

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

3 Xin Li^{1§}, Qiang Zhao^{1§}, Jia Chang¹, Guangyou Duan², Jinlong Bei³

4 Tung On Yau⁴, Jianyi Yang⁵, Jishou Ruan⁵, Bingjun He^{1*}, Shan Gao^{1*}

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

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

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

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

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

13

14

15

16

17 § These authors contributed equally to this paper.

18 * Corresponding authors.

19 SG : gao_shan@mail.nankai.edu.cn

20 BH : hebj@nankai.edu.cn

21

22 **Abstract**

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

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

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

44

45

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

47

48 Introduction

49 Coronavirus disease 2019 (COVID-19) is caused by severe acute respiratory
50 syndrome coronavirus 2 (SARS-CoV-2) [1] [2]. SARS-CoV-2 has a genome of ~30
51 kb [3], including the genes *spike (S)*, *envelope (E)*, *membrane (M)*, *nucleocapsid (N)*,
52 and *ORF1a*, *1b*, *3a*, *6*, *7a*, *7b*, *8* and *10* [4]. The *ORF1a* and *1b* genes encode 16
53 non-structural proteins (NSPs), named NSP1 through NSP16 [5]. The other 10 genes
54 encode 4 structural proteins (S, E, M and N) and 6 accessory proteins (ORF3a, 6, 7a,
55 7b, 8 and 10). NSP4-16 are significantly conserved in all known CoVs and have been
56 experimentally demonstrated or predicted to be critical enzymes in CoV RNA
57 synthesis and modification [6], including: NSP12, RNA-dependent RNA polymerase
58 (RdRp) [7]; NSP13, RNA helicase-ATPase (Hel); NSP14, exoribonuclease (ExoN),
59 and methyltransferase; NSP15 endoribonuclease (EndoU) [8]; and NSP16, RNA
60 2-O-methyltransferase (MT).

61 NSP1-16 assemble into a replication and transcription complex (RTC) in CoV
62 [7]. The basic function of the RTC is RNA synthesis: it synthesizes genomic RNAs
63 (gRNAs) for replication and subgenomic RNAs (sgRNAs) for transcription [9]. CoV
64 replication and transcription can be explained by the prevailing “leader-to-body
65 fusion” model [9]. For a complete understanding of CoV replication and transcription,
66 much research has been conducted to determine the global structure of the
67 SARS-CoV-2 RTC, since the outbreak of SARS-CoV-2. Although some single
68 protein structures (e.g. NSP15 [8]) and local structures of the RTC (i.e.
69 NSP7-NSP8-NSP12-NSP13 [7] and NSP7-NSP8-NSP12 [10]) have been determined,
70 these structures have not yet answered how the RTC functions in the “leader-to-body
71 fusion” process. As the global structure of the CoV RTC cannot be determined by
72 simple use of any current methods (i.e., NMR, X-ray and Cryo-EM), it is necessary to
73 ascertain the arrangement of all the RTC components, leading to the eventual
74 determination of its global structure.

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

82

83 **Results**

84 **Molecular basis of “leader-to-body fusion” model**

85 Here, we provide a brief introduction to the “leader-to-body fusion” model
86 proposed in an early study [9] and its molecular basis proposed in our recent study
87 [11]. CoV replication and transcription require gRNAs(+) as templates for the
88 synthesis of antisense genomic RNAs [gRNAs(-)] and antisense subgenomic RNAs
89 [sgRNAs(-)] by RdRP. When RdRP pauses, as it crosses a body transcription
90 regulatory sequence (TRS-B) and switches the template to the leader TRS (TRS-L),
91 sgRNAs(-) are formed through discontinuous transcription (also referred to as
92 polymerase jumping or template switching). Otherwise, RdRP reads gRNAs(+)
93 continuously, without interruption, resulting in gRNAs(-). Thereafter, gRNAs(-) and
94 sgRNAs(-) are used as templates to synthesize gRNAs(+) and sgRNAs(+),
95 respectively; gRNAs(+) and sgRNAs(+) are used as templates for the translation of
96 NSP1-16 and 10 other proteins (S, E, M, N, and ORF3a, 6, 7a, 7b, 8 and 10),
97 respectively. The molecular basis of the “leader-to-body fusion” model as proposed in
98 our previous study is that NSP15 cleaves gRNAs(-) and sgRNAs(-) at TRS-Bs(-).
99 Then, the free 3' ends (~6 nt) of TRS-Bs(-) hybridize the junction regions of TRS-Ls
100 for template switching. NSP15 may also cleave gRNAs(-) and sgRNAs(-) at
101 TRS-Ls(-), which is not necessary for template switching. This molecular basis

