01
SARS-CoV Tor2
GAAGATCCTTGTCCTACTGGTTACCAACCTGAATGGAATATAAGGTA
GAAGATCCTTGT----------AAGGTA
ORF8
F.
12-nt insertion
30-nt deletion
383-nt deletion
134-nt deletion
29-nt deletion


A.
B.
C.
S
S
S
+CH3
+CH3
A
B
C