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

105

106 **NSP15 cleavage, RNA methylation and 3' polyadenylation**

107 The cleavage sites of NSP15 contain the canonical TRS motif “GTTCGT” [11],
108 read on the antisense strands of CoV genomes. One study reported that RNA
109 methylation sites contain the “AAGAA-like” motif (including AAGAA and other
110 A/G-rich sequences) throughout the viral genome, particularly enriched in genomic
111 positions 28,500-29,500 [4]. Nanopore RNA-seq, a direct RNA sequencing method
112 [12], was used in the study that shares data for our reanalysis in the previous [11] and
113 present studies. By reanalysing the Nanopore RNA-seq data [4], we found that the
114 “AAGAA-like” motif co-occurred with the canonical TRS motif “GTTCGT” (**Figure**
115 **1A**) in TRS-Bs of eight genes (*S*, *E*, *M*, *N*, and *ORF3a*, *6*, *7a* and *8*). In addition, each
116 of these TRS-B contains many possible hairpins. These hairpins are encoded by
117 complemented palindrome sequences, which explained a finding reported in our
118 previous study: complemented palindromic small RNAs (cpsRNAs) with lengths
119 ranging from 14 to 31 nt are present throughout the SARS-CoV genome, however,
120 most of them are semipalindromic or heteropalindromic [13].

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

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

143 Another important phenomenon reported in the previous study [4] that merits
144 further analysis was that the “AAGAA-like” motif associates with the 3' poly(A)
145 lengths of gRNAs and sgRNAs. However, the study did not investigate the
146 association between the “AAGAA-like” motif and the nascent RNAs cleaved by
147 NSP15. The methylation at the “AAGAA-like” motif (read on the antisense strands of
148 CoV genomes) in TRS hairpins may affect the downstream 3' polyadenylation that
149 prevents the quick degradation of nascent RNAs (**Figure 1B**). Although the type of
150 methylation is unknown, a preliminary study was conducted, which revealed that
151 modified RNAs of SARS-CoV-2 have shorter 3' poly(A) tails than unmodified ones
152 [4]. However, there are two shortcomings in the interpretation of CoV RNA
153 methylation in this previous study: (1) it was not explained that many methylation
154 sites are far from 3' ends, which are unlikely to contribute to the 3' poly(A) tails; and
155 (2) the “AAGAA-like” motif on the antisense strand was not analyzed (**See below**), as
156 only a few antisense reads were obtained using Nanopore RNA-seq. Although the
157 associations between NSP15 cleavage, RNA methylation and 3' polyadenylation are

158 still unclear, they suggest that the RTC has a local structure composed of NSP15 and
159 16 to contain a TRS hairpin. This special local structure is able to facilitate the NSP15
160 cleavage and RNA methylation of the TRS hairpin at the opposite sides (**Figure 1B**).

161

162 **How RTC functions in “leader-to-body fusion”**

163 Since several A-rich and T-rich regions are alternatively present in each TRS-B,
164 it contains many possible hairpins. Thus, to determining which one is the TRS hairpin
165 needs decisive information. After comparing all possible hairpins in the TRS-Bs of
166 betacoronavirus subgroup B, we found that they can be classified into three classes.
167 Using the TRS-B of the S gene of SARS-CoV-2 as an example, the first class (**Figure**
168 **3A**) and the third class (**Figure 3C**) require the “AAGAA-like” motif involved in the
169 Watson-Crick pairing. However, the methylation at the “AAGAA-like” motif is not in
170 favour of Watson-Crick pairing. Further analysis of the “AAGAA-like” motif on the
171 antisense strand (**See above**) inspired us to propose a novel explanation of CoV RNA
172 methylation. RNA methylation of CoVs participates in the determination of the RNA
173 secondary structures by affecting the formation of hairpins. The methylation of
174 flanking sequences containing the “AAGAA-like” motif ensures that the NSP15
175 cleavage site resides in the loops of the second class of hairpins (**Figure 3B**).
176 Therefore, the second class of hairpin structures is the best choice for both the NSP15
177 cleavage and the “AAGAA-like” motif. The NSP15 cleavage site exposed in a small
178 loop, which facilitates the contacts of NSP15. This structure verified the results of
179 mutation experiments in a previous study [15] that the recognition of NSP15 cleavage
180 sites is independent on the TRS motif, but dependent on its context. These findings
181 confirmed that the recognition of NSP15 cleavage sites is structure-based rather than
182 sequence-based (**See above**).

183 NSP12-16 form the main structure of the RTC and work as a pipeline (**Figure**
184 **1B**). The RTC pipeline starts with NSP13 that unwind template RNAs [7]. Using
185 single-strand templates, NSP12 synthesizes RNAs with error correction by NSP14.

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

211

212 **Conclusion and Discussion**

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

226

227 **Materials and Methods**

228 1,265 genome sequences of betacoronaviruses (in subgroups A, B, C and D)
229 were downloaded from the NCBI Virus database
230 (<https://www.ncbi.nlm.nih.gov/labs/virus>) in our previous study [12]. Among these
231 genomes, 292 belongs to betacoronavirus subgroup B (including SARS-CoV and
232 SARS-CoV-2). Nanopore RNA-seq data was downloaded from the website
233 (<https://osf.io/8f6n9/files/>) for reanalysis. Protein structure data (PDB: 6X1B, 7BQ7,
234 7CXN) were used to analyzed NSP15, NSP10-NSP16 and
235 NSP7-NSP8-NSP12-NSP13, respectively. SARS-CoV were detected from 4 runs of
236 small RNA-seq data (NCBI SRA: SRR452404, SRR452406, SRR452408 and
237 SRR452410). Data cleaning and quality control were performed using Fastq_clean
238 [18]. Statistics and plotting were conducted using the software R v2.15.3 with the
239 Bioconductor packages [19]. The structures of NSP12-16 were predicted using
240 trRosetta [20].

241

242

243 **Supplementary information**

244

245 **Declarations**

246 **Ethics approval and consent to participate**

247 Not applicable.

248

249 **Consent to publish**

250 Not applicable.

251

252 **Availability of data and materials**

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

254

255 **Competing interests**

256 The authors declare that they have no competing interests.

257

258 **Funding**

259 This work was supported by the National Natural Science Foundation of China
260 (31871992) to Bingjun He, Tianjin Key Research and Development Program of China
261 (19YFZCSY00500) to Shan Gao and National Natural Science Foundation of China
262 (31700787) to Guangyou Duan. The funding bodies played no role in the study design,
263 data collection, analysis, interpretation or manuscript writing.

264

265 **Authors' contributions**

266 Shan Gao conceived the project. Shan Gao and Bingjun He supervised this study.

267 Guangyou Duan and Jia Chang performed programming. Xin Li and Qiang Zhao,

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

271

272 **Acknowledgments**

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

280

281 **REFERENCES**

- 282 [1] X. Li, G. Duan, W. Zhang, J. Shi, J. Chen, S. Chen, S. Gao, and J. Ruan. A Furin
283 Cleavage Site Was Discovered in the S Protein of the 2019 Novel Coronavirus.
284 Chinese Journal of Bioinformatics (In Chinese) 2020, 18(2): 103-108.
- 285 [2] G. Duan, J. Shi, Y. Xuan, J. Chen, C. Liu, J. Ruan, S. Gao, and X. Li. 5' UTR
286 Barcode of the 2019 Novel Coronavirus Leads to Insights into Its Virulence. Chinese
287 Journal of Virology (In Chinese) 2020, 36(3): 365-369.
- 288 [3] C. Jiayuan, S. Jinsong, O. Yau Tung, L. Chang, L. Xin, Z. Qiang, R. Jishou, and G.
289 Shan. Bioinformatics Analysis of the 2019 Novel Coronavirus Genome. Chinese
290 Journal of Bioinformatics (In Chinese) 2020, 18(2): 96-102.
- 291 [4] D. Kim, J.-Y. Lee, J.-S. Yang, J.W. Kim, V.N. Kim, and H. Chang. The
292 Architecture of SARS-CoV-2 Transcriptome. Cell 2020, 181(4): 914-921.
- 293 [5] S.J.R. da Silva, C.T. Alves da Silva, R.P.G. Mendes, and L. Pena. Role of
294 nonstructural proteins in the pathogenesis of SARS-CoV-2. J Med Virol 2020, 92:
295 1427-1429.
- 296 [6] Denison MR, Graham RL, Donaldson EF, Eckerle LD, Baric RS. Coronaviruses:
297 an RNA proofreading machine regulates replication fidelity and diversity. RNA Biol.
298 2011, 8(2):270-279.
- 299 [7] L. Yan, Y. Zhang, J. Ge, L. Zheng, Y. Gao, T. Wang, Z. Jia, H. Wang, Y. Huang,
300 M. Li, Q. Wang, Z. Rao, and Z. Lou, Architecture of a SARS-CoV-2 mini replication
301 and transcription complex. Nature Communications 2020, 2020(2020): 1-6.

- 302 [8] Y. Kim, R. Jedrzejczak, N.I. Maltseva, M. Wilamowski, M. Endres, A. Godzik, K.
303 Michalska, and A. Joachimiak. Crystal structure of NSP15 endoribonuclease NendoU
304 from SARS-CoV-2. *Protein Science* 2020, 29(7): 1596-1605.
- 305 [9] S.G. Sawicki, and D.L. Sawicki. A New Model for Coronavirus Transcription. in:
306 L. Enjuanes, S.G. Siddell, and W. Spaan, (Eds.). *Coronaviruses and Arteriviruses*,
307 Springer US, Boston, MA, 1998, pp. 215-219.
- 308 [10] H.S. Hillen, G. Kokic, L. Farnung, C. Dienemann, and P. Cramer. Structure of
309 replicating SARS-CoV-2 polymerase. *Nature* 2020, 584(7819): 1-6.
- 310 [11] Xin Li, Zhi Cheng, Fang Wang, Jia Chang, Qiang Zhao, Hao Zhou, Chang Liu,
311 Jishou Ruan, Guangyou Duan, Shan Gao. A negative feedback model to explain
312 regulation of SARS-CoV-2 replication and transcription. *Frontiers in Genetics* 2021,
313 10: 1-11.
- 314 [12] X. Xu, H. Ji, X. Jin, Z. Cheng, X. Yao, Y. Liu, Q. Zhao, T. Zhang, J. Ruan, W.
315 Bu, Z. Chen, and S. Gao. Using pan RNA-seq analysis to reveal the ubiquitous
316 existence of 5' and 3' end small RNAs. *Frontiers in Genetics* 2019, 10: 1-11.
- 317 [13] Liu C, Chen Z, Hu Y, Ji H, Yu D, Shen W, Li S, Ruan J, Bu W, Gao S.
318 Complemented Palindromic Small RNAs First Discovered from SARS Coronavirus.
319 *Genes* 2018, 9(9): 1-11.
- 320 [14] Xin Li, Jia Chang, Shunmei Chen, Liangge Wang, Tung On Yau, Qiang Zhao,
321 Zhangyong Hong, Jishou Ruan, Guangyou Duan and Shan Gao. Genomic feature
322 analysis of betacoronavirus provides insights into SARS and COVID-19 pandemics.
323 *bioRxiv*. 2020(2020): 1-8.
- 324 [15] Yount B , Roberts R , Lindesmith L, et al. Rewiring the severe acute respiratory
325 syndrome coronavirus (SARS-CoV) transcription circuit: Engineering a
326 recombination-resistant genome[J]. *Proceedings of the National Academy of Sciences*
327 *of the United States of America*, 2006, 103(33) : 12546–12551.
- 328 [16] Lianqi Z , Lei L , Liming Y, et al. Structural and Biochemical Characterization of
329 Endoribonuclease Nsp15 Encoded by Middle East Respiratory Syndrome
330 Coronavirus. *Journal of Virology*, 2018, 92.
- 331 [17] Krafcikova P , Silhan J , Nencka R , et al. Structural analysis of the SARS-CoV-2
332 methyltransferase complex involved in RNA cap creation bound to sinefungin. *Nature*
333 *Communications*, 2020, 11(1):3717.
- 334 [18] M. Zhang, F. Zhan, H. Sun, X. Gong, Z. Fei, and S. Gao. Fastq_clean: An
335 optimized pipeline to clean the Illumina sequencing data with quality control,
336 *Bioinformatics and Biomedicine (BIBM)*, 2014 IEEE International Conference on,
337 IEEE, 2014, pp. 44-48.
- 338 [19] S. Gao, J. Ou, and K. Xiao. R language and Bioconductor in bioinformatics
339 applications(Chinese Edition), Tianjin Science and Technology Translation
340 Publishing Ltd, Tianjin, 2014.

341 [20] Yang J, Anishchenko I, Park H, Peng Z, Baker D. Improved protein structure
342 prediction using predicted interresidue orientations. Proceedings of the National
343 Academy of Sciences 2020, 117(3), 1496-1503.
344

345 **Figure legends**

346 **Figure 1 How RTC functions in “leader-to-body fusion”**

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

357

358 **Figure 2 TRS hairpins in 5 typical recombination regions**

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

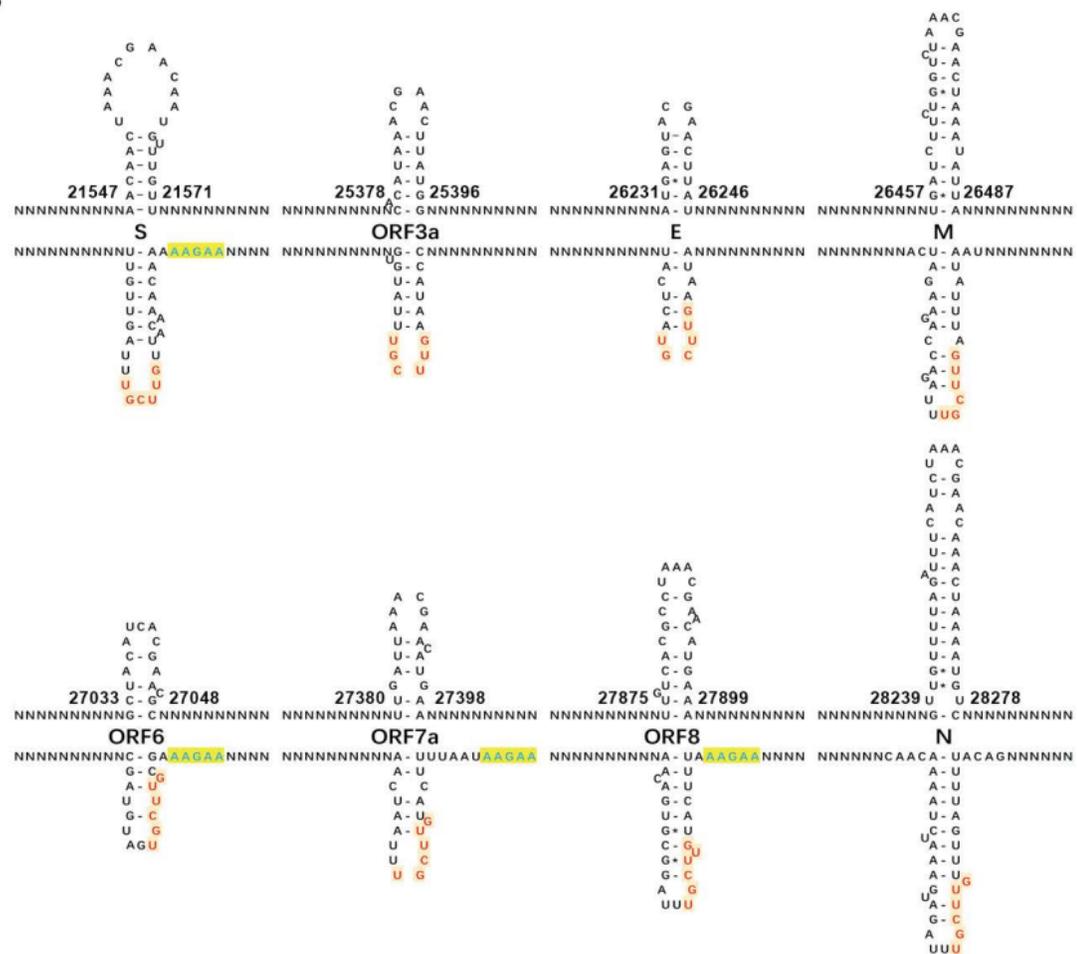
371

372 **Figure 3 Three classes of possible hairpins**

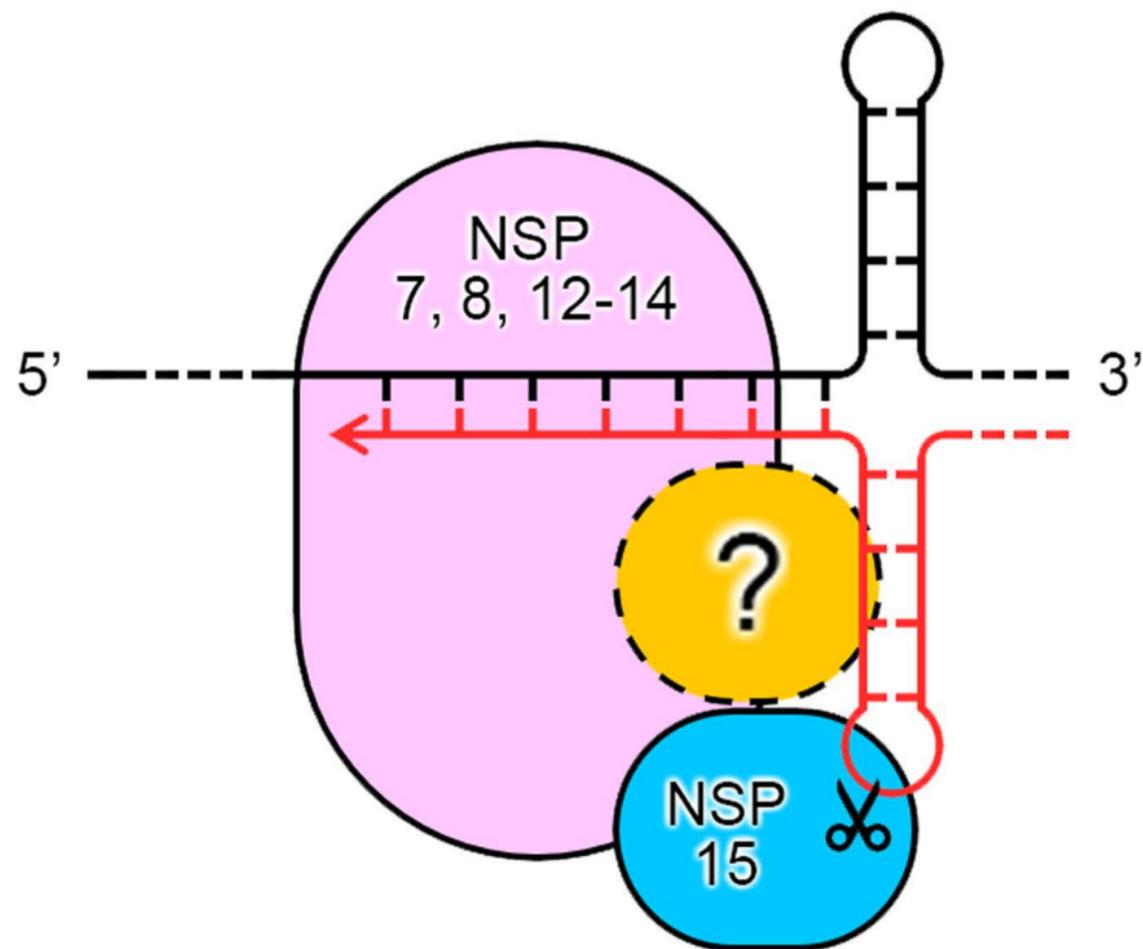
373 All possible hairpins in the TRS-Bs were classified into three classes. In the class 1
374 and 2. Using the TRS-B of the S gene of SARS-CoV-2 as an example, the first class
375 (**A**) and the third class (**C**) require the “AAGAA-like” motif involved in the
376 Watson-Crick pairing. However, the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

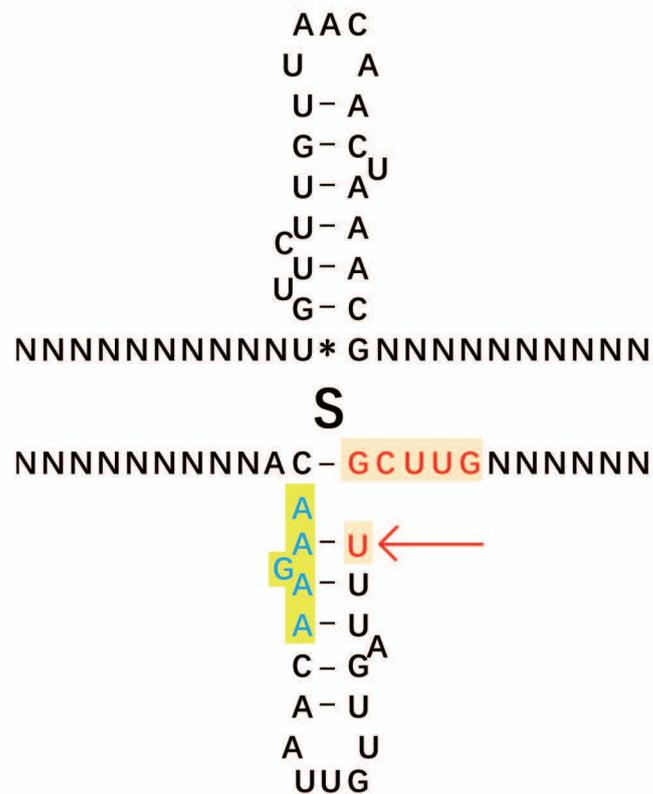
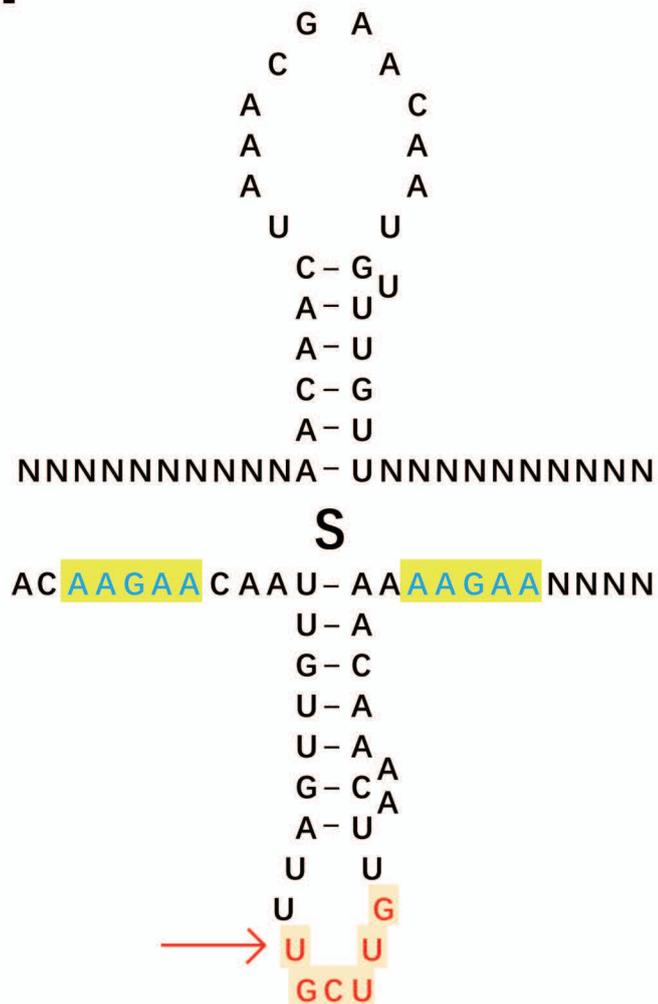
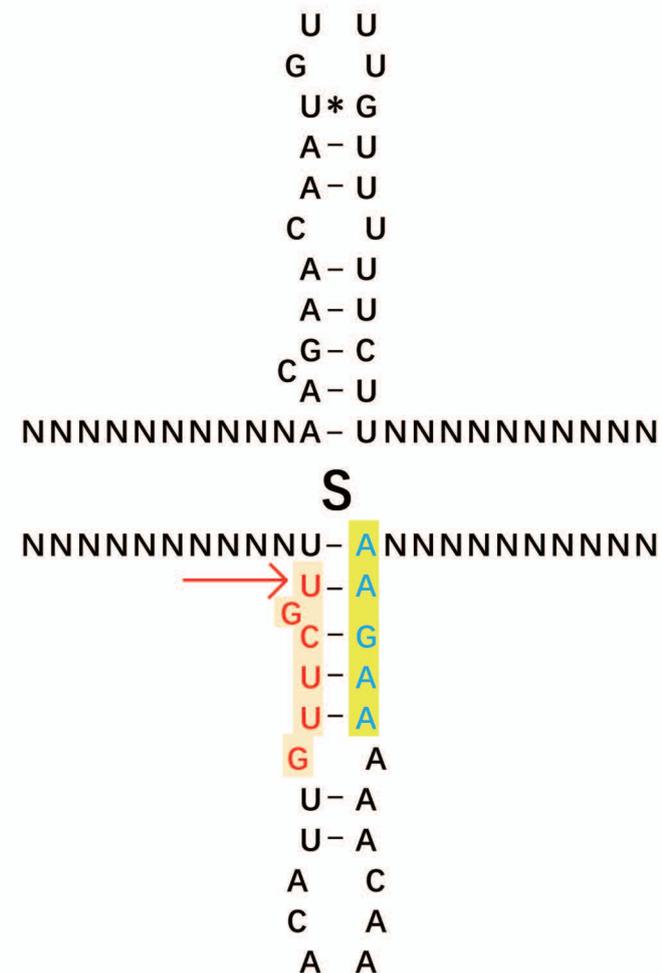
377 color) is not in favour of Watson-Crick pairing, which ensures that the NSP15
378 cleavage site (in red color) resides in the loops of the second class of hairpins (**B**).
379

A.



B.



A.**B.****C.**

ACAAGAACAAUUGUUGAUUUGCUUGUUACAAACAAAAGAA

+CH3+CH3

A

B

C